HUMAN FACTORS AND LINGUISTIC CONSIDERATIONS: KEYS TO HIGH-SPEED CHINESE CHARACTER INPUT

Paul L. King[*]
Cornell University
Ithaca, NY, USA

Abstract

With a keyboard and supporting system developed at Cornell University, input methods used to identify ideographs are adaptations of well-known schemes; innovation is in the addition of automatic machine selection of ambiguously identified characters.

The unique feature of the Cornell design is that a certain amount of intelligence has been built into the machine. This allows an operator to take advantage of the fact that about 60% of Chinese characters in text are paired with other characters to form two-syllable compounds or phrase words. In speech and writing these pairings eliminate about 95% of the ambiguities created by ambiguously identified syllables.

Introduction

For Chinese character input to computers, a Cornell research team has approached the problem from the point of view of the type of person who would most likely be operating a Chinese electronic typewriter, namely, a commercial Chinese typist with junior middle school education who would regularly be typing for eight to ten hours per day. For such a person, a word processing system should be easy to learn, fast (averaging 50 characters/minute), and capable of being used for several hours without inducing a high level of fatigue.

Of the many input systems that have been proposed in recent years, one based on Wang Yun-yu's four-corner numbering system[1] has best demonstrated, in the opinion of the Cornell team, the capability of meeting the criteria of ease of learning, speed, and low operator fatigue level. For example, operator training is simplified because keystrokes are assigned by "inspection" rather than rote memory.

Also, frequently used simplex characters such as particles are given unique identifiers so they can be inserted in text without going through a manual disambiguation process. Even more significantly, manual disambiguation has been eliminated entirely in nineteen cases out of twenty by attention to linguistic affinities of characters.

Shape Code

The four-corner system is a simple encoding scheme that native Chinese speakers learn in about half an hour. In this system, the peripheral forms of all Chinese characters are projected onto ten basic stroke shape classes (小, 一, 口, 扌, 十, 丿, 丶, 乛, 八, 亠), to which single-digit values are assigned. On the basis that ideograms are basically square in appearance, four-digit numbers can be read from stroke shapes at the corners, in the sequence top left, top right, bottom left, bottom right. Thus the shape classes describing a character such as 說 are 亠, 八, 口, 一.

I have made two refinements on Wang's original four-corner system. First of all, the Cornell code includes elimination of all four-corner null zeros, so that identifiers for characters vary from one to four digits. Thus, the identifier for the Chinese character "一" is simply "一", with the three null zeros eliminated.

As applied to the Cornell input system, middle-of-word ("comma") and end-of-word ("print") delimiter keys make possible the use of variable-length input codes for identification of Chinese characters. If null zeros were retained so that all characters were uniformly identified by four-digit numbers, there would be no need for either "enter" or delimit keys. However, new flexibility comes through the use of delimiters.

Specifically, there are three advantages to the Cornell code:

1) The variation of identifier size increases the number of identifier categories, thus somewhat reducing ambiguities.

2) An operator does not need to mentally add null zeros in order to read an identifier from an ideogram; he identifies only the shapes that are there.

3) On the average it takes fewer

key strokes to type the identifier for a given ideogram.

On this keyboard, all keys are taught in direct correspondence to stroke shapes, thus eliminating the need for operators to do any intermediate encoding into numbers. They do, however, use the number values of the keys in disambiguation. At any given instant the keyboard is thus being treated either as a collection of stroke shape class identifiers or as a collection of digits, but not both together.

The second refinement that the Cornell team has made in the four-corner system is in redefining the number values of some of the shapes in accordance with human factors considerations of keyboard design[2]. Figure 1 is a schematic representation of the new shape identifier keyboard.

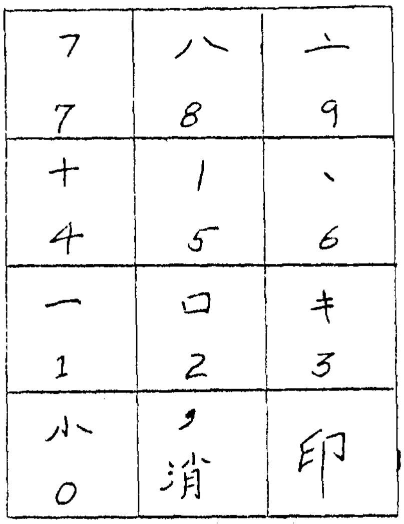


Figure 1

Placement of stroke keys is determined by shape association, frequency of use, and positions in characters.

"一" and "八" are the Chinese numbers 1 and 8, and "7" looks like a number 7, so those three shapes are placed on their respective number keys. Although no shape-number associations are taught to typists, placing the above shapes on their associated number keys is an attempt to forestall potential interference across modes.

The most frequently used shapes, "十", "丿", and "丶", are placed on the middle row of keys, so that an operator does not need to move his fingers from the normal rest position in order to depress those keys.

The remaining four shape identifiers are the least frequently used of the ten. They are placed on the keyboard in the approximate positions where they usually appear in Chinese ideograms.

Disambiguation

Lack of speed has, until recently, been a major drawback of shape identifiers for Chinese data entry. In a set of 8,000 characters, Wang Yunyu's code uniquely identifies only 11% of characters.

Thus, if the four-corner method is used to identify a simplex character, there is a two-step process that can be invoked in order to isolate the desired character. The key to speed in typing is to automate this process as often as possible, which is what we have done.

The non-automatic aspect is simple. The first step involves looking up all characters described by a particular code. In an electronic word processor, all the characters meeting a particular shape description are displayed on the CRT. The operator then implements the second step which consists of picking the desired character out of the displayed list for insertion in text.

In the Cornell device, manual two-step disambiguation is a straightforward process. If the identifier points to only one character, that character is inserted in text. In the event ambiguity remains after initial entry, the machine gives the operator an audible cue and displays the complete ambiguous list. Then the operator can make his choice by typing a number indicating which listed character he wants, followed by the "PRINT" command. That character is printed, and typing continues with the next entry.

However, the unique feature of the Cornell design is that a certain amount of intelligence has been built into the machine. This allows an operator to take advantage of the fact that about 60% of Chinese characters in text are paired with other characters to form two-syllable compounds or phrase words. In speech and writing these pairings eliminate about 95% of the ambiguities created by homophonous syllables.

The Cornell input scheme capitalizes on this characteristic of the language by allowing the operator to type in identifiers for two paired characters in sequence. A stored dictionary of pairings eliminates most ambiguities that arise in searching for simplex characters.

For example, Cornell code 一一 identifies such characters as 王，至，工，and 亞. However, the character 作 (Cornell code 丿八丿一) occurs in a pairing only with the last character in the above list. Thus, if an operator enters " 一一 , 丿八丿一 PRINT", using "," as a nonfinal character delimiter and "PRINT" as a final delimiter, the unique pair 工作 is retrieved immediately from the machine without need for any manual disambiguation.

In the event ambiguous pairs are still encountered, they are disambiguated manually in the manner first described. In any case, input speed is greatly increased through use of the dictionary.

An editor is constantly accessible as part of the system, so that changes can be made to any part of the text that is being typed at any time.

Development and Application

The Cornell design is meeting the criteria envisioned for commercial operators. Chinese test subjects require only half an hour of instruction to learn the shape keyboard of Figure 1 and the means of disambiguating. Thereafter, with about 80 hours of practice, typists are achieving speeds of more than 40 characters/ minute with typing error rates approaching 0 (See Figure 2). Moreover, over a period of a year, test subjects have maintained a high level of motivation with frequent long hours at the keyboard.

It is anticipated that mean typing speeds of 50-60 characters/ minute for uncorrected text will be achieved with further machine development and operator training.

Current areas of development are 1) implementation of simplified character sets in the machine along with associated shape identifiers, 2) isolation of specialized vocabulary into specified sets, and 3) continued testing of Chinese operators in the field.

Future development and applications include the following:

1) The 12-key keyboard can be implemented via a touch-tone telephone, enabling any touch-tone telephone to be used as a remote terminal for the electronic word processor.

2) Various output applications can be developed, including printer/ plotter, typesetting, and direct telegraph transmission.

3) Implementation of a chord keyboard can be studied.

Summary

In sum, the Cornell electronic word processor for Chinese has added machine disambiguation to an old input idea. By using a "friendly interface," we have enabled the machine to accept ambiguous input codes (four-corner shape identifiers) and use its limited intelligence to provide the desired output in a one-step process. With this system, the learning process is greatly simplified for Chinese typists, rapid typing speeds are achieved within a short period of time, and operator fatigue is kept at a low level.

The work on which this paper was based received support from the NCR Corporation.

References

[1] Herring, J. A., The Foursquare Dictionary, Taipei: Mei Ya Publications, Inc., 1969, pp. 6-12.

[2] Meguire, Patrick G., Human Factors Scientist, NCR Corp., Personal communication to author, 15 February 1979.

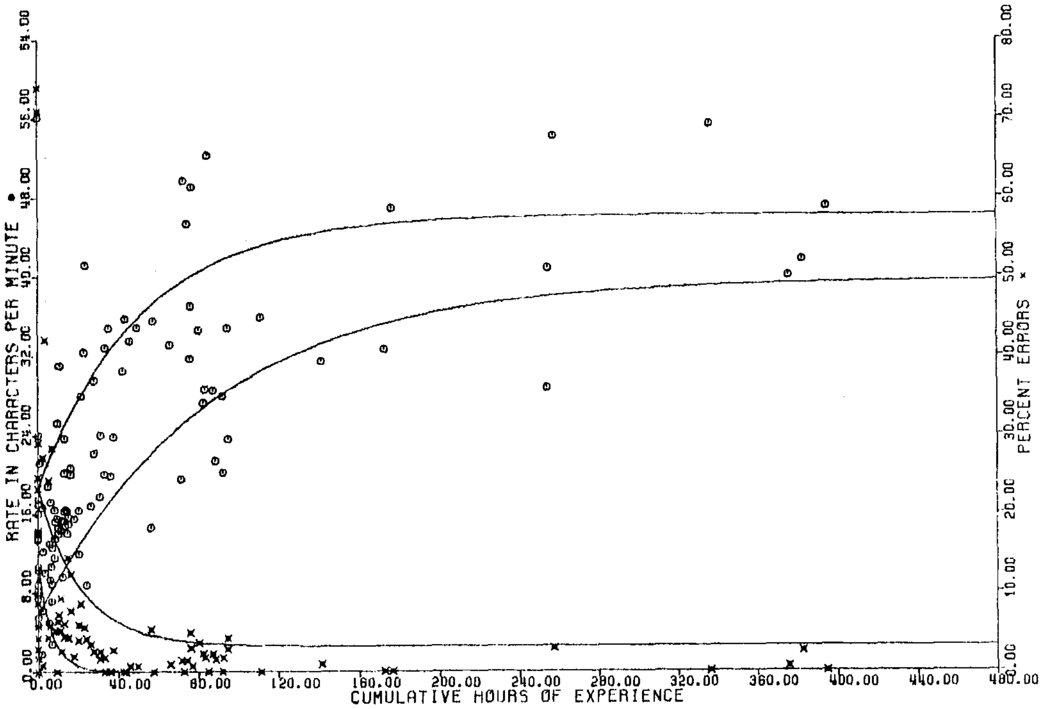
Bibliography

Kiang, Te-yao, "A Natural Way of Analysis of the Ideograms and its Application to Chinese Language Input System," Proceedings of International Computer Symposium 1977, p. 322.

Proceedings of International Computer Symposium 1977, Vol. 1, Taipei: National Taiwan University.

Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems, Taipei: Academia Sinica, 1973.

Yu, Wellington Chia-pier, "An Input Encoding Scheme for Chinese Characters," Proceedings of International Computer Symposium 1977.



NUMBER OF SUBJECTS = 30

TYPING RATE:

NUMBER OF TESTS = 109
TIME CONSTANT = 68. ± 17. HOURS.
ASYMPTOTE = 43. ± 3. CHAR/MIN
SCATTER = 9. CHAR/MIN

31 May 1980

ERROR RATE:

NUMBER OF TESTS = 83
TIME CONSTANT = 14. ± 7. HOURS
ASYMPTOTE = 1. ± 2. PER CENT
SCATTER = 11. PER CENT

Figure 2