John DeFrancis

How Efficient is the Chinese Writing System?

To help resolve disagreement regarding the relative efficiency of the Chinese system of writing, it is useful to take a close look at some of its specific applications. A good starting point is the arrangement of characters in dictionaries and the lookup procedures involved in locating entries. A closely related matter is composing text and reproducing it, processes which include typesetting, typewriting and digital composition. Composing text brings up the peculiarly difficult problem of segmenting text, which is rendered all the more acute by lack of agreement on how to standardize the orthography of the Pinyin alphabetic system that is acquiring new importance as an adjunct to handling characters on computers. Reformers increasingly emphasize the need for a policy of digraphia, the coexistence of two writing systems, Pinyin and the traditional characters, each to be used in the areas to which it is best suited. This trend throws further light on the efficiency of Chinese characters by bringing to the fore how they relate to reading and writing and where they fit into the classification of writing systems.

John DeFrancis, after a long and productive career as a Chinese language scholar teaching at notable American universities, continues to contribute to knowledge in this field with his recent publication of In the Footsteps of Jenghis Khan *(University of Hawaii Press, 1993). His next major publication, expected in spring 1996, is a compilation for the* ABC Dictionary *(short for Alphabetically Based Computerized Chinese-English Dictionary) on which he has been working with colleagues in the United States and China.*

Center for Chinese Studies
University of Hawaii at Manoa

Honolulu
Hawaii 96822

Visible Language, 30:1
John DeFrancis,
6-44

Rhode Island
School of Design

Providence
Rhode Island 02903

Chinese characters occupy the least amount of space. Their sound, shape and meaning are all concentrated within a square; the characters can be read and understood at a glance. This makes them distinctively superior to other writing systems in efficiency for reading and comprehension.

Li Youren 1991:69.[1]

CHARACTERS 'EASIER THAN ABC TO READ'

Headline in *China Daily 11/15/1984.*[2]

Many Chinese continue to hold opinions similar to those quoted above despite reformist efforts, now over one hundred years old, to advance a contrary view of the Chinese character system as excessively difficult, hopelessly inefficient and urgently needing to be supplemented, if not replaced, by an alphabetic system of writing. These disparate views touch upon everything from the actual use of characters in everyday life to the academic question of their classification among the writing systems of the world.

A decade and a half ago a leading Chinese linguist noted that attempts to determine optimal orthography for languages have generally suffered from a dearth of scientific investigation (Wang, 1981:234). Although some progress has been made since then, it remains true that the pertinent research that might help determine optimal orthography and facilitate comparisons between writing systems remains scanty. Nevertheless, it may still be possible to say something useful about the subject, particularly if we eschew facile generalizations and speculations in favor of the more arduous task of examining some concrete instances of graphic usage. It is only, I believe, by actually working through the details (here much simplified) of some specific applications that one can begin to acquire a firm basis for evaluating the efficiency of the Chinese script.

We can begin with Li's assertion that "the characters can be read and understood at a glance." Let us test this assertion by inviting some visiting Martians, equally unfamiliar with Chinese characters or latin scripts such as German and English, to glance at the expressions 欢迎 and *Willkomen!* and *Welcome!* Such a test would no doubt provide irrefutable proof that graphic symbols cannot be read and understood at a glance unless one has learned them beforehand or has gained familiarity with the system to which they belong.

1
Li Youren teaches at Dezhou Teacher' College in Shandong province.

2
The headline is based on an article by Zeng Xingchu of the Psychology Department of East China Normal University in Shanghai (Zeng Xingchu, 1983–1984).

Of course Li assumed possession of such familiarity, a blithe assumption that dodges the issue of efficiency by neglecting to consider how much time and effort is needed to acquire the knowledge, or, more broadly stated, what it takes to learn to read and write any script.

We are confronted here with the cat's–cradle problem of literacy. It appears hopeless to think that scientific tests can be devised which would make possible overall comparison across scripts of something that it has proved impossible even to define, much less measure, with any significant degree of agreement. We cannot do much better than to cite the frequently expressed belief that it takes at least two years more for Chinese to become literate in their script than it does for Westerners to achieve an equivalent level of literacy in theirs.

Although most Chinese would probably agree that their characters are indeed more difficult, without necessarily concurring in the precise degree of difference, there are others whose views are reflected in the headline quoted above reporting the claim of a Chinese psychologist that "children aged 2 to 4 can easily learn 3,000 characters." Advocates of writing reform have sharply criticized as scientifically unacceptable the methods used in arriving at such claims and they reject this and similar claims of simplicity as chauvinistic twaddle (Zheng, 1983; DeFrancis, 1989, 120–121).

In further pursuit of a common–sense evaluation of some specific applications of writing in everyday life, let us assume a basic knowledge of how the Chinese and Western systems work and then apply this knowledge to looking up some terms, among them those we presented to our Martian visitors, in dictionaries compiled for native speakers of the three languages. Although perhaps not entirely certain how the German and English words are to be pronounced (does *come* in *Welcome* rhyme with *some* or *dome*?), by mechanically following the sequence of letters we succeed, in a matter of seconds, in locating the terms and finding how they are pronounced and what they mean.

The spelling of the Chinese characters is an even less reliable guide to their actual pronunciation than in the case of the notorious English orthography, especially since the mid–fifties, when the People's Republic of China (PRC) introduced simplified forms for about a third of the older complicated characters. For example, if we look at the two characters that comprise the expression 欢迎, namely 欢 and 迎, we note that 欢 used to be written as 歡 (and still is in Taiwan and Hong Kong). The phonetic component 雚 located on the left hand side of character 歡 spells *guàn,* whereas the whole character 歡 has the pronunciation *huān.* The phonetic component 卬 located at the top right of the second character (i.e., 迎) in 欢迎 represents the sound *áng,* which has the same final as the sound *yíng* of the whole character 迎. So the spelling *guàn'áng* suggested by the two pre–simplification phonetic elements 雚 and 卬 actually represents the pronunciation *huānyíng* for the expression 欢迎, a disparity comparable to that between *psalm* and its actual pronunciation. However, simplification of the old phonetic element 雚 to the new element 又, which has the unhelpful pronunciation *yòu,* has eliminated the similarity that used to exist between *guàn:huān* for 雚 and 歡.

The Dictionary Problem

The introduction of simplified characters has complicated the problem of dictionary lookups. We first have to decide whether a character dates from before or after the PRC reform. If it is a simplified character, such as the first character 欢 in expression 欢迎, we have to consult a mainland dictionary or a conventional dictionary produced elsewhere that has a conversion table. If it is an original non–simplified form of a simplified character, such as 歡, we have to look it up in a Taiwan or other non–PRC dictionary, or check with a PRC dictionary that provides a conversion. If it is a non–simplified form of a traditional character, such as both characters in 激光, namely 激 and 光, we can look it up in any dictionary. It is often not readily apparent whether a particular character is a simplified or non–simplified graph, in which case we may have to try elsewhere after an initial failure to locate it.

It turns out, however, that even if we knew at first glance, or learned by consulting a dictionary, that the characters in 欢迎 in both their simplified and unsimplified form combine to make a term that is pronounced and transcribed as *huānyíng,* this knowledge would not enable us to emulate the German and English lookup procedures, for the Chinese have not produced a single dictionary in which the entries are arranged in simple alphabetic order. Instead they have contrived a host of other schemes all based on the shape of the characters.

Some dictionaries arrange the characters by total number of strokes. Most dictionaries based on some other arrangement include a stroke–order index, if not for all entries then at least for characters not easily found by other arrangements — generally about ten percent. The stroke–order arrangement involves a tedious counting of the number of strokes in the character (in the case of 歡, we count 1, 2, 3,...22 strokes) and checking in the general neighborhood of the supposed number in the numerous instances where the exact number is not clear; one can therefore easily make a mistake in arriving at the right total. As a result this is not a popular approach, but it remains useful, especially as a last resort when other approaches fail for one reason or another to lead to the item being searched.

The most popular arrangement is the so–called 'radical system' or 'radical–stroke system' initiated by the first Chinese dictionary, of the second century A.D., which arranged its 9,353 characters under 540 semantic keys, popularly but misleadingly called radicals, such as water, insect, vegetation. The number was later reduced to 214, which remained the standard until the PRC introduction of simplified forms for some of the traditional graphs. Since then PRC lexicography has been in a state of almost complete chaos, with dictionary–makers going their own way by variously arranging the characters under 186, 187, 188, 191, 201, 225, 226, 227, 242 and 250 keys. This is as if OED, Webster, and other dictionaries disagreed as to whether to retain the old A to Z order or change it by, say, placing the digraph *ph* after *f* rather than with *p* and grouping all words beginning with the *in* of direction separate from all those beginning with the *in* of negation and from another group containing *indigo* and *inkle.*

Leaving detailed differences aside, all Chinese dictionaries based on this arrangement order the keys by the number of strokes they contain; the number ranges from 1 to 17. Traditionally the keys with the same number of strokes used to be arranged in a fixed order, so that the 'water radical,' for example, was always the 85th out of 214. Dictionaries aimed at foreign users attached the appropriate sequential number to each key. Native dictionaries, however, like ours based on the familiar alphabetic sequence, usually counted on the user's ability to locate an item by its relative position; modern dictionaries are increasingly adding the sequential numbers.

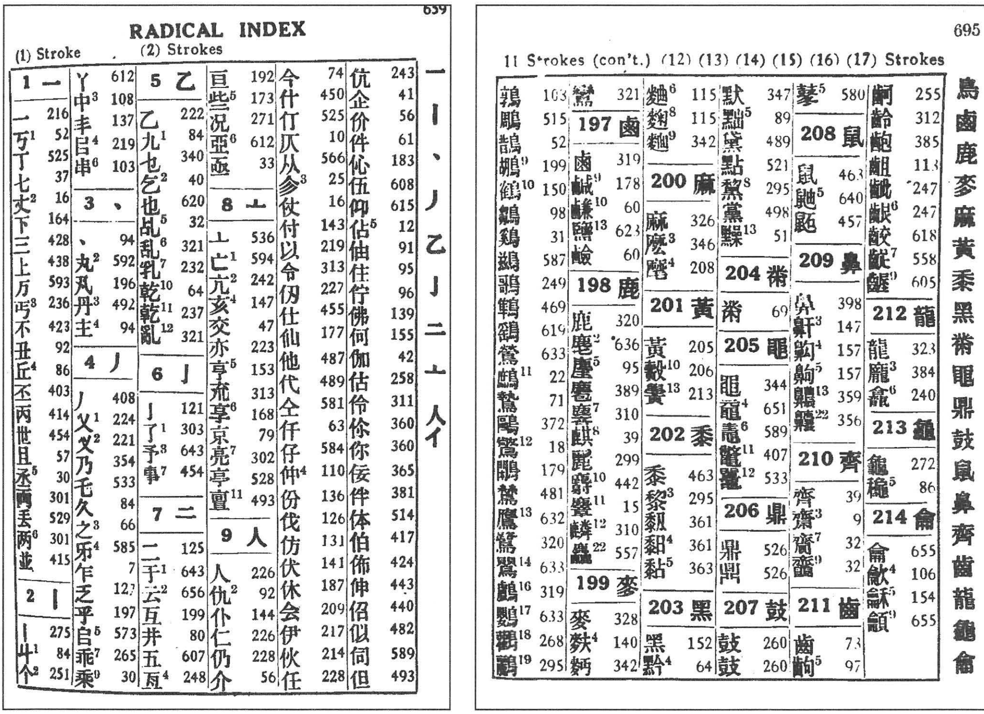

Figure 1

Radical Index of Traditional Characters

As it is often not easy to determine what part of a character might be considered its key, and hence might be used as the basis for filing it, most dictionaries use some other arrangement to lead one to these doubtful characters and note under what radical they will be found. In one such dictionary almost ten percent of its 7,773 entries are listed by the order of their total number of strokes in an appendix entitled "List of Characters having Obscure Radicals" (Mathews, 1945, 1222–1226).

The layout of this important radical–stroke method of ordering Chinese characters can be seen in the accompanying Radical Index of Traditional Characters (*figure 1*), which presents the first and last page of the radical index for 5,000 characters contained in a popular small Chinese–English dictionary (Fenn, 1926). The raised number after some characters refers to the number of additional strokes in the character apart from the key. The order of characters with the same number of residual strokes is not fixed, so that one has to run through all those grouped together before one can determine whether or not the dictionary being consulted has the desired character. The number after each character refers to the page where its pronunciation and definition may be found — in some dictionaries, to a specific identifying number for the character.

With the PRC babelization of dictionary arrangement, the hapless user of a Chinese dictionary must now exert even greater effort to determine which component of a character is likely to be taken by a dictionary–maker as the key under which to file the character. He must then track down where that key happens to occur in the particular order favored by the lexicographer, and finally must locate the character among those that share the same key by counting the number of additional strokes apart from the key and singling out the desired character from the random sequence of those that have the same number of residual strokes.

If we use a radical–stroke dictionary to look up, for example, the expression 激光, we are likely to guess, correctly, that the first character 激 will be filed under the left–hand key 氵 (conventionally number 85 "water") rather than the right–hand key 攵 (number 66 "to tap"). Next we count the number of additional strokes apart from the key and locate the character among the one or two dozen other characters with thirteen residual strokes under the "water" key. We are then referred to a page where we find the pronunciation rendered as *jī* in the Pinyin system and the definition "to surge; to stir up." (Note that the rest of the character apart from the water key, i.e., 敫, is the phonetic *jī*, which exactly represents, even as to tone, the pronunciation of the full character 激.)

The second character in 激光, namely 光, presents more of a problem and well illustrates some of the complications that have been introduced into the writing system by the PRC changes. As a common unsimplified character it can be found in any dictionary, all of which give its pronunciation as *guāng* and its meaning as "light." In conventional works, character 光 is to be found under 儿, which is key number 10 in the traditional list of 214; it is a variant of 亻, key number 9; both of these keys are given the pronunciation *rén* and are glossed as meaning "man." As a result of the PRC changes 儿 came to be used as the simplified form of 兒 and to acquire its pronunciation *ér* and its meaning "child." It also acquired a new place in PRC dictionaries: number 21 in one list of 189 keys, number 29 in another list of 227.

There is another fairly popular approach to ordering characters known as the Four Corner System. The creator of this system assigned a number from 0 to 9 to the different kinds of strokes that happen to occur in each of the four corners of a character, from upper left to upper right, then down to lower left and lower right corner. The resultant 4–digit number often did not uniquely identify a character, however, so that a fifth digit was added later to help resolve the ambiguity. But even this often does not suffice to solve all the difficulties (Coia, 1985, 81).

There are several versions of the system, including new ones that take account of the changes from traditional to simplified forms. In one such dictionary, a Chinese–Chinese work, after making what is likely to be several attempts to guess the right numbers, we will locate the first character in 激光, namely 激, under the number 3814 that is shared by 11 other characters. For the pronunciation and definition of the character we are referred to page 509, where we find the pronunciation *jī* and a definition that confirms the meaning we had already found in our Chinese–English dictionary (*Xiandai,* 1980).

There is still another method of serial arrangement of characters that is worth mention because of its use in an important dictionary that will be discussed later. That is the *Standard Telegraphic Code* or *S.T.C.* arrangement. In this adaptation of the "radical" ordering, Chinese telegraphy assigns a four–digit number to each of just under 10,000 characters. Referring to the code–book whenever necessary, that is for all but the numbers of frequently–occurring characters that they have memorized, dispatchers send the numbers corresponding to the characters of the message; receivers at the other end reverse the process. Thus the telegraphic code number for 激, the first character in 激光, is 3423 (*Miming*, 1946,34).

Quite a few other ways of ordering characters have been devised by lexicographers who think they have found simpler and more efficient solutions. These idiosyncratic approaches invariably turn out to be difficult to use because what may seem obvious to the compiler is seldom equally obvious to the user. That is a pity, because some of these dictionaries have considerable merit.

After we have consulted one or more dictionaries based on the approaches discussed above and have succeeded in finding that the first character in 激光 has the pronunciation *jī* , we are in a position to search for the English equivalent of the term *jīguāng* that transcribes the two characters whose respective meanings we have found to be “surge light.”

The simplest solution would be to find the term under the letter *j* in an alphabetically arranged dictionary. Here we must distinguish between two quite different kinds of dictionaries that are frequently referred to as alphabetically arranged. One is the fully alphabetic “one–sort dictionary” that arranges all entries in letter–by–letter sequence, so that we have the sequence

lìngwài in addition
línjū neighbor

with *ng* preceding *nj* despite the fact that the breakdown of the two entries by syllables corresponding to the characters is

lín jū
lìng wài

In the other kind of ostensibly “alphabetically arranged” dictionary that is actually only semi–alphabetic, the primary arrangement includes only the transcription corresponding to the head syllable, so that *lín* and its sub–entry *línjū* come before *líng* and its sub–entry *lìngwài*:

lín
—
—
línjū
—
—
lìng
—
—
lìngwài
—
—

All the character–combinations containing *lín* as head entry are grouped together, here also in alphabetic order, as is true also of *lìng* and its sub–entries. As a result the two entries are likely to be separated from each other by several pages, instead of only a few rows. What is even more serious, if one starts with a spoken expression or its Pinyin transcription, is that one must know what character is used to represent the first syllable of a term before one can even begin to look it up.

This last objection can be seen even more clearly in the case of the expressions

jīlǐ
jīlì
jìlǐ

Instead of being able to locate these items close to each other, as would be possible with a strictly alphabetic arrangement, one would have to know that the initial character of *jīlǐ* is 机, of *jīlì* 激, and of *jìlǐ* 祭 in order to track them down as much as thirteen pages apart in one well–known dictionary (Wu Jingrong, 1979, 308–321), where we find the entries

jīlǐ 机理 mechanism [on page 308]
—
—
jīlì 激励 encourage [on page 312]
—
—
jìlǐ 祭礼 memorial ceremony [on page 321]

Nevertheless, among the general–use bilingual dictionaries, this combination of sound–based and shape–based order is the most favored system. It is used to some extent in Chinese–Chinese dictionaries too, though the failure of many Chinese to retain command of the alphabetic system they learned in the first grade inhibits their use of even this compromise system.

There is no general–use dictionary, either Chinese–Chinese or Chinese–English, based on the completely alphabetic ordering of the entries. However, there are three special Chinese–English dictionaries based on this arrangement. Two, both aimed at beginning students, have only a small number of entries, so of course our search there for *jīguāng* results in expected failure (Simon, 1947; Wang, 1966). The other is a mammoth work that is surrounded by much mystery.

Modern Chinese–English General and Technical Dictionary was published by McGraw–Hill in 1963 without the usual preface indicating who produced the work, and also without informing us of a matter of more than usual interest in this case, that is, precisely how it was done; the introduction merely states that it "was prepared for the greater part by semi–automatic mechanical processes." I suspect that the anonymous compiler was the CIA.

Be that as it may, the work consists of three tomes each measuring 7.5 x 10.5 inches. Volume 3, the most useful one, has 1884 pages (including a supplement of 56 pages inadvertently omitted in the original printing) and weighs an unwieldy six pounds. Here are two sample entries:

BIDANWANG 443 6699 1734 4853
grenade net

BIDANYI 441 6699 1734 5902
bullet proof vests

The first group of numbers refers to the tones of the preceding term of three syllables, for we obviously have here an electronic printout that lacks the capacity to produce tonal diacritics for the normal transcriptions of *bìdànwǎng* and *bìdànyī*. The second group consisting of three 4–digit numbers refers to the *Standard Telegraphic Code* numbers corresponding to the three preceding syllables of the transcription. These numbers provide the means to look up each individual character in Volume 1, a smaller work of 152 pages that consists chiefly of a "Radical–Stroke Index to Characters" presenting character, code number and transcription. Of less utility for the general user is Volume 2, another tome about the size of Volume 1 that presents the material in the form

3423 *ji* 1 stir up....

It takes only a few seconds, less time than for any other arrangement, to look up in Volume 3 the sequence of letters that spell *jīguāng* in the position where the term should occur. Alas, it is not there. Following the second–fastest procedure, we turn to the *jī* transcription in a recently published semi–alphabetic dictionary, hunt among the almost three dozen *jī* characters to locate the one glossed "surge," run our eyes down the alphabetically arranged sub–entries, and find the entry *jīguāng* followed by 激光 and the gloss "laser," a term that apparently entered the language too late for inclusion in the 1963 McGraw–Hill dictionary.

Of course if one is confronted with an expression written only in characters, it is not possible to look it up in an alphabetically arranged dictionary unless one knows the pronunciation of the characters used to write the term. One might attempt to guess the pronunciation from the phonetic elements, which are often helpful enough to warrant an initial try at a sound–based lookup. Otherwise one must consult a radical–stroke dictionary or one based on some other approach to the shape and composition of the characters.

But most users of dictionaries, especially most Western users, have more occasion to look up terms containing characters they already know than those with characters they don't know. In a formerly popular dictionary containing 7,773 characters, a mere fifteen percent or 1,200 characters, those of highest frequency, entered into as many as seventy percent of the 104,000 multiple–character entries (DeFrancis, 1966, 13). Knowing the individual characters does not guarantee knowledge of the compounds they enter into, any more than knowing the meaning of *sweet* and *bread* guarantees knowing the meaning of *sweetbread.*

Some idea of the relative efficiency of the various dictionary arrangements can be grasped from the foregoing discussion. It is also attested by the results of dictionary lookup tests administered by Professor Victor Mair to speakers of Chinese, Japanese and Korean, all languages that still use characters in varying degrees. He states: "My personal experience and experiments with my students and colleagues have demonstrated that words can normally be found two to ten times faster in a single–sort alphabetically arranged list than in other types of arrangements" (Mair, 1986, 18–19).

The differences in efficiency that characterize the diverse schemes used in dictionaries of course carry over into other applications of serial arrangements. Telephone directories are generally based on the radical–stroke arrangement. Library catalogues also generally follow this arrangement, so that looking up a succession of characters may require several stroke–order lookups before the whole entry can be accessed or otherwise manipulated. A Chinese librarian remarked several decades ago:

> *In the absence of an alphabet, there is no system of filing Chinese characters which admits of their ready location. Those generally in use are time–consuming. They place an undue burden on the memory, and allow too many chances for error (Li, 1940, 10).*

As for indexes, these are so difficult to make and to use that most books solve the problem by simply omitting them altogether. Their absence here and in other areas is a serious lack. As one student of the subject remarks:

> *Indexing is perhaps the most common and economically important example of large–scale list making. Without indexes to serial publications, abstracts of reports and other*

synoptic aids, a modern researcher simply cannot cope with the flood of information in his or her field (Unger, 1987:55).

The maintenance of office files likewise presents enormous difficulties. Hence, as has been said of Japan, memory of the individual employees still plays a much greater role than in the West, so that "using human brains as filing cabinets is accepted as normal" (Unger, 1987, 56–57).

The Typographic Composition Problem

Among the problem areas deserving our particular attention is that of composing text and reproducing it. In this area China was for centuries far in advance of the West. At a time when Europe was still laboriously producing a limited number of copies of books by the inefficient method of copying them out by hand, China had print runs of thousands of copies for works printed first from woodblocks, invented in the seventh century, and later from moveable type, invented in the eleventh century. But with the spread of printing to the West, the simple European alphabetic scripts were able to make more efficient use of the Chinese invention than was possible by the Chinese with their more complicated character system.

Although the contrast between the two could be drawn in various ways, perhaps it can be made more vivid if I detail my own direct involvement in the printing process for both Chinese and English.

In 1936 Professor George A. Kennedy prepared to initiate the Chinese language program at Yale University by acquiring a font of Chinese type from Commercial Press in Shanghai. I helped him set up a miniature replica of a Chinese printing establishment in the basement of the graduate school building.

We filled a large room measuring about 30' x 40' by closely crowding together a number of inverted V–shaped frames measuring about 5' in height and 6' in width. The sloping sides of the frames were arranged to hold several rows of open–faced trays measuring about 8" x 10." Each tray was divided into a number of compartments containing pieces of type of various sizes ranging from 10.5 to 21 points. In the basic size of 10.5 points there were 10,000 different characters. The number of pieces for each character varied — 150 for the most common characters, at least 24 for the less common, and 5 for the rare (DeFrancis, 1940).

The small number of duplicate pieces meant that composition was limited to only a few pages at a time. When the type ran out, it was necessary to print what had been set up (or make electrotype plates in anticipation of later printing), break down the type, return the pieces to their proper location and then compose another section of text.

A printshop in China would have to do the same unless it could afford to keep as many pieces of each character as it might be called upon to use for works that might contain hundreds of thousands of characters of running text. Or it could cast new type as needed, including when these wore out from overuse, by using monotype molds.

The characters in our printshop, like those in China, were arranged in the order of a radical–stroke dictionary. In my role as compositor I attached a piece of handwritten text to a composing stick held in my left hand and then, following the text character by character, searched out the corresponding pieces of type much as one would look up characters in a dictionary. In an ordinary day of typesetting this involved several miles of walking in and out among the racks, from one end of the room to the other, bending down and reaching up in search of successive pieces of type.

My perambulatory stint as compositor of Chinese contrasts with my brief sedentary experience in handsetting a small amount of type in English. For that I sat with my composing stick and text before a single rack of trays each measuring about 2' x 3' and containing type that consisted of some hundred symbols (the twenty-six letters in upper and lower case, punctuation marks, numerals, etc.) as against the 10,000 of the Chinese font. The bulk of English composition was of course done by an experienced linotype operator using equipment that, prior to the advent of expensive photocomposing machines, could not be adapted to the huge number of Chinese characters, and later only by the usual clumsy method of dictionary lookups.

That inefficient traditional arrangement of characters was also used in Chinese typewriters. These consisted of a large bed of characters measuring about 2 1/2' x 3' surmounted by a movable carriage containing roller and paper. The bed contained about 4,000 pieces of type standing upright in both radical and stroke–order arrangement. Operators ("typists" doesn't seem quite appropriate) moved the carriage about with a handle in a horizontal plane over the bed in order to position a hammer with a square hole over each desired character. Then they pressed the hammer, which triggered a pin that pushed the character up into the square hole and also caused it to strike against the paper.

The total font of characters was 10,000. Apart from those in the bed the remaining 6000 were contained in boxes measuring about 8" x 10" each containing some 1500 of these less–used characters. There were a few empty spaces in the grid of basic characters in which operators could place rarer graphs whenever these were encountered. If there were too many of them, the operator first made a place for them in the bed by removing some of the characters already there, located the rarer characters in their separate boxes, and placed them in the vacated locations. Similar procedures were invoked to use characters of different fonts. There were usually two or three more beside the basic font, for such things as the equivalent of italics or boldface, but all of the same size, for the grid could only accommodate the generally used size of about 10.5 points.

Trained operators were able to type twenty to thirty characters per minute. Casual users could manage only two or three. If we consider that a character generally represents less than a Chinese word and roughly equate the latter with an English word, the comparison with typing speeds in English is derisory, for a teacher of typing informs me that the speed of a two–finger amateur is around forty-five words per minute, of a productive officeworker at least sixty words per minute, and of record–makers one-hundred words per minute.

The dismal performance of Chinese typewriters has now become irrelevant as the use of these Rube Goldberg contraptions approaches zero with the advent of computers.

Digital Composition

In the new information age the academic question of efficiency of scripts has assumed critical importance, especially for China in its self–proclaimed attempts at modernization and catching up with the advanced nations of the West. It does not appear, however, that the innovative means for handling data has inspired much innovative thinking about Chinese writing and its adaptation to the new technology. Instead we have the spectacle of Chinese, and foreigners too for that matter, rushing like lemmings along familiar paths that are strewn with the same roadblocks as those we have noted in dictionary lookups.

Over 500 schemes have been advanced in China alone for handling characters on computers (Li Xiuqin, 1990, 61). Worldwide, several dozen schemes of various types have actually entered into production. Not a single one, whether created by Chinese or foreigners, comes anywhere near the efficiency of those for the alphabetic scripts of the West, though some are enough better than the handwriting alternative to win a modicum of acceptance. What they are like can be seen if we look at the schemes in the light of the previous discussion and, in a few of the more important cases, approach them by a sort of hands–on simulation of some of the steps in Chinese wordprocessing.

A number of schemes can be grouped together as based on some variation or other of the Four Corner System. This was one of the earliest approaches to the problem, but also one of the first to be discredited, for at only twenty characters per minute it was even slower than the traditional Chinese typewriter.

There is another group of schemes based on an enlarged keyboard of up to several hundred keys in which each key represented a cluster of characters. Some of these Chinese schemes were similar in approach to the Japanese "Kanji Teletypewriter" adopted by the National Diet Library in which each of the 192 character keys was used to access a display of 14 characters. With one hand the operator accessed the cluster and with the other selected the desired character from the display.

The underlying principle here is similar to the radical–stroke dictionary lookup in which one first goes to a particular cluster (e.g., a group of characters listed under a radical) and then locates a specific character within the cluster (e.g., the third character with six residual strokes under that radical). The Chinese schemes of this sort were even more complicated and cumbersome than the Japanese because of the need for a larger number of characters. The high cost of the oversized keyboards and the slow speed for both professional typists and casual users, who must constantly look at the keyboard and refer to other aids, have prevented this approach from receiving much favor.

The "Ideographic Data Processing System" developed by Wang Laboratories on the basis of a coding technique called the "Three

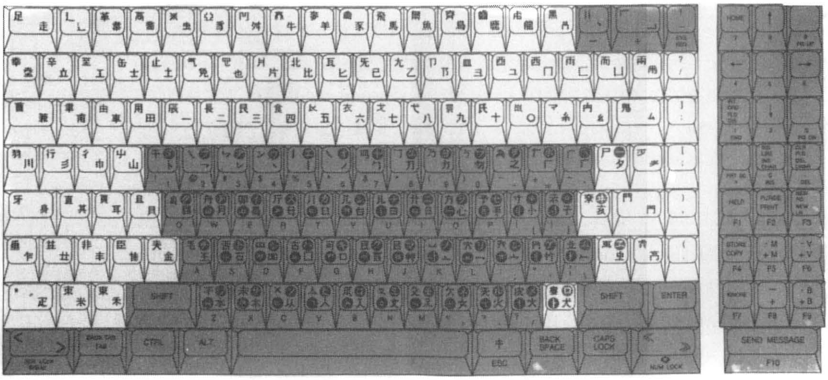

Figure 3

Enlarged Character Keyboard

Corner Coding Method" is representative of a number of schemes that identify characters by a numeric code related to their components. The Wang system analyzes each character into three components each of which is assigned a two–digit numeric code. Using only the numeric keypad, typists enter up to six digits to display on the screen all the characters corresponding to the code, in order of frequency, and then select the appropriate character. Objections that have been made to this approach center on the difficulty of learning a complex coding system. It therefore has little appeal to the casual user and in fact has had little success (Coia, 1986).

A somewhat better reception has been given to a number of schemes based on the idea of composing characters from components displayed on a greatly expanded keyboard. One of the most successful attempts to use this approach is the system developed by Transtech for the Research Libraries Information Network. Some one hundred twenty institutions with library holdings in Chinese, Japanese and Korean belong to this network.

Each of the member libraries has a multiscript workstation costing about $6,000 and pays an annual fee to RLIN. The centerpiece of the workstation is the enlarged CJK keyboard shown above. Leaving aside the simple Japanese *kana* syllabograms (shown in the top right black circles), the even simpler Korean Hangul alphabet (shown in the bottom left black circles), and the usual letters, numbers and assorted symbols, most of the space on the seven main banks of keys is given over to some 250 components of characters, usually two per key. The procedure here involves pressing the right keys, those containing the components of a character, to bring it onto the screen.

Character 欢 is simple enough. Operators will surely guess that its components are 欠 on the M key and 又 on the comma key; by pressing the appropriate keys they can call up character 欢 on the screen or present it in a printout. But 殷 is another matter. It is decomposed in a way that goes counter to how this character is built up and so requires a special effort of memorization. The ordinary sequence of strokes is: ˊ 厂 户 户 户 肙 肙 肞 朌 殷. Yet the Transtech system breaks it down into 白,勹 and 殳. The last is clear enough (it is on the M key), but 白(on the Y key) is derived by an unexpected dissection of the character, as is true also of 勹(on the 7 key), which is an even more jarring distortion of the shape.

If operators fail to guess how the system decomposes a character, they can find out by looking up the character in a radical–stroke index or call it up on the screen, using either the old Wade–Giles transcription or the new Pinyin transcription to type its pronunciation. This evokes a procedure that we noted in the Wang Laboratories system and is used in a great variety of other systems. We can illustrate it with the same character 殷.

This character is transcribed as *yin* without tone, as *yīn* with its proper tone indication. If we type *yin,* without tone, the screen will present seventy-six characters to choose from; if we type *yīn,* with tone, the screen will narrow the presentation to twenty-eight homophones. In both cases each character is followed by its components. We move the cursor to the desired character to single it out for whatever way we want to handle it, for example printing it out, which can be done on either a dot–matrix or laser printer. Ironically, in this system printouts of characters look better than those of transcriptions, as indicated by *Lu Hsün* (in Wade–Giles) appearing as *Lu Hs¨un* (Lu Xun in Pinyin).

Among the shortcomings of the character–decomposition approach just described are the need to memorize a large number of components and their location on the keyboard, to guess how the system decomposes characters and to resort to other lookup procedures in those not infrequent cases where one fails to guess at the sometimes idiosyncratic decompositions. Librarians remark that without constant use they have to keep rememorizing details. Nevertheless, for special usages such as library cataloguing, where Chinese characters do not form a large body of text, the system works so much better than the old alternative of doing things by hand that computerization along these lines undoubtedly represents an advance in efficiency.[3]

3
Professor Victor H. Mair of the University of Pennsylvania informs me (personal communication 3/12/91) that his library and many other member libraries have begun to use a new generation CJK terminal which employs a standard QWERTY keyboard and relies more heavily on phonetic inputting.

For wordprocessing, however, none of the systems described so far, all of which have a shape–based approach to Chinese characters, comes anywhere near matching the efficiency of the sound–based approach that characterizes Western systems of writing. We can go further and state quite categorically that they are all deadend systems having no chance whatsoever of eliminating or even greatly reducing the inefficiency inherent in a character system of writing.

It has been pointed out that "more than ninety percent of all documents in Japan are handwritten, or rather handcrafted" (Becker, 1984). The figure is doubtless even higher in China. In a work of major importance, J. Marshall Unger has made very clear why the Japanese writing system is incompatible with the computer age and why there is no hope of salvation in so–called "artificial intelligence" or other nostrums (Unger, 1987). The same applies with even greater force to the Chinese writing system. One can only hope to mitigate, not resolve, its basic inefficiency.

The greater success achieved by the sound–based approach used in the West, and increasingly even in Japan,[4] eventually led to attempts to apply it to Chinese. Basically this involves using a sound–based system, such as Pinyin, as the input device and developing a program for automatic conversion to characters — in short, Pinyin in, characters out.

Westerners initiated the attempts. The Chinese were slow to follow. As recently as 1986 two Chinese scholars, a linguist and a computer specialist, categorically dismissed the idea by stating that "designing a Chinese input system based on Pinyin or any other phonetic spelling is not practical at al." Their main arguments in support of this conclusion were that the population of China had not achieved sufficient mastery of Standard Mandarin, and that the problem of homophones would be either insurmountable or require excessively cumbersome and costly solutions (Sung and Kuo, 1986).

Mastery of Standard Mandarin does indeed constitute a serious problem in the area of wordprocessing, though not quite in the form that Sung and Kuo seem to suggest, namely mastery of its spoken form. Desirable though that is, it is not indispensable, since what is actually needed is mastery of the Pinyin orthography. Dialect speakers of Mandarin, i.e., the seventy percent of the population that speaks some form of Northern,

4
The preferred input method in Japan appears to be via *romaji*, not *kana* as I incorrectly stated in *Visible Speech* (p. 267). Apparently, though, neither input method solves the problem of handling *kanji* well enough to make computers attractive, for despite their affluence Japanese have not taken to computers anything near as much as Americans, and neither have the Taiwanese despite their economic progress. The *Free China Journal* of September 5, 1991, stated that "Taiwan's installation rate for PCs stands at 47 units per 1,000 people, compared with Japan's 76 and 231 in the U.S."

Southwestern and Southeastern Mandarin, are roughly in the same situation as British, American and Australian speakers of English, all of whom readily handle the same form of written English on computers. Speakers of Cantonese and other non–Mandarin forms of Chinese are in the more difficult situation of Italians and Spaniards having to work in French, or Indians in English. However, as my colleague Robert Cheng points out, continuous handling of Pinyin has the by–product of helping to improve command of Mandarin.

An important difference between the Chinese and English situations is that those who handle the written form of the latter have had a lifetime of exposure to alphabetic writing. Chinese have not, differing in this respect also from the Japanese with their lifetime of dealing with two simple phonetic scripts, one syllabic, the other alphabetic. To be sure, Chinese children learn Pinyin readily enough, becoming literate in it in a matter of weeks, but Chinese educators because of the emphasis on characters allow them to lose command of the system and revert to Pinyin illiteracy in the second grade (DeFrancis, 1984a, 268–269, 283). The failure to retain control of Pinyin, coupled with a lifetime of emphasis on the semantic as against the phonetic aspect of the characters, results in a mind–set that, as Professor Cheng acutely observes (personal communication), makes it psychologically difficult for adults to deal with their language in an alphabetic script. As a result, some Chinese educators, who insist on the need for a policy of digraphia, i.e., literacy in both Pinyin and characters, have given up on all but a few of the less hide–bound adults and are pinning their hopes on at least part of the younger generation growing up with a lasting command of Pinyin thanks to their innovative changes in teaching and promoting the alphabetic script.

Another difficulty in the application of a sound–based approach to the computerization of Chinese stems from the industry concentration on hardware and software that, in the words of a British report on a symposium devoted to French on computers, were "tailored to the needs of the English language" (*The Guardian* 12/8/1985). Subsequent developments have made it possible, albeit sometimes at considerable additional expense, to increase the potential of the hardware to handle not just the problem of French diacritics but even the more complex problem of Chinese characters. This increased potential is due mainly to advances in software. Indeed, it is generally agreed that the problem of computerizing Chinese is largely a software problem.

Efforts to produce the necessary software have been marked by piecemeal attacks on the multifarious difficulties that have to be overcome, so that only scattered bits and pieces of partial solutions have been found for some of the problems. Moreover, as noted in the previously cited article published in 1986, "all the Chinese input methods to date have been designed either by electrical engineers or computer scientists, leaving many linguistic factors out of consideration" (Sung and Kuo, 1986, 48). For their part linguistically oriented scholars have generally lacked the expertise to cope with the technical problems involved. Although there has been some improvement in this situation, there is still no full-scale cooperative effort for a comprehensive solution, notably on the part of the Chinese government, whose leadership has utterly failed to deal with this key aspect of its vaunted modernization drive. As a result, the Chinese computer scene presents the chaotic spectacle of half-baked systems rushing into the market and later trumpeting improvements that should have been made in the first place.

A case in point is the TianMa system, one of the best known and most widely used systems. Its 2.0 version produced in 1986 made the following claims for converting from Pinyin to Hanzi (Chinese characters):

> *With TianMa the text is typed in standard Pinyin Romanization on a normal English language keyboard. Because whole sentences, paragraphs and documents can be entered at once, the speed of input is limited only by the operator's typing speed....*
>
> *With most existing Chinese word processing systems, you type a Pinyin syllable and then select the correct Hanzi from a display of up to 200 characters. You continue to enter Pinyin, one syllable at a time, and selecting the correct Hanzi until the whole document is transcribed.*
>
> *The process is slow and tedious. Even a fast operator can enter no more than 15–20 characters per minute.*
>
> *In contrast, TianMa lets you transcribe at a rate of 100 or more characters per minute (TianMa, 1986).*

To test these claims I thought to start by handling some text from a poem dedicated to the celebrated writer Lu Xun (1886–1986) that is written in everyday spoken style, thus avoiding the problems posed by the Hybrid Vernacular style that dominates present-day Chinese writing. The first stanza of the poem, which appears both in its Pinyin and character versions (Yin and Felley, 1990, 524–528), reads as follows:

Yǒude rén huózhe	*Someone living on*
Tā yǐjing sǐ le;	*May already be dead;*
Yǒude rén sǐ le	*Someone dead*
Tā hái huózhe.	*May still live on.*

TianMa lacked the ability to indicate tones, so I typed the first two lines without tone marks and pressed a key to convert to characters. The screen showed the following characters (my translation on the right):

Someone perhaps	有的人或者
May already be silk;	他已经丝了

The fact that our first test produced gibberish out of a toneless Pinyin text that is perfectly intelligible to the human brain suggests, as later experience confirmed, that the TianMa claim of producing accurate text at the rate of one-hundred characters per minute was merely hype, such as pervades the Chinese as well as general wordprocessing scene.

TianMa made a fundamental error in limiting the discriminatory power of the system by attempting to handle characters without regard to the tones that must be part of their accurate transcription. Handling tones on computers has encountered such difficulties that many systems have been rushed into production, to get on the market ahead of competitors, before the development of acceptable solutions. In the advertisings for its upgraded version called New TianMa, tonal diacritics were promised by June, 1991, but it is only now, in 1995, that the system has acquired the capacity to handle tones. Tone indication had been provided earlier by some other systems, however, such as the Chinese Wordprocessing system of mainland China in the CW.3 version.[5]

Also involved here is a more general problem of tone indication. While the huge volume of publication of Chinese written alphabetically without tones, such as by the Dungans and others

5
This Pinyin–based system was initiated in 1987 as a joint project of the Electronics Instrument Works at Beijing University and the State Language Commission. Despite some good features, the system is of quite limited utility.

(Mair, 1990; DeFrancis, 1950), shows the feasibility of such an orthography for most purposes, the sometimes absolute necessity of tone indication, as in dictionaries and wordprocessing, runs up against resistance in Chinese that contrasts sharply with the situation in Vietnamese.

The French, who take their own diacritics as a given, in promoting the romanized system called Quoc Ngu with its fussy system of six superscripts and one subscript also passed on to the Vietnamese the habit of always adding these symbols, even in informal situations of writing their tonal language. In contrast, most transcription systems for Chinese, such as the long–dominant Wade–Giles romanization, which represented tones only under duress, were created under the influence of diacritic–less speakers of English, who apparently passed on to the Chinese a distaste for these pesky symbols. The literature is full of complaints of the added work involved in tone indication.

If we take tones into account, as we must in Chinese wordprocessing, the characters 或者 representing the word *huòzhě* "perhaps" will not be evoked if we correctly indicate tone 2 of the first syllable by typing *huózhe* for "live on." But we were unable do this on TianMa, since it was incapable of dealing with diacritics. The CW.3 system is able to, however, and does so in a simple fashion. One types *huo2zhe* to evoke *huózhe* on the screen and bring up the characters 活着 for "live on." Similarly by adding tone 3 to *si* and typing *si3,* we evoke *sǐ* on the screen and bring up the character 死 meaning "to die"; it turns out that this is the only character that matches this transcription. Thus by its use of tone indication CW.3 avoids both the errors generated by TianMa and outputs an accurate character version of the two lines of Pinyin.

There is also another difference between TianMa 2.0 and CW.3 that relates to one of the major problems in wordprocessing and one of the major factors in the inefficiency of the Chinese script. That is the question of parsing or segmentation of text.

Text Segmentation

Chinese texts consist of a succession of characters, which invariably represent syllables, with no segmentation of text except that provided by punctuation, which is a fairly recent addition in its present Westernized form. They therefore resemble European scripts before the introduction of word–division, which was developed late by the Greeks and did not become general in Europe until about the year 1000 (DeFrancis, 1989:256).

In English a word is in effect defined as the unit surrounded by white space on the printed page. It is the thing that gets listed in dictionaries and is bandied about in all sorts of situations, including wordprocessing, as the basic unit in the orthography. The treatment accorded characters is comparable to that of words in English, but the two are by no means linguistically equivalent. While all words in English have meaning and are considered free forms, in the case of Chinese characters, by my calculation based on a classification made by the eminent Chinese linguist Y.R. Chao, only some forty-four percent are free forms, forty-five are bound morphemes of the type *er* in *teacher* and eleven percent are meaningless syllables of the type *cor* and *al* in "coral" (DeFrancis, 1984a, 184–185).

Although careful scholars with enough linguistic sophistication to know the difference between word and morpheme label characters essentially as morphemes, most Chinese, and others as well, look upon them as words. To them a character *(zì* in Chinese) is thought of as a "word," but linguists properly distinguish *zi* "character" from *cí* "word," a new term that did not come into usage until this century and is still not widely known by the general public. Add to this a general inability to distinguish between speech and writing, an inability of course shared with many Westerners as well, and we get confusion much compounded in the case of Chinese.

The results of this confusion are far–reaching. The belief that every character is a word leads to the belief that the uniqueness of each character, which typically has a phonetic element combined with a distinguishing semantic element such as 氵 "water," would degenerate into insupportable ambiguity if the character was replaced by its transcription. The folklore of Chinese characters, and even at times scholarly studies of the script, are replete with such horror stories as "the ninety words pronounced *li.*" All too often homographobia reaches the extreme of citing Y.R. Chao's playful presentation of an intelligible, if rather bizarre, stringing together of twenty-four different characters, all tonal variations on the syllable *xi,* as proving that Chinese cannot be written in an alphabetic script (DeFrancis, 1984a, 192, 196–197; 1985). This is like proving that English and French cannot be written phonemically because in English *what* *wood a wood chuck do if a wood chuck wood chuck wood* cannot replace *what would a woodchuck do if a woodchuck would chuck wood* and in French *si si sã si si si si sã si siprɛ* cannot replace *si six cents six scies scient six cents six cyprès* "if six hundred six saws saw six hundred six cypresses."

Yet this obdurate silliness does point in the direction of a real problem, one that is comparable to converting phonemic English or French into their respective orthographies. The problem of choosing among the five graphic realizations of the syllable *si* is similar to the task of inputting a Pinyin syllable and converting it to a desired character.

In French, all except the last occurrence of *si* do present difficulties, though perhaps not excessively apart from made–up games. The problem of converting *si* to *cy* is a non–problem since its close combination with a succeeding syllable results in a unique succession of phonemes that can easily be matched with the conventional orthography.

This sort of solution is not so easy for Chinese to apply because of their belief in the autonomy of characters and their lack of experience in segmenting text and discriminating words. If asked to think about the Pinyin transcription for either spoken or written material, most Chinese (the literate minority, of course), would think of the characters involved and would then envisage either a succession of Pinyin syllables separated by white space or a succession of Pinyin letters like European writing before the introduction of word–division. Exasperated writing reformers constantly cite shop signs and building names written in the fashion of either ZHONG HUA REN MIN GONG HE GUO or ZHONGHUARENMINGONGHEGUO for Zhonghua Renmin Gongheguo "Chinese People's Republic" (Zhou Youguang, 1990).

In English and French we have had centuries of experience processing text using word–division, capitalization and other orthographic conventions that have evolved over the centuries in a rather haphazard fashion. Only occasionally do we have to refer to such authorities as the *Chicago Manual of Style* to determine whether to hyphenate, separate, join things together or adhere to other conventions.

The Problem of Orthography

Besides lacking the Westerners' experience with word–division and text segmentation, the Chinese, despite the official adoption of Pinyin in 1958, have not been taught how the system might function as an orthography. Indeed there has been continuing opposition to the use of Pinyin for anything more than the transcription of individual characters.

A breakthrough occurred in July 1988 with the official promulgation of rules for Pinyin orthography by the State Education Commission and the State Language Commission, the latter the successor of the Language Reform Commission, the renaming of which reflected the official downgrading of writing reform. For foreigners, and for Chinese who can read English, another advance was made in 1990 with the publication in the PRC of *Chinese Romanization: Pronunciation & Orthography*, which expands on the previously published rules by presenting detailed explanations of the rules and copious examples of how to write connected text in Pinyin (Yin and Felley, 1990). Although these works have the potential of accomplishing instant codification that took centuries to evolve in the West, disagreement within China concerning some of the rules has blocked their full acceptance as the official standard orthography for Chinese. Nevertheless, even as a first step toward the important goal of standardization, they offer the possibility of greatly increasing the efficiency of handling characters on computers for those users who prefer input systems based on Pinyin. Contrary to earlier predictions, such users now include virtually all Chinese and non–Chinese outside of China, and eighty to ninety percent of individual users of computers within the country. Shape–based inputting, which has received heavy government support despite outraged objections from many computer specialists, is now largely confined to specially trained professional typists.

But for improvements to become consolidated and generalized it will be necessary for producers of software to follow a standardized orthography and for those involved in wordprocessing to acquire familiarity with the system to at least some degree approximating the command by Westerners of their own orthographies.

This will be difficult to achieve in a situation where Taiwan objects to the Communist transcription system (along with the simplified characters), the PRC emphasizes simplified over traditional characters, Westerners don't bother with tone marks, Chinese are hobbled by mental blocks, and no one is strong enough to carry the day in any area, including the basic problem of orthography that was settled long ago in the West.

Standardization of orthography, if its need is accepted, may require considerable revamping of current software. For example, TianMa would have to change the Pinyin for the number "twenty-seven" from *èrshí qī* to the CR system's *èrshíqī*. It would also have to adopt the CR rule of attaching verb suffixes, as in *huózhe*, which uniquely brings up the characters 活 for "living," in place of its practice of separating *huo* and *zhe*, which requires a further selection of the desired graphs from among the two groups of homographs displayed on the TianMa screen.

Also needed are alphabetically-arranged aids of various kinds, such as dictionaries and thesauruses. At present, thanks to the yeoman efforts of writing reformers like Zhou Youguang and other members of the now defunct Writing Reform Commission, and with little thanks to the paltry government support, a word-list called *Hanyu Pinyin Cihu,* containing 60,400 entries, has been produced with a strictly alphabetic arrangement of Pinyin followed by characters (*HPC,* 1989).

A useful but neglected feature of this important but little-known work is its flagging of homographs by single-asterisking those due to non-indication of tones and double-asterisking those which occur even with tone indication. While there is much wasteful duplication of effort to produce marginally better dictionaries in one or another of the old inefficient arrangements, it has not been possible to gain sufficient support from scholars or the government to turn the *HPC* word-list into the sort of alphabetized dictionary that is desperately needed to help standardize Chinese orthography and make it more adaptable to computers and other modern information processing devices.

A project is close to completion, however, to produce a Chinese-English dictionary along these lines based on the Pinyin orthography as spelled out in *Chinese Romanization.* Initiated by Professor Victor H. Mair of the University of Pennsylvania and joined by several PRC colleagues and volunteers in the United States, including myself, the project has produced an *ABC (Alphabetically Based Computerized) Chinese-English Dictionary* of over 71,300 entries that will go far toward meeting an urgent need of scholars of Chinese and all those concerned with the computerization of the language and its system of writing. In addition to its strict alphabetical arrangement, the *ABC Dictionary* is also unique in making use of *HPC* and frequency data to distinguish monosyllabic homographs and the much less frequent polysyllabic homographs by utilizing raised numbers before the transcription in the fashion of English dictionaries, thereby producing such distinctive transcriptions as ^{1}ba ^{2}ba ^{3}ba ^{4}ba ^{5}ba for 罢 霸 坝 耙 爸 respectively.

While this dictionary is aimed primarily at Westerners, among them those concerned with the computerization of Chinese, it will doubtless also be of help to the Chinese themselves, especially to those reformers who have long been battling to push Chinese lexicography in this direction and to develop other tools based on the alphabetic principle. Such forward–looking reformers are in the forefront of the wordprocessing advances that have already been made, including research on the use of contextual clues in order to seek solutions to the problems that currently defy automatic conversion based on single characters and combinations of characters.

Eventually the efficiency of computerized handling of characters can be maximized only if there is a change in writing style in the direction of bringing it closer to actual speech. As Y.R. Chao pointed out years ago, intelligible romanized writing can be produced only if it is based on intelligible speech, speech that is "clear and full in sound" (Chao, 1934). And only writing in characters that is intelligible when read aloud can be converted into intelligible romanized writing for inputting into computers. This means exorcising the bugaboo of homography by writing polysyllabically to reflect the polysyllabicity of speech. So long as Chinese insist on writing in a Hybrid Vernacular, showing off their knowledge of the "unspeakable" classical style and their ability to play games with characters, a character–by–character transcription of text will continue to make for excessive homography and hence for inefficient use of computers. The more classical the style, the more inefficient the inputting, so that for fully classical texts there can be no hope of even a modicum of efficiency on computers.

For those Chinese who do succeed in writing more or less as they speak, the possibility is presented of their being able to employ Pinyin with maximum efficiency by touch–typing a whole text, whether article or even book, whether copied from a handwritten manuscript or composed in their head as they type along, and then converting the whole thing by pressing one or two conversion keys. This is unachievable by convoluted shape–based approaches which may make do for mechanical copying of text by professional typists but are incompatible with fluent touch–typing by ordinary users of computers who must concentrate on content. (The prominent Chinese reformer Zhou Youguang makes an important contrast between "copy–typing" and "think–typing.") And it is not too far–fetched to envisage that some Chinese, perhaps some current first–graders growing up digraphic, may eventually discover that for certain purposes they can dispense with characters altogether, in which case their efficiency with a Pinyin–in–Pinyin–out approach may well exceed that of Westerners working with English, French, Russian and other alphabetic scripts.[6]

6
For further discussion of input techniques and other matters involved in handling Chinese characters on computers, see Unger, 1987. Unger's discussion centers on Japanese, but much of what he has to say applies also to Chinese.

Writing and . . . Reading

Yet if this sort of emphasis on Pinyin may make it possible to computerize the Chinese language with maximum efficiency, it may have the opposite result, if we accept the claim cited at the beginning of this paper, of losing the greater efficiency allegedly enjoyed by characters for reading and comprehension. The view is widely held not only by Chinese but also by many Westerners, including students of the reading process, that although a simple phonemic system of writing may be easier to learn, in the long run it may not serve experienced readers as well as a more complex orthography. In a conclusion based on arguments that they admit are "speculative," two reading researchers state that

> *the ideal orthography for spelling is incompatible with the ideal orthography for reading. Any useful orthography must be a compromise between these two requirements (Frith and Frith, 1980,295).*

Another scholar, in a context dealing specifically with Chinese, states that

> *there is no evidence to suggest that a writing system which is efficient in terms of grapheme–phoneme correspondence is superior to other systems and is easier to read. Kyöstiö's (1980) study on the Finnish language, which is considered to have one of the more efficient writing systems in this respect, indicates that it is only the mechanical aspect of*

reading that is easy for Finnish children to master at an early age (King, 1985, 111).

In these remarks King has conflated and confused several aspects of the reading process which are related but should be clearly distinguished. One is "reading" in the sense of being able to sound out graphic text. (Of course I do not mean that we always verbalize in reading, but that we must have the ability to use phonetic clues to the extent needed to identify a piece of text.) We might give this a dictionary listing as "1read *verbalize text*." Another is "reading" in the sense of deriving meaning from text. Dictionaries might list this as "2read *understand text*." Still another aspect is "reading" in either sense as a subject of comparative study.

For the first of these aspects the evidence is overwhelming that a close grapheme–phoneme correspondence makes for a writing system that is easier to 1read, though this accomplishment is devalued by Kyöstiö and King as merely "mechanical." Kyöstiö states that "the answer to the question of easiness in reading the Finnish language is affirmative as far as mechanical reading is concerned." Then he adds: "But if by reading we mean a higher level skill, the answer *might* [emphasis added] be the same as in other languages" (Kyöstiö, 1980, 89).

However, leaving aside the obviously identical difficulty in all languages due to content (ordinary people might be able to 1read but not 2read scientific treatises), the answer is clearly not the same in Finnish as in other languages. From Kyöstiö's own account it is apparent that the problems he cited for Finnish children are largely related to auditory, visual, psychological and other disabilities. Their problems cannot be treated as comparable to those of otherwise perfectly normal American and Chinese children who are still unable to 1read well enough to 2read with comparable efficiency long after their normal Finnish counterparts or Chinese first–graders literate in Pinyin (DeFrancis, 1984a, 168, 283).

For normal readers reading in their native orthography and at their cognitive level, given their command over a huge vocabulary and a complex linguistic structure, 1read = 2read. It is precisely because of this equation that reform–minded Chinese teachers are insisting that school children must retain command of Pinyin so they can be provided with mind–stretching content years before they can handle it in characters.

This equation is true, of course, only for native speakers and readers. While non-natives commanding one latin-based script, if confronted by a similarly based foreign script, can in just a few hours learn to 1read it aloud well enough to be understood by native listeners of that language, in attempting to 2read in that script they would experience difficulty proportionate to their lack of command of its spoken form. For them it would doubtless be helpful to have non-phonetic supplements, such as capitalization in German, which as broadly applied in that language is not needed by native speakers and is considered by many to be a nuisance that should be circumscribed.

But the views regarding the efficiency of Chinese writing, the system most extensively endowed with non-phonetic supplements, as expressed by King, Li Youren and others, and of writing in general by students of the subject, are usually concerned less with the acquisition of "mechanical" reading by children and other novices than with the act of reading by adults skilled in the art. In their view writing systems like Chinese and English make up for any weaknesses that may exist in the sound-symbol relationship by the greater efficiency supposedly provided by spelling distinctions and semantic embellishments.

These unproven claims are made all the more suspect by the fact that they are most often advanced by persons who compare their greater facility in handling a complex orthography, which they most likely learned first and became habituated to by a lifetime of practice, with their lesser facility in handling a less complex orthography. Chinese born and raised with their traditional script, if they know English but have had only the usual minimal exposure to Pinyin, or none at all, say they can read characters faster than English or other alphabet-based scripts. Americans habituated to their traditional orthography insist that its vagaries are actually essential for enabling them to read faster as compared to a purely phonemic rendering of the language. Their claims are probably right, for the same reason that trained athletes can out-perform untrained amateurs, but as scientific proof for the superiority of one or another script such subjective observations and assumptions are virtually worthless except to confirm what I have facetiously called "The Law of the First Script"; this law states that the first script learned is generally considered the most satisfactory, if not most perfect, of all possible systems (DeFrancis, 1950, 200).

In opposition to these views and to the claim that the ideal orthography for reading must be different from an ideal orthography for spelling, and leaving aside the practical question of social acceptance, I should like to propose the counter theory that, given the phenomenal verbal resources possessed by normal native speakers, the optimal orthography in any language for readers at any level is one which maximally capitalizes on this capacity by matching written symbols with spoken utterances in the simplest possible way, either phonemic or syllabic depending on the language, with no contamination whatsoever by non–phonetic accretions.

I don't know how to set up scientific tests for any of these theories related to optimum orthography. Robert Cheng (personal communication) suggests a cross–systems test involving three groups each consisting of one or two dozen matched individuals who are accomplished monolingual readers of Finnish, English and Chinese. But how can the individuals be truly matched? My wife, like me a native reader of English, could read novels two or three times faster than I could, and retained far more details than I did.

But such testing, even if perfected and showing, say, that Americans read fifty percent faster than Finns and Chinese read twice as fast as Americans, is not the end of the story. Those who claim that complex scripts are superior for reading concentrate on this one aspect and hardly ever mention the many areas where they are obviously inferior to simple scripts, as in dictionary lookups and wordprocessing, not to mention basic acquisition of literacy. A leading Chinese psycholinguist, Ovid J.L. Tzeng, has shown the enormous complexity of any cross–script comparisons and the need to factor in many things. Arguing against scholars claiming superiority of *kanji* over *kana* for initial reading instruction in Japanese, which is similar to the claim of superiority of Chinese characters over Pinyin and English, he states that

> *in a real life situation, learning to write is almost an integral part of learning to read. Thus, it may not be very realistic to compare the ease of learning to read* kanji *with that of learning to read* kana *without also considering the difficulty of writing associated with these two types of scripts (Tzeng and Singer, 1978–1979, 664).*

Again, with respect to the idea that unique features in the Chinese script lead to its being processed by the brain differently from alphabetic scripts, he forcefully asserts that

> *words written in an alphabetic script require the dominant left–hemisphere's verbal awareness for their perception and production. So do words written in a logographic script. We think it is time for scientists to base their judgments on empirical data such as we have presented here, rather than on wild speculations and on naive views of orthographic structures (Tzeng et al, 1986, 372).*

And finally, in a review of my own classification of writing systems that rejects Chinese as a meaning–based category and classifies it as morphosyllabic, he states that

> *under such a conceptualization, the Chinese writing system is very much sound–based and accordingly, its reading comprehension depends on the success of recovering its morphosyllabic representation. Indeed, experimental results of recent psycholinguistic and neurolingistic studies on reading Chinese are very much consistent with DeFrancis' analysis (Tzeng, 1991).*

Classification of the Chinese Writing System

The question of the classification of the Chinese writing system is central to understanding its structure and evaluating its efficiency. In their arguments for the superiority of Chinese characters, Li Youren and others of like mind classify Chinese as a unique semantically based system. This view of Chinese as an ideographic (if not indeed a pictographic) system is not very far removed from the more academic classification of Chinese writing as a logographic or morphemic system that is uniquely well adapted to the peculiarities of the Chinese language.

Now while Chinese characters per se may possibly be considered logographic or morphemic, they are not, contrary to widespread misconception, the same thing as the *writing system*, any more than words are the basis for classifying alphabetic systems. The characters, I have argued, comprise frames, that is lexemes or dictionary items akin to our words and morphemes, and they are based on graphemes consisting of phonetic elements representing syllables, in contrast to our graphemes consisting of letters representing phonemes. The so–called "radical" component of characters is a relatively late addition to the phonetic element, whose vitally important role has been minimized owing to the infatuation with the more striking but less useful semantic accretion (DeFrancis, 1984a, 1989).

While avoiding the general mistake of taking the characters rather than the phonetic elements as the basis for classifying the Chinese writing system, William Boltz, a student of early Chinese writing, presents a more sophisticated, but to my mind, still incorrect defense of classifying Chinese as a morphemic rather than morphosyllabic script on the grounds that "non–morphemic syllables...do not exist" and that the phonetic elements stand for both a syllable and a morpheme (Boltz, 1989). Nevertheless, to the extent that that is true (I do not accept the non–existence of non–morphemic syllables), the phonetic aspect, in my view, remains the more basic aspect. As my colleague Y.C. Li has noted (personal communication), while syllables have always been limited, morphemes were not, and in adding morphemes to the language Chinese based them on pre–existing syllables.

It is my basic contention that Chinese writing is primarily sound–based and only secondarily semantically oriented, and that the inefficiency of the system stems precisely from its clumsy method of sound representation and the added complication of an even more clumsy system of semantic determinatives. The greater inadequacy of the semantic elements, those unique features which have so captivated many people, can be seen in a study comparing semantic and phonetic predictability in Chinese.

Useful semantic clues are to be found in less than half of the characters. In not a single case do they unequivocally pinpoint a precise meaning, and at best they point only to broad thesaurus–like areas of meaning. At their far too frequent worst, the semantic elements can be downright misleading. To give one typical example, the "insect" radical is the misleading determinative in characters with such non–insect meanings as clam, egg, snake, frog, jellyfish, hedgehog, rainbow, barbarian, stupid and melt (DeFrancis, 1984b).

As to the phonetic elements, bad as they are, they provide more precise information than do the semantic determinatives. No less than twenty-five percent of the phonetics accurately predict the pronunciation, even as to tone, of the full characters of which they form part, and another seventeen percent precisely indicate the segmental phonemes but not the tones (DeFrancis, 1984, 1989). But they accomplish all this in an exceedingly cumbersome fashion.

How the phonetic elements contribute to the inefficiency of Chinese characters can be seen if we look at the roughly parallel ways in which Chinese and English represent a more or less common syllable, chosen at random, that we can transcribe as /piy/ in English and as *pi* in Pinyin orthography. I present first the fourteen ways (1.0–14.0) in which the syllable is spelled in English, together with one or two words (1.1–14.2) that illustrate their usage:

1.	***-pe***	**5.**	***-pea***	**8.**	***-pie***	**12.**	***-ppy***
	pe*nal*		***pea****cock*		*shar****pie***		*ha****ppy***
	pe*ony*		***pea****nut*		***Pie****gan*		*slo****ppy***
2.	***-pae***	**6.**	***-pee***	**9.**	***-py***	**13.**	***-ppi***
	pae*an*		***pee****wee*		*skim****py***		*ha****ppi****ness*
	pae*diatrics*		***pee****vish*		*ras****py***		*slo****ppi****ness*
3.	***-poe***	**7.**	***-pi***	**10.**	***-pey***	**14.**	***-ppie***
	*sub****poe****na*		***pi****ano*		*do****pey***		*pre****ppie***
	*onomato****poe****ia*		*Ho****pi***		*Pom****pey***		*yu****ppie***
4.	***-peo***			**11.**	***-pei***		
	peo*ple*				*Pom****pei****an*		

Below I present a list of phonetic elements and their derivatives (characters derived from the phonetics) that show the ten ways in which the syllable is spelled in Chinese.[7]

1.	*pi*	辟	**3.**	*pi*	皮	**6.**	*pi*	匹	**9.**	*bi*	毘
	pi	噼		*pi*	披		*pi*	鴄		*pi*	貔
	pi	霹		*pi*	狓	**7.**	*pi*	罷		*pi*	媲
	pi	劈		*pi*	鈹		*pi*	犤	**10.**	*bei*	卑
	pi	擗		*pi*	疲	**8.**	*pi*	比		*pi*	椑
	pi	癖		*pi*	魾		*pi*	批		*pi*	啤
	pi	譬		*pi*	陂		*pi*	紕		*pi*	脾
	pi	鷿		*pi*	詖		*pi*	砒		*pi*	裨
	pi	僻		*pi*	被		*pi*	枇		*pi*	蜱
	pi	闢	**4.**	*pi*	丕		*pi*	毗		*pi*	鼙
	pi	澼		*pi*	坯		*pi*	琵		*pi*	埤
	pi	嬖		*pi*	狉		*pi*	蚍		*pi*	陴
	pi	甓	**5.**	*pi*	否		*pi*	仳		*pi*	郫
2.	*pi*	圮		*pi*	痞		*pi*	阰		*pi*	睥

Note the partial parallels with the English syllables: in English,

7
The list has as its base all the characters listed under *pi* in a fairly comprehensive modern dictionary of 6,000 single–character entries (Wu, 1979), together with a few additional derivatives from other dictionaries formed with the same phonetic elements. A more extensive search, especially among rarer characters, would doubtless turn up more phonetic elements and many more derivatives having the pronunciation *pi*. To maintain uniformity of presentation, I have converted Wu's simplified characters into the traditional forms. Note also that in addition to the deviant pronunciations of the phonetics numbered 8.0, 9.0, and 10.0, number 1.0 is also pronounced as bi and 4.0 as *fou*.

the syllables are combined with other letters to form distinctive words of various lengths (e.g., combining *peo* with *ple* to form *people*); in Chinese, the phonetics are combined with "radicals" to form distinctive morphemes, always of one syllable. Note also that Chinese 8.0 *bi*, 9.0 *bi*, and 10.0 *bei* are like English *pie* in having one pronunciation when alone, another (or several) in combination, as in *piebald* and *Piegan*.

A major difference between the two systems is that of phonemic representation of syllables versus integral representation of that unit. While some English graphemes, especially those representing vowels, may display considerable complexity, others may be fairly simple, as in the case of *p* and *pp* in our English sample. This facilitates guessing the pronunciation of the whole syllable. But if whole syllables are represented by integral graphemes, these are necessarily more numerous and hence more complex than phonemic graphemes. Unless standardized as a simple syllabary like *kana*, any system of syllabic graphemes is inevitably far more complex and hence far less efficient than even a complex phonemic system. In the case of Chinese, its phonetic graphemes are both unstandardized and saddled with the addition of semantic determinatives that would be superfluous in a writing style truly based on speech. All this enormously increases the difficulty of the system.

The inefficiency that the more complex character system brings to many areas of Chinese life, some of which have been noted above, would seem to entail a price tag of staggering proportions. A number of years ago, prompted by the discussion of a cost–benefit approach in *Can Language be Planned?* (Rubin and Jernudd, 1971), I asked a specialist in the economy of China whether that approach could be applied to estimating the comparative costs of the traditional script and an alphabetic system. The reply was negative.

However, the inability to carry out a full–scale cost–benefit analysis does not prevent us from reaching some commonsense conclusions about the inefficiency of the Chinese system of writing and the need to mitigate that inefficiency by a policy of digraphia.

Acknowledgement

I am grateful to Robert Cheng, Susie Cheng, Y.C. Li, Victor H. Mair and J. Marshall Unger for helpful information and critical readings of this paper. The Chinese characters were adapted and created by Chao-jen Wang, a graduate student at the Institute of Design from Taiwan.

References

Anon. 1963. *Modern Chinese–English Technical and General Dictionary.* 3 vols. New York: McGraw–Hill.

Becker, Joseph. 1984. Multilingual Word Processing. *Scientific American,* 251, 96–107.

Boltz, William G. 1989. Review of *Visible Speech: The Diverse Oneness of Writing Systems,* by John DeFrancis. Honolulu: University of Hawaii Press, 1989. In *Sino–Platonic Papers,* 14, A1–A9.

Chao, Y.R. 1934. The Idea of a System of Basic Chinese. *Quarterly Bulletin of Chinese Bibliography,* 1:4, 171–183.

Coia, David A. 1985. The Chinese Language and Computers: An Introduction. *Journal of the Chinese Language Teachers Association,* 20:3, 79–92.

DeFrancis, John. 1940. Far Eastern Type for American Scholarly Publications. *Notes on Far Eastern Studies in America,* No. 7. Washington, DC: American Council of Learned Societies,1–8.

——. 1950. *Nationalism and Language Reform in China.* Princeton: Princeton University Press.

——. 1966. Why Johnny Can't Read Chinese. *Journal of the Chinese Language Teachers Association,* 1:1,1–20.

——. 1984a. *The Chinese Language: Fact and Fantasy.* Honolulu: University of Hawaii Press.

——. 1984b. Phonetic Versus Semantic Predictability in Chinese Characters. *Journal of the Chinese Language Teachers Association,* 1:.1,1–21.

——. 1985. Homographobia. *Xin Tang,* 6 (November), 2–17.

——. 1989. *Visible Speech: The Diverse Oneness of Writing Systems.* Honolulu: University of Hawaii Press.

——, editor. (in press). *ABC (Alphabetically Based Computerized) Chinese–English Dictionary.* Honolulu: University of Hawaii Press.

Frith, Uta and Christopher Frith. 1980. Relationships Between Reading and Spelling. In *Orthography, Reading and Dyslexia,* by James F. Kavanagh and Richard L. Venezky, editors, 287–317. Baltimore: University Park Press.

HPC. 1989. *Hanyu Pinyin Cihui [Chinese Pinyin Wordlist].* Revised edition. Beijing: Yuwen chubanshe.

Kyöstiö, O.K. 1980. Is Learning to Read Easy in a Language in which the Grapheme–Phoneme Correspondences are Regular? In *Orthography, Reading and Dyslexia,* by James F. Kavanagh and Richard L. Venezky, editors, 35–49. Baltimore: University Park Press.

Li, Augustine. 1940. An Experiment in Cataloguing Chinese Books. *Notes on Far Eastern Studies in America,* No. 7. Washington, DC: American Council of Learned Societies, 10–19.

Li Xiuqin. 1990. *Evolution de l'écriture chinoise.* Beijing: Shangwu yinshuguan.

Li Youren. 1991. Tan Hanyu Hanzi de tedian [On the Distinctive Features of the Chinese Language and Chinese Characters]. *Yuwen jianshe tongxun,* 33, 68–70.

Mair, Victor H. 1986. *The Need for an Alphabetically Arranged General Usage Dictionary of Mandarin Chinese: A Review Article of Some Recent Dictionaries and Current Lexicographic Projects = Sino–Platonic Papers,* No. 1. Philadelphia: University of Pennsylvania.

——. 1990. *Two Non–Tetragraphic Northern Sinitic Languages = Sino–Platonic Papers,* No. 19. Philadelphia: University of Pennsylvania.

Mathews, R.H. 1945. *Mathew's Chinese–English Dictionary*. Revised American edition. Cambridge: Harvard University Press.

Miming. 1946. *Miming ma dianbao shu [Covert and Open Telegraphic Code]*. Shanghai: Shangwu yinshuguan.

Rubin, Joan and Björn H. Jernudd. 1971. *Can Language be Planned?* Honolulu: University Press of Hawaii.

Simon, Walter. 1947. *A Beginner's Chinese–English Dictionary of the National Language (Gwoyeu)*. London: Lund Humphries.

Sung, Margaret and Alex Kuo. 1986. A Linguistic Evaluation of Chinese Computer Input Systems. *Journal of the Chinese Language Teachers Association*, 21:2, 43–57.

TianMa. 1986. *TianMa Handbook* Version 2.0. Vancouver: International Geosystems Corporation.

Tzeng, Ovid J.L. 1991. Review of *Visible Speech: The Diverse Oneness of Writing Systems*, by John DeFrancis. Honolulu: University of Hawaii Press. In *Contemporary Psychology*, 36:11. 982.

——, Daisy L. Hung, Sylvia Chen, Jori Wu, and Mao–Song Hsi. 1986. Processing Chinese Logographs by Chinese Brain Damaged Patients. In *Graphonomics*, by H.S.R. Kao et al, editors. Amsterdam: North Holland, 357–374.

—— and Harry Singer. 1979. Failure of Steinberg and Yamada to Demonstrate Superiority of *Kanji* over *Kana* for Initial Reading Instruction in Japan. *Reading Research Quarterly*, 14:4, 661–667.

Unger, J. Marshall. 1987. *The Fifth Generation Fallacy. Why Japan is Betting its Future on Artificial Intelligence*. New York and Oxford: Oxford University Press.

Wang, Fred Fangyu. 1966. *Mandarin Chinese Dictionary. Chinese–English*. South Orange, New Jersey: Seton Hall University Press.

Wang, William S.–Y. 1981. Language Structure and Optimal Orthography. In *Perception of Print: Reading Research in Experimental Psychology*, by Ovid J.L. Tzeng and Harry Singer, 223–36. Hillside, New Jersey: Lawrence Erlbaum Associates.

Wu Jingrong. 1979. *The Pinyin Chinese–English Dictionary*. Hong Kong: Commercial Press.

Xiandai. 1980. *Xiandai Hanyu Cidian [Contemporary Chinese Dictionary]*. Beijing: Shangwu yinshuguan.

Yin Binyong and Mary Felley. 1990. *Chinese Romanization: Pronunciation & Orthography*. Beijing: Sinolingua.

Zeng Xingchu. 1983–1984. An Appeal for Justice for Characters. *Chinese Education*, 16:4, 50–81. This is a translation of Zeng's 1983 article "Hanzi hao xue hao yong zheng [Proof that Chinese Characters are Easy to Learn and Easy to Use]." *Jiaoyu yanjiu*, 1, 73–79; 2, 58–63.

Zheng Linxi. 1983. *"Hanzi hao xue hao yong zheng" de yixie lunju bu he shishi* [Some of the Arguments for "Proof that Chinese Characters are Easy to Learn and Easy to Use" do not accord with the Facts].*Jiaoyu yanjiu*, 11, 53–57,72.

Zhou Youguang. 1990. Foreword. In Yin Binyong and Mary Felley. *Chinese Romanization: Pronunciation & Orthography*. Beijing: Sinolingua


方も最先端のインズ
入会した方にもれ
10R
WBA J・バンタム級9位
鬼塚勝
13戦13勝(12KO)
フリーダイヤル
たら金銀
ラン
対決
新
進撃!!
ヤル
KA