Character Entry Efficiency for the Chinese Language

Jay Lee

Follow this and additional works at: https://scholarsarchive.byu.edu/jeal

BYU ScholarsArchive Citation
Available at: https://scholarsarchive.byu.edu/jeal/vol1987/iss82/7
I. The Chinese Language and Electronic Data Processing

The Chinese language is indisputably one of the most complicated languages in the world because of the tremendous number of characters and the "unscientific" ideographic structures. Still, a computer can process and manipulate these elements without much difficulty—provided they are entered first. Data input devices, therefore, become an especially critical path for the full utilization of the computer's capability in the Chinese language automation.

Even though several state-of-the-art input devices, e.g. vocal, optical, and artificial intelligence (AI), have been researched and experimented with for some time, they have not reached the stage of economical and effective application for the daily needs of electronic data processing. Even for English-language data processing, the QWERTY keyboard remains the most popular and cost-effective device. To date, this holds true not only for English, but even more so for the Chinese language. This paper discusses the keyboard character entry device (input method) for Chinese characters called the Tsang-chieh input method.

II. Chinese Character Input Methods

Chinese character input methods can be efficient, and in fact, are already adequately efficient. My experience suggests that they are relatively easy to learn for the user who is familiar with English typing. The barrier for Chinese language automation has been more a psychological, and less a technological, issue. For those familiar with English touch typing, acquiring Chinese input skill takes no more than half the effort it takes a beginner to acquire English typing skill. (This estimate is hypothetical; readers are encouraged to conduct research to verify this estimate. The actual effort could be much less.)

Nevertheless, Chinese character input methods sound "user-hostile" when encountered. Confusing user-friendliness with efficiency is a mistake made by many potential users. This common mistake has been the major reason preventing many from objectively assessing the method; and, therefore, they may refrain from learning at all. The following analogy can demonstrates that user-friendliness (simplicity) and efficiency are two different, if related, elements. Handwriting is very easy for literate adults, but typewriting can be much more efficient though it requires serious initial training. An ideal design would combine these two elements. In most cases, however, there tends to be constraints on each design. This is particularly true for Chinese language input methods.

Because of the structure of the language, keyboard input methods for Chinese probably can never be as efficient as English typing. Nonetheless, a reasonable
degree of efficiency (hypothetically, 50% to 75% of one's English typing speed) can easily be achieved. I believe the same can be said about Japanese kana inputting since it has a small character set to deal with.

III. Learning the Tsang-ch'ieh Input Method

The Tsang-ch'ieh input method has been adopted by many major Chinese automation systems on the market.

A certain three operators who input (type) Chinese characters as professional users have been tracked at inputting sixty to eighty characters per minute. Two of them use the graphic-based Tsang-ch'ieh input method (TC), while the other uses the pronunciation-based pin-yin input method. They manipulate these "user-hostile" designs with a high degree of proficiency. Learning these input methods is a modest challenge to anybody with average manual dexterity. A typist with a skill of thirty words a minute can memorize himself the eighty-nine graphic patterns of TC in three hours and familiarize himself with the code-forming rules (漢字法) in three two-hour practice sessions. The following paragraphs give a simple mnemonic association technique for learning the patterns.

The mnemonic associations for learning the first seven patterns, represented by the English letters A through G, sort out the first two patterns (ㄖ and 甲) from the rest. These two symbols can be memorized in only seconds.

The mnemonic associations for 鈀 is C because it is romanized as chin according to the Wade-Giles transliteration; and D is for 木 since a tree is deep-rooted; E for 水 because in French, water is spelled as eau. (An inputter who does not know French must create his own association.) There is no difficulty in memorizing 火 and 土 since in English, the former is "fire" and the latter, "ground". By creating these associations one can memorize them in a few minutes. With minimal practice one can touch-type the seven patterns in a few minutes.

金 木 日 土 火 水

Using the same approach, one can memorize all twenty-four primary patterns in about thirty minutes, and all eighty-nine patterns in less than three hours. Creating random lists of the patterns for practice allows one to practice the combinations with both hands.

The more patterns memorized, the easier it becomes, there being more elements to create associations with. For instance, once ㄖ, 甲, and 鈀 are memorized ㄖ, 甲, and 鈀 can be picked up easily because of their visual similarity or etymological relationship. Once one remembers 火 (an easy task since fire elements are associated with the F key), 木 and 金 can be memorized in no time due to the visual similarity. Using visual, auditory, and etymological associations, the following patterns come easily:

1. 水 = ㄑ.

2. 甲 = ㄈ. And since ㄌ is often used as a variant form of ㄈ, the former is easy to retain.
3. Though perhaps arcane, \( \text{月} \) can be memorized as \( \text{田} \) without the two horizontal internal strokes. Once \( \text{月} \) is associated with the B key, \( \text{日} \) can be easily remembered as a shorter version of \( \text{月} \).

4. \( \text{¥} = \text{木} \). Because the latter is the former without the two side strokes.

5. \( \text{士} = \text{±} \) and \( \text{匕} = \text{匕} \).

6. \( \text{忄} = \text{小} = \text{扌} = \text{P} \). Because of the Greek prefix "psycho-".

Looking at the following random patterns on the left one can now type their English codes on the right as follows.

金木日土月火水心士手言忄戈月刀門戶七夕

Some of these associations are rather "universal" and can be employed by anybody who has the language knowledge. For instance, F represents \( \text{火} \) because of FIRE; T, \( \text{ト} \), because of the visual similarity. Others of them are rather esoteric; for instance, E representing \( \text{水} \) because of the French word eau, and P representing \( \text{心} \) because of its Greek etymological root "psycho-", as in "psychology". Users may wish to create their own associations that work equally well.

There are several patterns for which it is difficult to create mnemonic associations. If one blends one or two of them into each set of patterns, they can be memorized easily. By expanding one's random list with each set of new patterns, the first set of patterns becomes 金木日土月火水心士手言忄戈月刀門戶七夕. With the second set, it became 金木日土月火水心士手言忄戈月刀門戶七夕. The revision and expansion of these patterns reinforces memory retention.

The more time-consuming part of learning the TC input method is practicing to familiarize oneself with the code-forming rules. After three two-hour sessions of practice one can match 50% of one's English typing speed. It was very easy to select characters for practice using input code dictionaries supplied by vendors. Some vendors even supply sample characters under each rule for practice.

After these self-training sessions, one may practice the TC input method with a focus on the code-forming rules. Some of the unnatural codes repeat themselves frequently enough that they soon become familiar by virtue of repetition. For instance, \( \text{応} \) is encoded as a combination of \( \text{応} \) and \( \text{忇} \); \( \text{応} \), \( \text{忇} \), and \( \text{忇} \). Therefore, the code for \( \text{応} \) is PD, while that for \( \text{忇} \) is TT.

One drawback of the Tsang-chieh input method is a constraint imposed by the structure of the guidelines. About ten to twenty percent of the characters do not fit the code-forming rules unequivocally. For instance, one could input \( \text{生} \) as OG with O representing \( \text{月} \) and G, \( \text{±} \). It is, however, encoded in the dictionary, as HQM representing \( \text{火} \), \( \text{扌} \), and \( \text{戈} \). Since one code-forming guideline dictates that each letter should represent as many strokes as possible when more than one composition scheme is possible, one may, therefore, contend that \( \text{生} \) should be encoded as OG because O represent two strokes, \( \text{月} \), while H only represents one stroke, \( \text{戈} \). In this case OG would clearly be preferable to HQM, since the former has only two letters while the latter, three.
Another constraint is that calligraphic and font type variations may cause misinterpretation. For instance, жд calls for the letter I as the initial code while м calls for H, probably due to different calligraphical preferences. Fortunately, common characters repeat themselves so frequently that in a short time the user becomes familiar with these preferences.

These drawbacks can be lessened due to multiple input methods. Since vendors provide more than one input method, including the user-friendly (if slower) pronunciation input methods, characters with troublesome codes can be easily accommodated. If two attempts to input a character fail, leave a blank space for this character and move on to the next character; then use the pin-yin or Wade-Giles romanization to input these characters while proof-reading the text rather than attempting more codes.

IV. Observations on the Tsang-chieh Input Method

1. The character-by-character input method based on the graphics composition scheme is the most efficient input method for the Chinese language, (and other languages associated with it), because it makes touch-typing possible. Several other input methods are available on the market, e.g. a mouse or script conversion software. These are user-friendly but not as efficient.

2. In contrast to initial impressions, the TC input method is easy to learn.

3. The Tsang-chieh is very efficient. For users who are familiar with English typing, it is possible to achieve the same typing speed with only half of the learning effort. This observation is borne out by the fact that the key positions of the QWERTY keyboard was intentionally arranged to slow down typing speed. Early typewriters were not efficient in design; as a result, many typists would outperform the machine and jam the keys. Since the QWERTY keyboard was arranged to create awkward key positions, the more convenient Tsang-chieh design compensates for some deficiencies of TC imposed by the Chinese language.

4. The advantage in acquiring TC inputting skill is great. Considering the computer's capabilities in word processing, file manipulation, database management, and type setting, achieving fifty percent of one's English typing efficiency, (an easy goal to achieve) is a very high payoff.

V. Other Input Devices

For users who still find it difficult to commit themselves for two or three weeks to learn TC, there will be alternatives: optical or voice recognition, or artificial intelligence. They are, however, still far from as efficient and cost-effective as a keyboard.

The scanner, an optical reader, can be used to input large quantities of printed data very fast and accurately. It is, however, already limited by the number of font types that it can recognize. The enormous number of the Chinese characters and their calligraphic variations present even more limitations. The greatest flaw in scanning devices is that they are of course useless for documents yet to be composed.
Voice recognition has its serious drawbacks too. First, can one mouth words faster than ten fingers can type them? Second, how can users read fast and accurately without fatigue? Third, how will homonyms, accents, and dialectical variations affect accuracy? Finally, how can we deal with noise in the office when more than one person is inputting text at the same time?

Whether AI can be perfected at all is still being debated; and costs are unknown.

At present, and for some time to come, these alternatives are not viable ones, but the keyboard input method remains for us. Though it requires a commitment in learning, it is not difficult to master.

I am interested in giving training workshops on the Tsang-chieh input method if there is sufficient interest. If you are interested, please send a note to:

Jay Lee, Marketing Specialist
OCLC
6565 Frantz Rd.
Dublin, Ohio 43017-0702