Original Paper *Forma*, **15**, 409–414, 2000

Complexity of Chinese Characters

Tatsuo TOGAWA[1], Kimio OTSUKA[1], Shizuo HIKI[2] and Hiroko KITAOKA[3]

[1]*Institute of Biomaterials and Bioeng., Tokyo Medical and Dental University, 2-3-10 Kanda Surugadai, Chiyoda-ku, Tokyo 101-0062, Japan*
[2]*School of Human Sciences, Waseda University, Tokorozawa, 2-579-15 Mitsugashima, Tokorozawa-shi 359-1192, Japan*
[3]*Biomedical Research Center, Osaka University Medical School, 2-2 Yamadaoka, Suita-shi 565-0871, Japan*
E-mail: togawa@inst.i-mde.tmd.ac.jp, hiki@human.waseda.ac.jp, kitaokah@image.med.osaka-u.ac.jp

(Received May 1, 2000; Accepted June 13, 2000)

Keywords: Chinese Character, Logogram, Complexity, Stroke Number, Coding

Abstract. Chinese characters are logograms each of which corresponds to one or only a few ideas. To express many ideas, many characters have to be used. Thus, in order to distinguish different characters, great complexity of form is necessary. If Chinese characters had evolved so as to maximize efficiency in conveying ideas, character complexity would depend on the necessary number of characters. Therefore we examined the distributions of stroke numbers of Chinese characters taught from the first to sixth grade of Japanese elementary school, Chinese characters for normal daily use, and all the Chinese characters listed in large dictionaries. A fairly linear relationship was found between the logarithm of the number of characters and the average number of strokes in each groups ($r = 0.998$). This result suggests that the morphology of the Chinese character system is well organized so as to distinguish given number of ideas with minimum complexity.

1. Introduction

In human culture, many kinds of symbols have been used to distinguish specific objects or ideas from others. For example, emblems or family crests have played important roles in securing the individual identities of discrete social units. Such symbols have evolved over long periods, making the form of each symbol simple enough for daily use, and correctly distinguishable from others. It would be expected that the forms of symbols become simple when only a few objects need to be distinguished, whereas their complexity would have to increase when the number of objects is higher. Chinese characters are logograms, each character being a symbol representing a specific idea, and these have evolved over a long period of history to form a large-scale system of symbols.

Using many different characters, many ideas can be represented in principle by a character even though two or more characters are often combined to compose an idiom. In Japanese, phonograms are also used in combination with Chinese characters so that major terms are represented by the Chinese characters and other terms such as particles to identify

case and tense are represented by phonograms.

The simplest Chinese character can be written as one stroke whereas the most complicated character needs more than 30 strokes as seen in Fig. 1. It is known that fewer strokes are required in frequently used characters and that the number increases as the frequency of use decreases (KAWAI, 1970). In the Japanese school system, children are taught Chinese characters systematically. According to the guidelines provided by the Ministry of Education, Science, and Culture, Japan, Chinese characters are taught from the first grade of elementary school, so that children know about 1,000 characters by the end of 6th grade. However, this level is still insufficient for ordinary intellectual activities such as to reading a newspaper. About 1,900 characters are designated for daily use, and always accessible in most Japanese computer operating systems. However, more than 8,000 characters are listed in large dictionaries.

Because Chinese characters evolved over many thousands of years in China, and more than a thousand years have passed since their introduction into the Japanese language, it is likely that Chinese characters and their usage had evolved so as to maximize efficiency in representing necessary information. In fact, it is known that Chinese characters representing fundamental ideas are always simpler than those representing more sophisticated concepts, and thus Chinese characters with smaller number of strokes are used more frequently than those with larger number of strokes (MIYAZIMA, 1978). However, more detailed quantitative analysis of stroke numbers from a theoretical viewpoint is still lacking. It is considered that during progression through elementary school, the necessary number of characters increases to accommodate the increased complexity of the information which is necessary for children to handle.

If each of the strokes making up a Chinese character contributes to distinguishing the conveyed information independently, each additional stroke multiplies the possible number of objects represented by the characters a constant factor corresponding to the conveyed information capacity of a stroke, like an additional bit multiplies the possible number of representation a factor 2 in the binary code, and thus, the possible number of objects will increase exponentially with stroke number. Therefore, it is suspected that the average stroke number at each grade of elementary school will be proportional to the logarithm of the cumulative number of characters to be taught until the end of that grade. Therefore, we attempted to examine whether or not the average number of strokes is consistently related to the information processing capacity at different levels.

Fig. 1. Examples of Chinese characters with different stroke numbers. The meaning of each character is shown in parenthesis.

2. Method

The 1,006 Chinese characters that need to be taught from the 1st to the 6th grade listed in the *Guideline for Elementary School* (MINISTRY OF EDUCATION, SCIENCE, AND CULTURE, JAPAN, 1977), the 1,945 Chinese characters designated for daily use (MINISTRY OF EDUCATION, SCIENCE, AND CULTURE, JAPAN, 1981), and the total of 8,160 Chinese characters listed in a standard dictionary (KANJIGEN, 1993) were classified according to stroke numbers, and the average stroke numbers as well as the standard deviations were plotted against the logarithm of the number of characters used at these levels.

3. Result

Figure 2 shows the distributions of average stroke numbers of cumulative groups of Chinese characters taught up to different grades of elementary school, as well as Chinese characters designated for daily use and Chinese characters listed in the dictionary. For example, the 5th group means all Chinese characters taught from 1st to 5th grades. In Fig. 3, average stroke numbers and standard deviations of those groups of Chinese characters are plotted against the number of characters in a logarithmic scale. The regression line between the logarithm of the number of characters, $\log_{10} x$, and average stroke number, y, is

$$y = 3.74 \log_{10} x - 1.88 \quad (1)$$

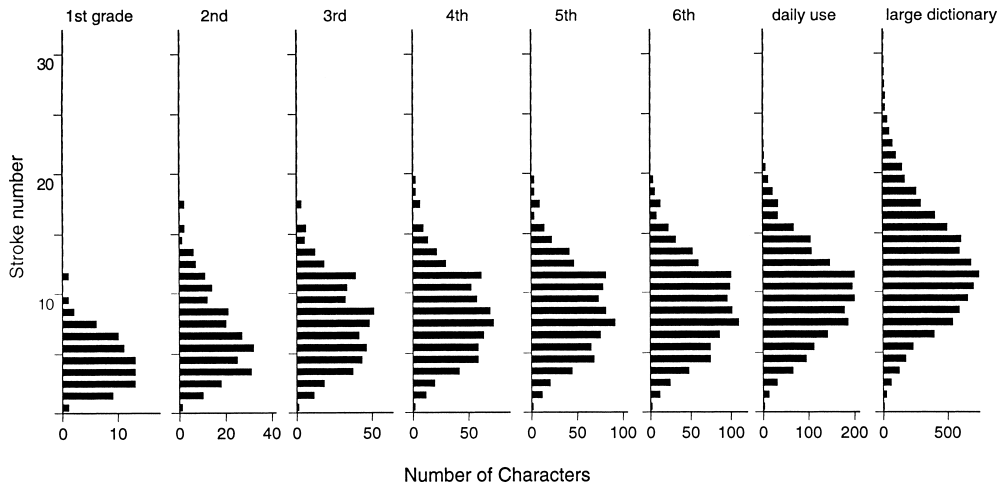

Fig. 2. Distributions of stroke numbers of cumulative groups of Chinese characters. For example, the 5th group means the characters taught from 1st to 5th grades. Distributions of stroke numbers of Chinese characters designated for daily use and the characters listed in a large dictionary are also shown.

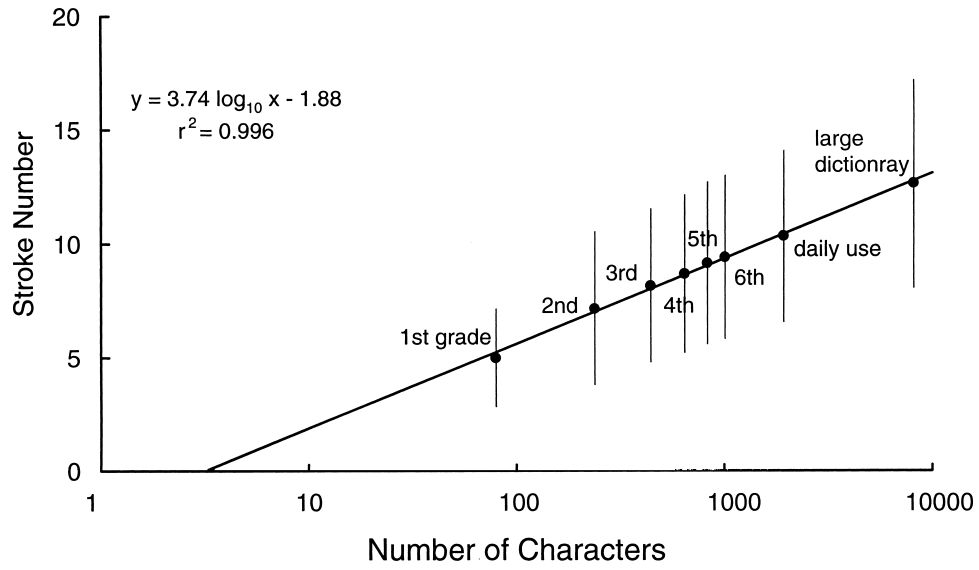

Fig. 3. Relationship between average stroke numbers of Chinese characters and number of Chinese characters at different cumulative groups from 1st to 6th grades in elementary schools as well as number of Chinese characters designated for daily use and the characters listed in a large dictionary on a logarithmic scale.

with correlation coefficient of 0.998. If

$$x = ca^y \tag{2}$$

then a = 1.85 which is close to 2. This means that each stroke in each Chinese character conveys almost one bit.

4. Discussion

The above result demonstrates that simple characters are always taught earlier and complex characters later. It is known that Chinese characters with smaller number of strokes are used more frequently than those with larger number of stroke, as mentioned before. Therefore, the result suggests that frequently used characters are taught earlier and rarely used characters later. This results is reasonable in the sense that frequently-used fundamental concepts represented by simple characters are taught earlier, and that complex ideas are introduced and corresponding characters are taught according to the development of mental activities.

The fact that the average stroke number increases in proportion to the logarithm of the number of characters to be used suggests that the Chinese character system is an efficient coding system like optimal binary or decimal coding in which the number of bits or digits increases in proportion to the logarithm of the number of objects to be coded.

Each stroke in a complex Chinese character does not always contribute independently to distinguishing that character from others, but many characters are composed of two or more components which are also Chinese characters having fewer strokes. If a set of characters consists of those with stroke number y, which are composed of two components having stroke number y_1 and y_2 so that $y_1 + y_2 = y$, and if each component can be regarded as a simpler Chinese character, then, from Eq. (2), the number of components having stroke number y_1 and y_2 will be $x_1 = ca^{y_1}$, $x_2 = ca^{y_2}$. Even though all possible combinations of these two sets of components can not compose characters, a small fraction, say k (<1), are able to do so. Then, the total number of characters of stroke number y composed of two components having stroke number y_1 and y_2 will be

$$x = kx_1x_2 = kc^2a^{y_1+y_2} = kc^2a^y \tag{3}$$

which is similar to Eq. (2).

The logarithmic relationship between code complexity and the number of objects encoded is not a feature specific to Chinese characters but can be applied to coding systems in general. The logarithmic relationship between the complexity and the number of objects suggests an underlying efficiency of the coding.

In European languages, each word consists of many characters while each character is quite simple. Such a character system will be efficient if the average word length is proportional to the logarithm of the total number of words. This argument suggests that both character systems had evolved so as to maximize the efficiency of their utility, while Chinese and European languages employ different character systems.

Coding is used not only in artificial systems but also in a natural one. In fact, we have previously suggested that a relation between the form and the coding system is recognized in the morphology of the dendritic structure of cortical neurons and the size of the cerebral cortex (TOGAWA and OTSUKA, 1999). The total length of dendrites, which corresponds to the complexity of the form of each neuron, seems to increase with an increase in the number of cortical neurons, which in turn is proportional to the surface area of the cortex. Although the representational scheme for information in the cortex is still unknown, it is likely that large brain will process more information than a smaller one. As long as all information is encoded in the cortex, the coding scheme used by a larger brain will inevitably become more complex. We have pointed out previously that the total length of dendrites of each pyramidal cell in animals with different evolution levels increases roughly in proportion to the logarithm of the cortical surface area which in turn correspond to the total number of neurons in the cortex (TOGAWA and OTSUKA, 1999).

We have also proposed a model of the cortical neural network in which the single-cell-representation hypothesis was introduced, and shown that the number of synapses on each neuron is proportional to the logarithm of the total number of neurons in the cortex (TOGAWA and OTSUKA, 1998, 2000). The single-cell-representation hypothesis has been seriously criticized and most investigators have abandoned it (ROLLS and TREVES, 1998). However, asking how many neurons are needed to represent an idea in the brain is similar to asking how many characters are needed to represent an idea in a language, and the answer to the latter question is that it depends on the language system. The fact that total length of the dendrites of each neuron increases in the evolution stage suggests that the neural

representation scheme is similar to the scheme in the Chinese or Japanese language, in which many ideas can be represented by a single Chinese character while the complexity of each character increases inevitably.

5. Conclusions

It was found that the averaged stroke numbers of cumulative group of Chinese characters taught at every grade in elementary schools, as well as the characters for daily use and the total characters listed in large dictionaries, are proportional to the logarithm of the cumulative number of characters at each level. This suggests that Chinese characters can be regarded as an efficient coding system that has evolved over a long period, giving each character a form that enables it to distinguish the necessary number of ideas with minimal complexity.

REFERENCES

KAWAI, Y. (1970) Paradoxical relations between reading and spelling of Chinese characters, *Mathematical Linguistics*, **No. 55**, 23–29 (in Japanese).

MIYAZIMA, T. (1978) The new style Chinese characters and stroke numbers, *Mathematical Linguistics*, **11**, 301–306 (in Japanese).

MINISTRY OF EDUCATION, SCIENCE, AND CULTURE ed., (1977) *Guideline for Teaching in Elementary School* (in Japanese).

MINISTRY OF EDUCATION, SCIENCE, AND CULTURE ed., (1981) *Table of Chinese Characters Designated for Daily Use* (in Japanese).

ROLLS, E. T. and TREVES, A. (1998) *Neural Networks and Brain Function*, Oxford Univ. Press, New York.

TODO, A., MATSUMOTO, A. and MATSUMOTO. A. eds., (1993) *Kanjigen*, Kenkyusha.

TOGAWA, T. and OTSUKA, K. (1998) A model of cortical neural network structure, *Proc. IEEE EMB 20th Annual Int. Conf.*, 2066–2069.

TOGAWA, T. and OTSUKA, K. (1999) Simulation of dendritic structure of pyramidal cells in the cerebral cortex, *Proc. IEEE EMB 21st Annual Int. Conf.*, 402.

TOGAWA, T. and OTSUKA, K. (2000) A model of cortical neural network structure, *Biocybernetics and Biomedical Engineering*, **20** (in press).