Running head: A SOFTWARE USED TO ANALYZE AND SELECT CHINESE CHARACTERS

C-CAT: A Computer Software Used to Analyze and Select Chinese Characters and Character Components for Psychological Research

Ming Lo and Chih-Wei Hue

National Taiwan University

Corresponding Author: Chih-Wei Hue

Department of Psychology, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei, 10617, Taiwan.

Telephone: 886-2-33663086

E-Mail: hue@ntu.edu.tw

Abstract

This computer software was designed to assist researchers to construct experimental materials using traditional Chinese characters. In the software package, there are two sets of character stocks in which one is suitable for research using literate adults as subjects and the other for research using school children as subjects. The software can identify the linguistic properties of a character, or a character component, such as the number of strokes contained, character/component pronunciation regularity, and the arrangement of character components in a character. Moreover, it can compute the character's linguistic frequency, neighborhood size, and phonetic validity with respect to a user selected character stock. It can also search the character stock selected for similar characters, or character components with the linguistic properties specified by user.

Introduction

Most psycholinguistic theories of visual word recognition have been built on alphabetic writing systems – mainly English (Barber & Kutas, 2007). However, it has been shown that the recognition processes of visual words are affected by the principles that the words represent phonology (Frost, Katz, & Bentin, 1987). In modern Chinese, most words consist of two syllables. In writing the words, each syllable is represented by a character which is roughly equal to a morpheme. On the surface, a character is a logo like symbol composed of pen strokes. Although, there are indeed a small number of characters which are logograms, for most of the frequently used characters, they are compounds of simpler characters. For these characters, the mapping from orthography to phonology and semantics is not opaque. According Karlgren's computation (see DeFrancis, 1984), about 90% of the 6000 common traditional Chinese characters[1] he analyzed are semantic phonological compounds which are composed of a radical, and a phonetic component. The radical is the semantic component of a semantic phonological compound, and often provides clues to the compound character's meaning, and the phonetic component hints to the pronunciation of the character. Karlgren estimated that in semantic phonological compounds, 23% show a clear relationship in meaning between radical and character, and 64% show a relationship in pronunciation between phonetic component and character. Both a radical and a phonetic component can appear in more than one character, which creates a complex relationship in semantics or in

phonology among the characters sharing a common component.

There are studies designed to investigate how the different principles used in character construction and alphabetic word spelling affect the two types of writings processed psychologically. In these studies, some were designed to test the universality of the existing theories of visual word recognition, and some were designed to see if the logographical characters afford cognitive processes different from that of the alphabetical words. For instance, to test if the activation-synthesis model of English word recognition proposed by Glushko (1979) is applicable to explain Chinese character recognition, a neighborhood size (or combinability of the stimulus characters) was manipulated in studies by Hsiao, Shillcock, and Lavidor (2006), Feldman and Siok (1997, 1999), and Taft and Zhu (1997). “Neighbors” share a same component, and therefore, to compute the neighborhood size of a character, a researcher has to examine thousands of characters, one by one, to identify the character’s neighbors. Moreover, if character pronunciation consistency is of concern in an experiment, the neighbors of the stimulus characters have to be identified, and their pronunciations have to be checked. It is not difficult to see that the laborious procedure is error prone.

Character-Component Analysis Toolkit (C-CAT)

C-CAT is a computer software package designed to assist a researcher in constructing experimental materials using Chinese characters. The functions of C-CAT and the principles

that the software uses to decompose a character into its constituent components were designed in reference to the variables manipulated in a set of 26 recently published or frequently cited studies accessing the cognitive processes involved in character recognition (Table 1). The variables manipulated in these studies could be categorized into three types. The variables of the first type represent certain characteristic of a character (e.g., character type, character complexity, and linguistic frequency). The variables of the second type represent the orthographic, phonological or semantic relationships among a group of characters (e.g., character consistency). The third type of variable represents the relationships between a character and its constituent components (e.g., character regularity). Including a second or a third type of variable in an experiment indicates that while considering the property of a character, a researcher will not only take the characteristics of the character into account but also the characteristics of its constituent components. In the studies reviewed, most of the researchers, except Chen, Allport and Marshall (1996) and Taft and Zhu (1997), treated a character as composed of two components, a radical plus another component, despite that the "other" component is also a compound character. For example, the character "subjectively" (Fig. 1a, /yi4/) will be treated as a composite of the characters "moon" (Fig. 1b, /yue4/) and "meaning" (Fig. 1c, /yi4/), instead of "moon", "sound" (Fig. 1d, /yin1/) and "heart" (Fig. 1e, /xin1/). In light of the results of the analysis, C-CAT will analyze a character at most into two components according to the following rules. First, a character is treated as

composed of a radical and another component, if striping the radical down, the remaining part of the character is a radical, a legal character, or a constituent component of another character. For example, the character “elder sister” (Fig. 1f, /jie3/) is treated as a composite of the radical “female” (Fig. 1g, /nv3/) and the remaining part (Fig. 1h). The remaining part is neither a radical nor a character, but C-CAT treats it as a constituent component because it is in the characters “one trillion” (Fig. 1i, /zi3/) and “bed” (Fig. 1j, /zi3/). Secondly, a character is treated as one component unit if it is a radical or when striping the radical down, the remaining part of the character is neither an independent character nor a constituent component of other characters. For example, character “several” (Fig. 1k, /ji3/) is treated as a one component character. The radical of the character is shown in Fig. 1m. Striping the radical down from the character “several”, the remaining part is neither a character nor a component of any other character. Thirdly, variant written forms of a radical are treated as different radicals. For example, the radical shown in Fig. 1m and Fig. 1n are variant forms of the radical “hand” (/shou3/), C-CAT treats them as two different radicals.

Please insert Table 1 here

Please insert Figure 1 here

C-CAT is designed to provide two types of functions. First, it can analyze the properties of a character or a character component as specified by a user. For example, when a user inputs a character, C-CAT will compute the number of strokes that the character is comprised of, its linguistic frequency and neighborhood size. If a component is specified, C-CAT will compute the component's neighborhood size and phonetic validity (see below). Secondly, C-CAT can select a small set of characters from a character stock according to the conditions (e.g., linguistic frequency, phonological relationship between a character and its phonetic component, etc.) specified by a user.

C-CAT consists of an execute program and a database, which are packed in a setup file. The program performs character (and component) analysis (and selection) and user interface. The database contains the pronunciations of the 13047 traditional Chinese characters included in the Microsoft East Asian languages files. In addition, based on the works of Chuang and Hsieh (2005) and Hsieh and Chuang (1995), the assemblage of components of each of the 13047 characters are analyzed and listed in the database. Thirdly, there is a set of internal codes representing the components which are used to construct Chinese characters (Chuang & Hsieh, 2005), but are not included in the Microsoft East Asian languages files. To install C-CAT, one needs to simply double click on the setup file, and the installation will

proceed and finish automatically. A C-CAT icon will be created on the desktop once the software is installed.

To run the software, one needs to double-click on the C-CAT icon, which will activate a function selection window. In the window, there are four buttons representing the four functions the software performs. A parameter setting window will be presented after a user selects a function (Figure 2). In the window, C-CAT will prompt the user to enter two types of information. The first concerns with the "character stock" that the stimulus characters of an experiment are to be selected from. The second concerns with the linguistic properties of a to be analyzed or selected character (or component). In other words, "Character stock" defines the targeting set of the characters that C-CAT operates on. There are seven character stocks stored in C-CAT, and thus, a user has to specify which stock is to be used before character (component) analysis and selection can be performed.

Please insert Figure 2 here

Character Stocks

Characters are not only the building blocks of Chinese words, but also the main entries of the mental lexicon of a native Chinese speaker (Hoosian, 1992). However, not all the

Chinese characters are suitable to be used as stimulus materials in a study investigating the cognitive processes involved in character recognition. There are nearly fifty thousands characters listed in the Kang-Xi Dictionary, and a common character dictionary lists about 15000. However, it is estimated that a Chinese college student knows only about 5150 characters (Hue, 2003). This implies that in a study investigating the content or the operation of mental lexicon, the stimulus characters have to be selected from a set of characters representing the sight vocabulary of the subjects. With this consideration in mind, two types of character stocks are stored in C-CAT. One is suitable for a study using college students as subjects and the other can be used when elementary students are subjects.

The corpus prepared by the Chinese Knowledge Information Processing Group (CKIPG, 1993) is a collection of 20 million Chinese characters sampled from the articles published in three major Taiwanese news papers and a magazine. According to the CKIPG, there are 5656 different characters used to compose the articles collected (Table 2). In C-CAT, the 5656 characters were ranked according to their frequency of occurrence in the corpus. Judging form the number of characters in the character stock and the nature of the source where the articles included in the corpus were sampled, the CKIPG character stock is suitable for study using college students, or people with equivalent education, as subjects.

The electronic files of the articles contained the Chinese textbooks used in the elementary schools in Taiwan were obtained from the publisher (NICT; National Institution

for Compilation and Translation, 1995). Analysis of the characters used in the articles showed that cumulatively, there were 399, 896, 1322, 1814, 2306, and 2687 different characters used for the textbooks from the first to the sixth grade respectively. The characters of a NITC character stock were ranked according to their frequencies of occurrence in a respective corpus. For example, in the NICT 1st-grade corpus and character stock, the character "home" (Fig. 1o, /jia1/) occurs 25 times and ranks 24th. The same character occurs 398 times and ranks 30th in the NICT 1st-6th-grade corpus and character stock, respectively. The six character stocks are suitable for study using elementary school students as subjects.

Please insert Table 2 here

Character Analysis Function

The function is capable of taking a character input by a user, and computes the number of strokes the character contained and its linguistic frequency with respect to a corpus specified by the user. In addition, C-CAT is able to perform two kinds of computation relating to the pronunciation of the character. First, it can analyze the components of the target character, in terms of how the components are arranged in the character, and with respect to a component, the character's phonological regularity and neighborhood size. Secondly, it can

fetch the characters from a character stock in accordance with the phonological constraint specified by the user.

To use the function, a user needs to input a character and specify the "character stock" containing the to-be analyzed character (Fig. 2a). To analyze the regularity of the character, the user needs to specify the phonological relationship between the character and a component (e.g., same vowel) that defines the regularity. To find a set of characters with similar pronunciation as the input character, the user needs to define what similar pronunciation is (e.g., homophone excluding tone).

Character/component phonological relationship. In English, word regularity refers to the degree to which words follow the spelling-sound rules. In Chinese, Hue (1992) defined the character regularity in a similar manner. A regular character has a pronunciation similar to the pronunciation of one of its components (i.e., the phonetic component), and an irregular character does not. C-CAT allows its user to define character regularity by specifying the phonological relationship between a character and its components. A user can define "regular" as that a character has the same pronunciation (with or without tone) as one of its components, or as that the pronunciations of a character and one of its components share the same vowel (or same consonant).

Between characters phonological relationship. If a user needs to locate the characters in the character stock, which share certain phonological feature with the input character,

C-CAT allows the user to specify the condition of retrieval. The user has five options to select from. The first four represents the four types of phonological relationships between characters, i.e., homophones, homophones excluding tone, same vowel, and same consonant. If such function is not needed, the user can select the fifth option, which is “no specification”.

In Table 3, an example of using this function to analyze the character “dried up” (Fig. 1p, /ku1/) is presented. According to the linguistic properties specified by the user, the to-be analyzed character is sampled from the NICT 1st-6th-grade character stock, and “regular” is defined as “the pronunciations of a character and one of its components share a same vowel”. In addition, it is required that the homophones (excluding tone) of the input character are to be retrieved from the character stock. Table 3 also presents the results of analysis. In the output, C-CAT prints the numbers that the characters contained in the corpus and character stock selected, then, the results of analysis. The results showed that in the character corpus that the character stock was derived from, “dried up” occurs 5 times, and the frequency of character ranks 1442nd in the stock of 2687 characters. The character is comprised of 9 strokes, and has two left-right arranged components “wood” (Fig. 1q, /mu4/) and “ancient” (Fig. 1r, /ku3/). The pronunciations of both components share the same vowel with that of the character “dried up”. Finally, in the character stock, the component “wood” is included as a component in another 110 characters (i.e., the neighbors of “dried up”), and the component “ancient” is included in another 11 characters. Finally, C-CAT finds five characters matching

the between-character phonological relation specified. In the output, the characters are presented along with their frequencies and frequency ranks.

Please insert Table 3 here

Character Selection Function

In research studying Chinese character recognition, a researcher is restricted by the purpose of the study concerning what kind of characters can be used as stimuli. Usually, s/he is looking for a set of characters with certain linguistic properties which is difficult to obtain because s/he has to check thousands of characters in order to search for the ones meeting their needs. C-CAT is able to search a character stock for the characters with the linguistic properties specified by a user. The following are the linguistic properties that C-CAT recognizes: character/component phonological relationship, frequency rank, arrangement of constituent components in a character, and number of strokes (Fig. 2b).

Character/component phonological relationship. A user can specifically ask C-CAT to select regular characters from a character stock. If this feature is to be used, the user should define regularity in terms of the phonological relationship between a character and its constituent components (see the section on character analysis function).

Frequency. In C-CAT, the characters of a character stock are ranked according to their frequencies of occurrence in a respective character corpus. Thus, a user can use either “rank” or “frequency” to indicate the familiarity value of the characters to be selected.

Arrangement of components in a character. Most of the common characters in use are compound characters composed of a few constituent components. The components of a compound character are often arranged side by side, in a left-and-right arrangement. To a lesser extent, the components are arranged in a top-down fashion. In C-CAT, a user can specify what kind of compound characters are to be selected. There are four choices to select from: left-and-right, top-down, surrounded (one component encompasses the other components), and no specification.

An example is presented in Table 4. In this example, the user wants to select the characters with the following properties from the 1000 most frequently used characters in the CKIPG character stock. The selected characters should be composed of 8 to 12 strokes, and their constituent components are left-and-right arranged. In addition, the user defines “regular” as “the pronunciations of a character and one of its components are the same (excluding tone)”. Thus, a selected character should have a component which has the same pronunciation, except the tone, as that of the other character. In the 5656 characters contained in the CKIPG character stock, there are 53 characters meeting the criterion. In Table 4, one of the selected characters, “old” (Fig. 1s, /gu4/), is used as an example. The character “old” is

comprised of 9 strokes, and in the CKIPG corpus, its linguistic frequency and frequency rank are 3980 and 732nd, respectively. The character has 19 neighbors in the CKIPG character stock, each of the neighbors has the character "ancient" (Fig. 1r, /ku3/) as a constructing components.

Please insert Table 4 here

Character-Component Analysis Function

The information required for performing this function is shown in Figure 2c. When a user inputs a character component, C-CAT is able to compute the component's phonetic validity and to search through a character stock specified by the user for all the characters containing the component. The phonetic validity is the number of characters in the stock containing the component dividing the number of regular characters containing the component (Lo, Hue, & Tsai, 2007), where regularity is defined by the "Character/component phonological relationship". In addition, a user can restrict the range of characters to be searched for in terms of the specific position of the component as found in a character, e.g., the right half of a left-right character or otherwise.

As shown in Table 5, in the NICT 1st-6th-grade character stock, the component

"ancient" (Fig. 1r, /ku3/) is contained in 11 characters without specifying its position in a character. Using the user's definition of regularity (i.e., the pronunciations of a character and one of its components share the same vowel), the phonetic validity of this component is 0.8. In addition, the frequency-weighted value of phonetic consistency can be computed from the output (Jared, McRae, & Seidenberg, 1990). For the component "ancient", the value is 0.8.

Please insert Table 5 here

Character-Component Selection Function

C-CAT is able to search through a character stock for those characters which can be used to construct other characters. C-CAT can perform the search based upon the restriction of the "character stock" and the three linguistic properties specified by a user: number of strokes, neighborhood size, and phonetic validity (Fig. 2d). In the example presented in Table 6, the user asked C-CAT to search through the CKIPG character stock for all the character components which are comprised of 5 to 10 strokes and are included in 5 to 15 different characters as a component. In addition, the user defined that the pronunciation of a regular character shares the same consonant with one of its components. With this definition of regularity, the phonetic validity values of each of the selected "components" (or rather,

characters) are computed. As a result, C-CAT found 194 character components meeting the criteria. For example, the component “bureau” (Fig. 1t, /si1/) is a “simple”, but independent, Chinese character, which is composed of 5 strokes, and is a constituent component of 5 “complex” characters. In three of the five characters, their pronunciations share the same consonant /s/ with “bureau”, and thus, the phonetic validity of “bureau” is 0.6.

Please insert Table 6 here

System Requirements and Saving C-CAT’s Output

C-CAT can be installed in a personal computer (PC) with or without a Traditional-Chinese version of WindowsXP. If C-CAT is to be installed in a PC running a non-Traditional-Chinese WindowsXP, the software requires the following system configuration: (1) an WindowsXP operating system, (2) the Microsoft East Asian languages files, and (3) the Microsoft Global IME (Input Method Editor). Although the files for the East Asian languages and the Global IME are not routinely installed along with WindowsXP, they are included in the WindowsXP package. The procedure to install them can be found on the Microsoft official web site. The default text editor for the output of C-CAT is the Windows Notepad. It is suggested that a user save an output in a comma-separated format (i.e., the

CSV file type). In this file format, the output can be read by MS Excel which will present them in a spreadsheet.

Limitations

In C-CAT, there is a set of principles guiding how a character can be decomposed into components. These principles are necessary, because without their constraints, it is difficult to design an executable and efficient algorithm for character analysis. However, these principles also create limitations for C-CAT. First, research results indicate that the radical of a phonetic component affects lexical decision of the character containing the phonetic component (Taft & Zhu, 1997). In an experiment studying further the issue, the characteristics of the sub-components of a stimulus character's constituent components have to be taken into account while designing the experiment. However, C-CAT is only able to analyze a character into two components, and computes the characteristics of the three.

Secondly, in C-CAT, variant forms of a character are treated as different characters. On the other hand, a character and a radical sharing a same form are treated as a same character by C-CAT and share a same entry. This may create miscalculation for certain characteristics of a character containing a component with variant written form. For example, the components in Fig. 1u and 2e are variant written forms of a same radical "heart" (/xin1/). In the NICT 1st-2nd-grade character stock, there are 10 characters containing the component in

Fig. 1u and 11 containing the one in Fig. 1e. Because the component in Fig. 1e is also a character, and thus a miscalculation will be resulted when computing the neighborhood size of a character using it as the phonetic component. Similarly, it is not know from the output of C-CAT that how many character uses “heart” as radical.

Thirdly, in C-CAT, a character with multiple pronunciations has multiple entries. Thus, a character with two pronunciations will be processed two times while the character is the target of analysis. In such circumstance, redundant information is often presented in C-CAT’s output, and this will increase reading difficulty.

A revised edition of C-CAT is planning with the above mentioned limitations under consideration. However, we have to admit that the task is difficult, because new algorithms are to be invented so that a character can be analyzed into a set of “meaningful” components and that C-CAT can recognize a character’s variant forms and a component’s role in a character (e.g., a radical).

References

Barber, H. A., & Kutas, M. (2007). Interplay between computational models and cognitive electrophysiology in visual word recognition. *Brain Research Reviews, 52,* 98-123.

Chen, Y. P., Allport, D. A., & Marshall, J. C. (1996). What are the functional orthographic units in Chinese word recognition: The stroke or the stroke pattern? *Quarterly Journal of Experimental Psychology: Section a-Human Experimental Psychology, 49,* 1024-1043.

Chinese Knowledge Information Processing Group (1993). *Corpus-based frequency count of characters in journal Chinese: Corpus-based research* (no.1). Taipei, Taiwan: Academia Sinica Institute of Information Science.

Chuang, D. M., & Hsieh , C. C. (2005). Design and application of a Chinese character database. *Proceedings of International Conference on Chinese Characters and Globalization* (pp. 119-133), Taipei, Taiwan: Taipei City Government.

Defrancis, J. (1984). *The Chinese language: Fact and fantasy*. Honolulu, HI: University of Hawaii Press.

Ding, G. S., Peng, D. L., & Taft, M. (2004). The nature of the mental representation of radicals in Chinese: A priming study. *Journal of Experimental Psychology: Learning Memory & Cognition, 30,* 530-539.

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. Gao & R. Hoosain (Eds.), *Linguistics, psychology, and the Chinese Language* (pp. 11-21). Hong Kong: University of Hong Kong, Center of Asian Studies.

Feldman, L. B., & Siok, W. W. T. (1997). The role of component function in visual recognition of Chinese characters. *Journal of Experimental Psychology: Learning Memory & Cognition, 23,* 776-781.

Feldman, L. B., & Siok, W. W. T. (1999). Radicals contribute to the visual identification of Chinese characters. *Journal of Memory & Language, 40,* 599-576.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception & Performance, 13,* 104-115.

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

Hoosian, R. (1992). Psychological reality of the word in Chinese. In H. C. Chen & O. J. L. Tzeng (Eds.), *Language processing in Chinese* (pp. 111-130). Amsterdam: North-Holland.

Hsieh , C. C., Chuang, D. M., Chang, T. L., & Hsu, W. J. (1995). Design and application of a Chinese character database. *Proceedings of the 6th National Conference on the Study*

of Written Chinese Characters (pp. 9-22), Taipei, Taiwan: National Chung Hsing University.

Hsiao, J. H. W., Shillcock, R., & Lee, C. Y. (2007). Neural correlates of foveal splitting in reading: Evidence from an ERP study of Chinese character recognition. *Neuropsychologia, 45*, 1280-1292.

Hsiao, J. H. W., Shillcock, R., & Lavidor, M. (2006). A TMS examination of radical combinability effects in Chinese character recognition. *Brain Research, 1078*, 159-167.

Hue, C. W. (1992). Recognition processing in character naming. In H. C. Chen & O. J. L. Tzeng (Eds.), *Language processing in Chinese* (pp. 93-107). Amsterdam: North-Holland.

Hue, C.W. (2003). Number of characters a college student knows. *Journal of Chinese Linguistics, 31*, 300-339.

Jared, D., McRae, K., & Seidenberg, M. S. (1990). The basis of consistency effect in word naming. *Journal of Memory and Cognition, 29,* 687-715.

Kuo, W. J., Yeh, T. C., Lee, C. Y., Wu, Y. T., Chou, C. C., Ho, L. T., Hung, D. L., Tzeng, O. J. L., & Hsieh, J. C. (2003). Frequency effects of Chinese character processing in the brain: An event-related fMRI study. *Neuroimage, 18*, 720-730.

Leck, K. J., Weekes, B. S., & Chen, M. J. (1995). Visual and Phonological Pathways to the

Lexicon: Evidence from Chinese Readers. *Memory & Cognition, 23,* 468-476.

Lee, C. Y., Tsai, J. L., Su, E. C. I., Tzeng, O. J. L., & Hung, D. L. (2005). Consistency, regularity, and frequency effects in naming Chinese characters. *Language & Linguistics, 6,* 75-107.

Lee, C. Y., Tsai, J. L. ,Kuo, W. J. ,Yeh, T. C. ,Wu, Y. T. ,Ho, L. T. ,Hung, D. L., Tzeng, O. J. L., & Hsieh, J. C. (2004). Neuronal correlates of consistency and frequency effects on Chinese character naming: An event-related MRI study, *Neuroimage, 23,* 1235-1245.

Liu, C. L., Hue, C. W., Chen, C. C., Chuang, K. H., Liang, K. C. Wang, Y. H., Wu, C. W., & Chen, J. H. (2006). Dissociated roles of the middle frontal gyri in the processing of Chinese characters. *Neuroreport, 17,* 1397-1401.

Lo, M., Hue, C. W., & Tsai, F. Z. (2007). Chinese readers' knowledge of how Chinese orthography represents phonology. *Chinese Journal of Psychology, 49,* 315-334.

National Institution for Compilation and Translation. (1995). *Elementary School Chinese Textbooks*. Taipei, Taiwan: National Institution for Compilation and Translation.

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 & Cognition, 24,* 101-118.

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

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

Taft, M., & Zhu, X. P. (1997). Submorphemic processing in reading Chinese. *Journal of Experimental Psychology: Learning Memory & Cognition, 23,* 761-775.

Tan, L. H., Feng, C. M., Fox, P. T., & Gao, J. H. (2001). An fMRI study with written Chinese. *Neuroreport, 12,* 83-88.

Wang, M. Y. (2002). The nature of character-component interaction in Chinese character perception. *Psychologia, 45,* 162-175.

Wang, M. Y. (2006). Examining the bias for orthographic components using an apparent motion detection task. *Psychologia, 49,* 193-213.

Weekes, B., & Zhang, B. Y. (1999). Chinese character recognition in the left and right visual fields. *Brain & Cognition, 40,* 269-272.

Weekes, B. S., Chen, M. J., & Lin, Y. B. (1998). Differential effects of phonological priming on Chinese character recognition. *Reading & Writing, 10,* 201-222.

Author Note

We wish to thank the two anonymous reviewers of this article for their useful comments. This research was supported by a grant (NSC93-2413-H002-019) to Chih-Wei Hue from the National Science Council of the Republic of China. The software and the user's manual of C-CAT can be retrieved from http://140.112.62.230/C-CAT/. Correspondence concerning this article should be addressed to Chih-Wei Hue, Department of Psychology, National Taiwan University, Taipei, 106, Taiwan. Email: hue@ntu.edu.tw

Footnotes

[1] Traditional Chinese characters are used in Taiwan, Hong Kong, and Macau.

In the mainland China and Singapore, simplified characters are used.

Table 1

The properties of Chinese characters manipulated in a number of studies using characters as experimental materials.

Variable	Study
Linguistic frequency	Ding et al., 2004; Hue, 1992, 2003; Kuo et al., 2003; Lee et al., 2004; Lee, et al, 2005; Liu et al., 1996; Lo et al., 2007; Seidenberg, 1985; Taft & Zhu, 1997; Wang, 2002, 2006.
Character complexity	Chen et al., 1996; Perfetti & Tan, 1998; Seidenberg, 1985; Taft & Zhu, 1997; Tan et al., 2001; Wang, 2002, 2006; Weekes et al., 1998.
Type of characters	Fang et al., 1986; Hue, 1992; Leck et al., 1995; Wang, 2002, 2006; Weekes & Zhang, 1999; Weekes et al., 1998.
Orthographic, semantic, and phonological relationship in pairs of characters	Leck et al., 1995; Liu, et al., 2006; Perfetti & Tan, 1998; Weekes et al., 1998.
Position of radical or phonetic component in character	Chen et al., 1996; Ding et al., 2004; Feldman & Siok, 1997; Hisao et al., 2007; Lo et al., 2007; Wang, 2002, 2006.
Character regularity	Hue, 1992; Lee et al., 2005; Seidenberg, 1985; Shu et al., 2000; Tan et al., 2001.
Character consistency	Fang et al., 1986; Hue, 1992; Lee et al., 2004; Lee et al., 2005. Lo et al., 2007.
Neighborhood size	Feldman & Siok, 1997, 1999; Hsiao et al., 2006; Taft & Zhu, 1997.

Table 2

The number of total characters contained and the number of different characters used in the seven character corpuses stored in C-CAT.

Character corpus	Total number of characters contained in the corpus	Number of different characters used in the corpus (Character stock)
CKIPG	20,698,116	5,656
NICT 1st-6th-grade	80,651	2,687
NICT 1st-5th-grade	60,331	2,306
NICT 1st-4th-grade	40,448	1,814
NICT 1st-3rd-grade	22,983	1,322
NICT 1st-2nd-grade	10,617	896
NICT 1st-grade	2,775	399

Table 3

An example of using C-CAT to analyze a Chinese character.

	User's input
Character to be analyzed	see Fig. 1p
Character stock	NICT 1st-6th-grade
Character/component phonological relationship	Same Vowel
Between-character phonological relationship	Same (excluding tone)
	C-CAT's output
Corpus: NICT 1st-6th-grade	
Corpus size: 80651	
Number of different characters used in the corpus: 2687	
Frequency	5
Frequency rank	1442
Number of strokes	9
Structure	left-right
Component	see Fig. 1q
Regularity	yes
Neighborhood size	110
Component	see Fig. 1r
Regularity	yes
Neighborhood size	11
Characters fit the phonological condition specified	"bitter" (Fig. 1v1, /ku3/, 45, 356)*, "cry" (Fig. 1v2, /ku1/, 27, 508), "warehouse" (Fig. 1v3, /ku4/, 8, 1150), "brutal" (Fig. 1v4, /ku4/, 1, 2641), "pants" (Fig. 1v5, /ku4/, 1, 2605)

* The two numbers represent the frequency and frequency rank of the character.

Table 4

An example of using C-CAT to select characters.

	User's input
Character stock	CKIPG
Frequency rank	1 to 1000
Structure	Left-right
Number of strokes	8 to 12
Character/component phonological relationship	Same (excluding tone)
	C-CAT's output
Corpus: CKIPG	
Corpus size: 20698116	
Number of different characters used in the corpus: 5656	
Character	see Fig. 1s*
Frequency	3980
Frequency rank	732
Number of strokes	9
Component	see Fig. 1r
Regularity	yes
Neighborhood size	19

*C-CAT will print out all the characters that meet the linguistic properties specified by the user. The table presents only one of them as an example.

Table 5

An example of using C-CAT to analyze a character component.

	User's input
Component to be analyzed	see Fig. 1r
Character stock	NICT 1st-6th-grade
Character/component phonological relationship	Same Vowel
Position of the component	Any
	C-CAT's output
Corpus: NICT 1st-6th-grade	
Corpus size: 80651	
Number of different characters used in the corpus: 2687	
Phonetic validity	0.8
Neighborhood size	11
Neighbors	"old" (Fig. 1w1 , /gu4/, 64*), "bitter" (Fig. 1w2, /ku3/,45), "an aunt" (Fig. 1w3, /gu1/, 24), "solid" (Fig. 1w4, /gu4/, 8), "why" (Fig. 1w5, /hu2/, 2), "dried up" (Fig. 1w6, /ku1/, 5), "to estimate" (Fig. 1w7, /gu1/, 1), "to sell" (Fig. 1w8, /gu4/, 1), "guilt" (Fig. 1w9, /gu1/, 1), "to subdue" (Fig. 1w10, /ke4/, 11), "to live at" (Fig. 1w11, /jiu1/, 20)

*Character frequency.

Table 6

An example of using C-CAT to select character components.

	User's input
Character stock	CKIPG
Number of strokes	5 to 10
Character/component phonological relationship	Same Consonant
Phonetic validity	0.0 to 1.0
Neighborhood size	5 to 15
	C-CAT's output
Corpus: CKIPG	
Corpus size: 20698116	
Number of different characters used in the corpus: 5656	
Component	see Fig. 1t*
Number of strokes	5
Phonetic validity	0.6
Neighborhood size	5
Neighbors	"words" (Fig. 1x1, /ci2/, 1701†),
	"to feed" (Fig. 1x2 , /si4/, 525),
	"to serve" (Fig. 1x3, /si4/, 181),
	"an ancestral shrine" (Fig. 1x4, /ci2/, 81),
	"a descendant" (Fig. 1x5, /si4/, 76)

*C-CAT will print out all the components that meet the linguistic properties specified by the user. The table presents only one of them as an example.

†Character frequency.

Figure Caption

Figure 1. Sample characters used in the present article.

Figure 2. The parameter setting windows corresponding to the four C-CAT functions.

Figure 1

Figure 2