HOW TO OVERCOME ANALPHABETISM IN READING CHINESE CHARACTER

Livio C. Piccinini
Ting Fa Margherita Chang
Giovanni Tubaro
Department of Civil Engineering and Architecture
University of Udine, Italy

Abstract. *Most languages in the world use some system of alphabetical characters: Latin, Greek, Cyrillic, Hebrew, Arabic, Hindi and so on. A foreigner that does not know the language can get some information from a written text provided he knows the alphabet and disposes of a dictionary. When there is no longer an alphabet, but only pictorial characters, the problem becomes at first unsolvable. Chinese is the main language where an ignorant foreigner is completely analphabet. Fortunately there are methods that after some training allow the recognition of pictorial characters. In our university some twenty pupils of the Excellence School participated to an experiment of Chinese alphabetization gluing a traditional practical Chinese first course with information theory methods for dealing with image data bases. In this article first we discuss both the theoretical foundations. Then we give a report of the merging of the two conceptual schemes as it was performed at the excellence school. Finally we draw some conclusions about improvements of the method.*

Keywords: *Analphabetism in Chinese characters, Chinese writing, Data mining in pictorial data bases, Decomposition of symbolic drawing, Udine's Scuola Superiore*

Introduction

Nowadays the problem of analphabetism in reading Chinese characters is mainly restricted to non Chinese people, and we are dealing with this situation. Likewise we do not deal with professional students of Chinese language, who must become alphabetized like a Chinese pupil[36], though some didactical differences might arise in view of the different age.

A common European experience is the fact that in western European countries you can read what you see along a road, or you can read the title of a placard or of a newspaper and, *using the dictionary*, you can at least understand the topic and/or the main information that is conveyed, even if you do not know the language. Some capability of distinguishing between grammatical particles and semantic items may be required in order to divide properly words into components, what allows not to lose time in obvious words, and also allows the use of small pocket dictionaries[37].

The problem becomes more complicated where Greek or Cyrillic characters are used, since they require some training in reading. Furthermore the number and

[36] For the Italian standard compare [1] and [14].

[37] A typical case is given by composed words of German, like dementsprechend (dem-ent-sprech-end = correspondingly to that).

order of letters in the alphabet no longer coincide with western alphabet. The number is no problem, but the order becomes relevant in the physical use of a dictionary. The problem is relevant even if you use the computer, because you must get used to a modified keyboard.

Defective alphabets totally different from Latin present increasing difficulties. Hebraic and Arabian alphabets at least retain much of the original alphabetical sequence[38], while for example Indian alphabet shows a fully different sequence. It is still possible to make an alphabetical search in the dictionary, but much more exercise and a minimal knowledge of the language is required.

Chinese symbols or, at least, a fixed set of its symbols, build a finite list of images that can be learnt like an alphabet, with two main differences: the number of pictures is very high and for most of them there is no rule for the pronounce. The second fact is not relevant when searching in the dictionary, even if psychologically when you cannot design a symbol by its name you find recognition and memorization more difficult. What indeed puts a dreadful barrier to entry for a non professional user is the first fact. Remember that, while western pupils need few months to learn the alphabet, Chinese pupils require many years for the same achievement. Foreign students of Chinese language at the university or higher schools usually require two years of intense study to learn the standard set of characters. The official HSK test (Hanyu Shuiping Kaoshi) for foreigners includes 2063 characters, but there is anyhow the possibility of finding characters not known, hence to be looked for in the dictionary. A powerful help, also for Chinese pupils, is given by the Latin transcription called *pinyin,* that means *to note sounds*, and is official since 1956 [39]. Unlike Japanese, where an analogous transcription is satisfactory and widely used, for Chinese language the pure transcription in Latin characters cannot distinguish satisfactorily between all the homophones, hence it cannot replace the ideograms. A further hindrance is that each character does not represent a sound but fundamentally represents a (monosyllabic) word, even if nowadays many words are bisyllabic or multisyllabic, and glue together different concepts in a "creative" way[40]. The dictionary lists multisyllabic words according to the first symbol (hence it uses the lexicographic order like the alphabetic dictionaries, but in a very restricted way).

Therefore the main problems remain: how can a dictionary of pictures be ordered; how can a picture be recognized in the list in a reasonable time? Both problems arise and are widely studied in computer science. The first problem is typical of geometric and pictorial data bases, while the second has been studied also in artificial intelligence by OCR (Optical character recognition) experts. In

[38] Remember that the first alphabet was Phoenician, that is actually Semitic.

[39] Remark that it includes four diacritics accents, that, writing on the computer, must be ignored or replaced by other conventions such as numbers following the syllables. Compare [1], pp. 15-18.

[40] For example the "phonetic" ideographic transcription of Milan (Mi- Lan) was chosen as "rice orchid".

this case the solving path seems to be similar to what can be used by human intelligence.

Structure of Chinese characters

We start with the description of the problem to be solved, so that information theoretical section may be customized according to the actual needs.
First of all Chinese characters are not so pictorial as, for example, the complete set of road signs. The reason is that rather than being mechanically reproduced they used to be handwritten (by a thin brush). The evolution tended to standardization, achieved by Kaishu in 3 rd century[41]. The contemporary printed characters are mainly founded on this standard[42], even if in 1956 a simplified form of some characters became official in People Republic of China. Some cursive forms exist in handwritten Chinese, namely Caoshu (almost a shorthand) and Xingshu, that still allows a good readability (for an experienced reader!). From now on anyhow we refer to printed characters.
Characters are always composed by a fixed number of single strokes. There is a variety of strokes but they are essentially linear (horizontal, vertical, descending to the left, descending to the right, ascending); there are some forms of "point"; some strokes may be formed by a sequence of linear components with sharp angles, and finally some hooks may end the stroke.
There are strict rules according which, viewing a character, it is possible to enumerate its strokes and recognize them. The number of strokes allows also to check the correct recognitions of the character. One of the authors remembers that an uncle of hers, who was dean of an high school in China, knew by heart the exact number of strokes of each ideogram. The sequence of strokes follows some very general rules, so that each character should be handwritten following a fixed order.
We show some examples: in the first four of them the count of strokes is straightforward. Some more care is required in the last two. The number of strokes is listed after the colon.

十 [shí] ten: 2,　　王 wang[2] (king): 4,　　大da[4] (big): 3,
子zi [3](son): 3　　口kou[3] (mouth): 3!,　　女nu"[3] (woman): 3!.

There are four main ways of creating a character:
1. Pictogram (xiangxin) originates from the picture (sun=日ri[4])

[41] We refer to [5], pp. 27-30 and pp. 31-34 for this and the following information.
[42] Already de Saussure in sections 45-47 of [16] pointed out the exceptional prestige of Chinese calligraphy, that was adopted in the whole of China and even in Japan, for a completely different language!

2. Meaningful (zhishi) adding symbolic elements to a pictogram in order to mean some abstract concept (root = 本ben[3] from tree = 木mu[4], adding the basis)
3. Logical compounds (huiyi) where two ideograms are joined in a logical way to define a new concept (good = 好hao[3] from woman = 女nu"[3] and son = 子zi[3])
4. Ideophonetic compounds (xingsheng) where one ideogram recalls the pronounce (but does not mean anything) and one broadly states the meaning, or rather the semantic class to which the word belongs[43]. Sometimes phonetic elements appear tied with many different radicals; that is the reason why a simple transcription in pinyin may not be sufficient to a correct understanding. We give some ideograms (not all of them) derived by qing[1] = green.

精 jing[1] =essence, 睛 jing[1] =eye, 靖 jing[4] =quiet, 青qing[1] =green, 清 qing[1] =clear, 情 qing[3] = love, 晴 qing[3] = clear

In the cases 3 and 4 the ideogram that supports the meaning is called radical; it can be an ideogram that actually exists by itself, or a symbol used only in compound characters. What complicates the analysis is the fact that the radical can appear in any position (left, right, top, bottom, around) and that anyhow there are a little more than 200 radical classically recognized. Also non radical-supported ideograms of types 1 and 2 are assigned by convention to a radical, but in this case the decomposition is not evident and both experience and luck are required. This depends on the fact that radicals may not be isolated in the whole of the character, but they may be interconnected. Examples are 中 zhong[1] (center, China), where the radical is the vertical line (no corresponding ideogram) or 我 wo[3] (I) and 成 cheng[2] (to become) where the radical 戈 ge[1] (hatchet) is mixed with other strokes.

But what is the advantage of the decomposition? Type 4 gives hints on the pronounce, what does not happen in cases 1 to 3. The main advantages are connected with the use of the dictionary. Nowadays the characters are listed according to the pinyin (Latin) transcription. In case of tie they are listed according to the tone (indicated by the four diacritic symbols). But for an analphabet of Chinese characters that does not help, since a graphical index is required; unfortunately many commercial dictionaries do not contain any index, so that they are of no use to the illiterate pupil.

We refer to three Chinese-Italian dictionaries that have an index: in [22] and [21] the index is based firstly on the radicals (as in the tradition) and secondly on the number of strokes that are added to the radical, while in [20]

[43] We can consider it as a severe form of what Eco calls hypocodification (compare [7], pag. 191).

classification ignores the radicals and is based primarily on the total number of strokes and secondly on the starting stroke.

The search of the radicals is organized according to the number of strokes, and follows a rather consolidated order. The user must know which ideograms are already radicals, because they should not appear in other parts of the list. The handy [22] fortunately has a certain redundancy, so that uncommon radicals appears also as compounded ideograms that start from easier radicals incorporated (usually one-stroke radicals).

From the side of computer science

The theoretical foundations required to deal with pictorial data bases are much more sophisticated than alphanumeric data bases. Much effort has been made in the last decades to make them efficient. We can recall some classical books where the main approaches have been thoroughly presented: in [3] conceptual data bases are discussed, with further specialization based on the knowledge in [8]. [10] was more specialized in modeling objects and environment. A good pictorial data base must be efficient in its structure, and must be accessed from pictorial information, either directly or by means of preliminary partitions and/or analysis. A good exercise for the reader is to rethink the way of listing the flags in order to recognize them by their image, or to think how to enter and decode a full (and somewhat cumbersome) list of road signs.

The comparison of images is a long lasting problem of computer vision. In particular OCR (optical character recognition) performs satisfactorily, though not yet perfectly (not even for printed characters) as it could be expected. For handwritten texts it depends whether they are personal notes or if they are meant to be read, and in some cases a training procedure is required.

For printed characters the most popular method is template matching or even a comparison with many items of different types, what helps also for handwritten characters. As Mc Cormick states in chapter 6 of [12] OCR is one of the algorithms that "change our future", since it can favorably compete with human intelligence. But for handwritten Chinese characters it seems not to work. On the contrary it works for printed characters, but it cannot be transferred to human understanding because the number of models is too high to ensure a first strike recognition, with the exception of a limited number of characters that are either very simple or much different from any other[44] .

The good hint was given from the very beginning by Chang [6] that pointed the role of radicals. In this case the cumbersome decomposition in elementary strokes and their recollection ([19]) can be overcome by some more global methods as it was pointed out in [18].

[44] It holds de Saussure's mutual exclusion principle ([16], sections 160-161).

Many suggestions that come from artificial intelligence can be used to organize also human intelligence, taking of course into account the differences due on one side to the high capacity of memory and the high speed of computers and on the other to the great parallelism of neural circuits of human brain. Actually for human intelligence a mix of the last two methods seems to be adapt. A first global attempt can be made to find out simple radicals, in case of negative exit a more detailed analysis is performed on the potential radicals, with a check in the list. Clearly it is a struggle with time. Therefore it is needed that our a priori knowledge follows the laws of information theory about frequency and productivity of information[45]. Another question arises: is the classical Chinese ordering of radicals optimal according to the modern principles of pictorial data bases? Of course not, since some radicals have a lot of extensions, and even if their dependence list is segmented according to the number of strokes, there is a lot of characters that must be detected by direct inspection of the whole sub-list (up to 30 for radicals like kou[3] (mouth) or mu[4] (wood). Others on the contrary generate very few characters or even no other character (shu[3] = mouse in [22] generates only itself). For practical uses many other lists could be created, both more functional and more essential, as usually the pupil prepares for his own use while he is learning. The method of the first stroke used in [20] is joined with the number of characters: this method is quite suitable for simple ideograms when the lazy learner forgets some fundamental ideograms and reverts to analphabetism, but huge sub-lists (more than 100 non ordered items) easily arise. Since the order of strokes is easily recognizable, the second and the third strike could be added, so that some form of alphabetic sequence could be generated and joined with the number of strokes.

The experiment at Udine’s Excellence School

The experiment was performed in 2010 with a group of pupils of the Excellence School of Udine University (“Scuola Superiore” dell’Università di Udine)[46]. The general structure was similar to a beginners’ practical course in Chinese[47] and the text followed was a classical in its line [11]; some integration was taken from the textbook of the course in “Chinese Cultural Intermediation” of Rome [14]. From the very beginning a personal copy of dictionary [22] was given to each pupil, and enlarged copies of the index of characters (pp. 9-28) were made available.

In usual courses *pinyin* transcription is used for the pronounce, and, for the most diligent students, for the check of ideograms, but no help is given for

[45] Criteria follow the fundamental work of Shannon and Weaver [17], but keep into account also information evolving according to experience (and authority) like it happens with Google as Mc Cormick remarks in ch. 2 & 3 of [12]. For a classical reference book see [13].

[46] For further information see [15].

[47] Our long experience trainer Dr. Gao Peiqi directed the course.

recognizing ideograms from their picture. On the contrary in our alphabetization course each pupil had to make an individual search in the index for each new word. In the first lectures the teacher suggested possible radicals, but after the tenth lesson this help was reserved only for difficult cases, and in most cases there was no longer any help at all.

Since the tenth lesson the students began simple exercises of decoding sentences with unknown words. The sentences were written at first carefully in Kaishu and later on directly in Xingshu, so that the pupil had to reconstruct the correct number of strokes (without the help of *pinyin*). At the end a colloquial text of well known meaning was divided into small sections for the translation by the pupils[48].

Of course the emphasis on reading reduced somewhat the time dedicated to speak and to learn grammar, what would be a hindrance in a course intended to prepare for efficient conversation, but was reasonable in this context where the object was a long lasting technique that could eventually overcome the loss of memory due to non intensive practice.

We could check the response time of the pupils to different types of ideograms. When the radical could be easily detected, the time essentially followed a Gaussian distribution. The variance was due mainly to the speed in counting the strokes and to the speed of recognition in the list. Remark that the question was not the recognition of the meaning of the ideogram (this could have been known in advance!) but the actual discovery in the index. A restricted group of pupils with previous experience in Chinese did perform somewhat better that the fresh pupils, but the improvement was statistically meaningful only in the case of Xingshu writing, where previous practice allowed a faster recognition of the single strokes.

In the case of two-radical ideograms (with unique entry in the dictionary) the time followed a bimodal distribution according to the practice in finding the leading radical. In more complicated cases the distribution was almost uniform until the limit time.

The reading of ideograms already known depended heavily on exercise and practice, hence it was meaningful on the practical side, but did not particularly add information on the best reading strategy. To achieve this aim some students prepared reduced indexes of known ideograms, with fast access; this solution shortened the response time, but heavily depended on the phase of the study, since personalized indexes would require a continuous updating, during both the learning phase and especially the latent (i.e. forgetting) phase.

[48] We chose a tourist guide of Roma [2], where at page 73 it was particularly interesting the translation of the Pantheon (*of all gods*) that in Chinese becomes *Temple of Ten thousand Gods*.

The discussion with pupils of mathematics showed that their strategy for finding an ideogram seemed to satisfy the rules of Bellman's dynamic programming[49]. Remark that according Bellman different strategies arise if you want to maximize the probability of solving the problem in a given time or if you want to minimize the average time required. An improvement (somehow related) is to learn from the beginning some tricky ideograms, such as the above mentioned *wo* and *cheng*, coming from the radical *ge*.

After two years we met some students that did not practice any longer. Most ideograms were forgotten, but the capability of alphabetization with dictionary was essentially preserved, with a time for complicated ideograms comparable to the original. On the contrary simple ideograms or radical ideograms, where practice prevails, took a time much longer, unless it is used a dictionary centered on the total number of strokes like [20]. The ways how the sets of foreign words are learnt and later on forgotten could be of some interest and should be compared with the studies begun by Jakobson [9].

A dictionary that for each character refers back to its radical (as it is done for a limited number of words in [11] in its section about writing) would help the training phase. Incidentally we recall also a problem that is teasing for foreigners: in some cases (numbers, demonstrative adjectives) a noun must be preceded by a suitable classificator. This is a word lexically connected, but redundant (actually you know that it is required, so you could skip the ideogram, if you have courage enough). Anyhow when speaking you require it, and many different classes of words require different classificators; there are some fuzzy rules, but it would be much better to find in the dictionary the choice of classificators together with the noun. Unfortunately this does not happen; in Italian books only Abbiati's grammar ([2], pp. 329-341) gives a list (though not exhaustive), but its use is difficult and thus restricted to professional students.

Conclusions

Nowadays analphabetism must be divided into absolute analphabetism and relative analphabetism. Relative analphabetism means that one code is not known, but some other code is known. Provided a computer can translate from a code to another, and input and output are available, the result mainly depends on the possibility of conversion between different codes. If the output code allows a sufficient discrimination, or post processing allows to solve ambiguities, the required knowledge of the original code is highly reduced.

In the case of printed Chinese ideograms a good snapshot, followed by OCR, should be able to give sufficient information about the translation of a single ideogram, at least as much as the indexed dictionary can give.

[49] In particular the famous problem ([4] Chapter I, 49) about a fuzzy search of a document in an office, originally due to Mosteller

Absolute analphabetism would mean that one is not even able to understand the "translation", that is that no code exists rich enough to translate. In practice this would be the situation of a Chinese analphabet who cannot read pinyin or in general a translation written with Latin characters. Really there was always a phonetic possibility of representation, using the first half of a known character and the second half of another known character, hence reducing the set of characters required to represent a Chinese text, and this was an old time expedient. Remark that the big number of homophones (compare fig.2) may require a revision founded on the original ideographic code. Bisyllabic words reduce this need, but they cannot totally solve the problem. The availability of machine reading and translating means that soon indexed dictionaries and personal capabilities of their use could be partially overcome in everyday texts, even if the actual problem of translation will remain relevant as it happens with all automatic translators. As a matter of fact some capability of reading the original text may still be required in order to be able to translate correctly, but the computer can supply many helps, since the couple ideogram *plus* pinyin allows to find out easily the word in its proper place in the dictionary, so that the recognition of the ideogram is restricted only to homophones (usually no more than ten).

The practice in alphabetization allows to perform this last operation much faster than a non skilled pupil, hence the technique we have tested seems to be highly recommendable as a useful complement to practical courses of introduction to Chinese, since the time it requires does not overcome ten per cent of the total time of each lecture, and the acquired capabilities can be preserved even with little exercise.

Summary

In pictorial data bases there are two main problems: the creation of a suitable (non alphabetical) organization and the way of accessing images stored in the data base starting from their image. When pictorial images belong to an alphabetical code the first problem is solved *a priori* and the second is referred as OCR (optical character recognition). Since there exist many alphabets and many variants, and sometimes images are blurred, the problem is not always straightforward. In some cases the alphabet becomes huge, or even is not recognized as an alphabet: this happens with Chinese ideograms, whose number is very high (more than 2000 are officially used in newspapers). Moreover handwriting adds severe problems of recognition.

A human reader that is originally analphabet of Chinese has the same problems, and they can be solved much in the same way as computer programs recognize handwritten characters. It is easy to find out the dictionaries that allow a search and those that are not structured for this aim.

With this information, both methodological and specific for Chinese language, an experiment of alphabetization was made with a group of pupils of the Excellence School of Udine University, during a practical first course of Chinese language. The

description of this work is the main object of section 4. At the end some conclusions are drawn, in view of an extension of the experiment to larger classes of pupils. We added some considerations about follow up, that show that the technique of indexed alphabetization is much more stable than the pure learning by heart of Chinese characters. In particular, unlike pure memorization, it allows fast recovery after latent periods.

Bibliography

1. Abbiati, Magda. (1998). *Grammatica di Cinese moderno.* Venezia: Cafoscarina.
2. AA.VV. (2005). *Luo Ma* [*Rome Guidebook in Chinese*]. Roma: Lozzi Editore.
3. Batini, C.; & Ceri, S.; & Navate, S.B. (1992). *Conceptual Database Design.* Redwood City: The Benjamin/Cummings Publishing Company.
4. Bellman, Richard E. (1957). *Dynamic Programming.* Princeton: Princeton University Press.
5. Biasco, Margherita; & Mao, Wen; & Banfi, Emanuele. (2003). *Introduzione allo studio della lingua cinese*. Roma: Carocci.
6. Chang S. K. (1973). *An Interactive System for Chinese Character Generation and Retrieval.* IEEE Trans. Systems, Man and Cybernetics, 3, 257-265.
7. Eco, Umberto. (1975). *Trattato di semiotica generale.* Milano: Bompiani.
8. Frost, R. (1986). *Introduction to knowledge based systems.* London: Collins.
9. Jacobson, Roman. (1944-1971). *Il farsi e il disfarsi del linguaggio* (Italian Compilation that extends *Kindersprache und Aphasie*). Torino: Einaudi. 149-163.
10. Kalay, Y.E. (1989). *Modeling Object and Environments.* New York: John Wiley& Sons.
11. Kantor, Philippe. (2005). *Chinese with ease volume 1.* Assimil France.
12. MacCormick, John. (2012). *Nine Algorithms That Changed the Future.* Princeton: Princeton University Press. 31-44.
13. McEliece, R.J. (1977). *The theory of information and of coding*. Reading, MA: Addison-Wesley.
14. Masini, Federico; & al. (2006). *Il cinese per gli Italiani.* Milano: Hoepli Editore.
15. Piccinini, Livio C.; & Rinaldis, Francesca. (2008). *Alle radici della Scuola Superiore di Udine.* in L.C. Piccinini Ed. *Superiore. La parola agli studenti.* Udine: FORUM Ed. 5-9 and 60-63.
16. de Saussure, Ferdinand. (1915). *Course de linguistique general.* Paris: Editions Payot.
17. Shannon, C.E.; & Weaver. W. (1949). *The Mathematical Theory of Communication.* University of Illinois Press.
18. Shi, D.; & Gunn, S.R.; & Damper. R.I. (2003). *Handwritten Chinese Radical Recognition Using Nonlinear Active Shape Models.* IEEE Transactions on Pattern Analysis and Machine Intelligence 25. 277-280.
19. Wang, A. B.; & Fan, K.C. (2001). *Optical Recognition of Handwritten Chinese Characters by Hierarchical Radical Matching Method.* Pattern Recognition 34. 15-35.
20. Wang, Huanbao; & Wang, Jun; & Shen, Emei; & Ke, Baotai. (2006). *Dizionario Italiano-Cinese Cinese-Italiano* Foreign Language Teaching and Research Press / Novara: IGDA.
21. Zhang, Shihua. (2006). *Dizionario conciso Italiano-Cinese Cinese-Italiano.* Shanghai Foreign Language Education Press / Milano: Hoepli Editore.
22. Zhao, Xiuying; & Gatti, Franco. (1996). *Dizionario compatto cinese italiano.* Bologna: Zanichelli.

Livio C. Piccinini	Department of Civil Engineering and Architecture University of Udine, Italy E-mail: piccinini@uniud.it
Ting Fa Margherita Chang	Department of Civil Engineering and Architecture University of Udine, Italy E-mail: chang@uniud.it
Giovanni Tubaro	Department of Civil Engineering and Architecture University of Udine, Italy E-mail: giovanni.tubaro@uniud.it