AN APPROACH TO THE ABBREVIATION OF CODE-STRINGS FOR CHINESE CHARACTER IDENTIFICATION

Ryohei Kagaya and Yo Kobayashi

1. Introduction

In our Chinese character generation system, a unit of a Chinese character is specified by an I-rep which is a linear string of alternating stroke-identifiers (Table I) and operators (i. e., concatenators and pseudo-concatenators, Table II). When a pair of strokes (see infra) is given in an I-rep, an operator specified a way of concatenating two strokes in terms of the coincidence of the functional points of the strokes (see Fujimura & Kagaya, 1969 and also Fujimura & Kagaya, 1972). For practical purposes the I-rep tends to become too long when a unit consists of more than several strokes. For example, the I-rep of the unit "田" is /21S11P21C21X11E11/.

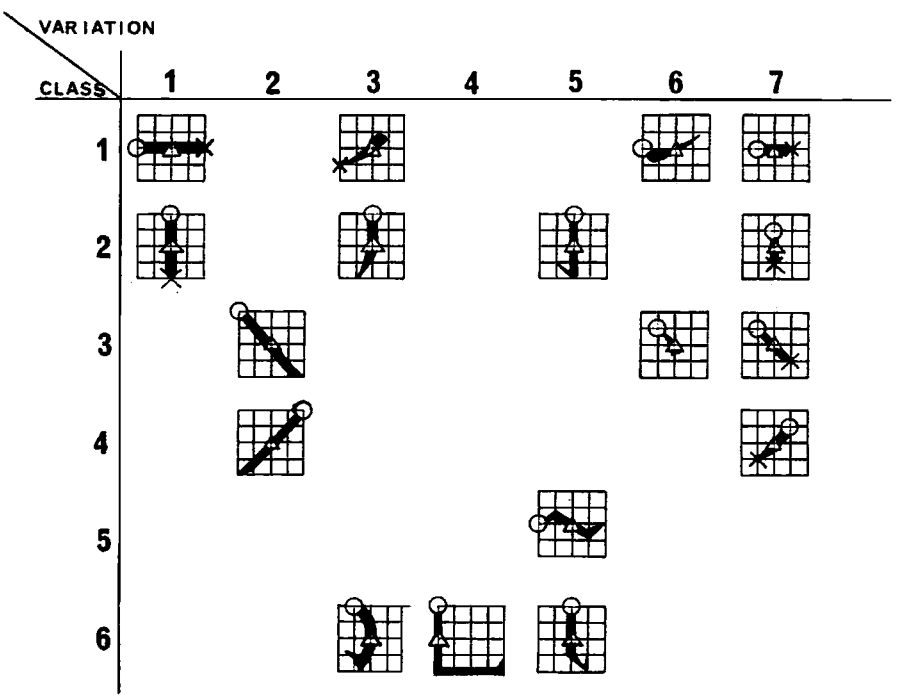

Table I: Stroke table. Each stroke is identified by a two-place number called the stroke-identifier, the first representing the class of the stroke and the second its variation. On each stroke pattern, "O" stands for the functional point α , the "Δ" the μ and the "X" the ω

General principles for the abbreviations of constituents of an I-rep have been discussed previously (cf. Fujimura & Kagaya, 1973 and Fujimura, 1973). In this paper, an actual method for abbreviating the I-rep wil be described.

In the following, we assume that, when a sequence of strokes is given, there is a most "natural" way of concatenating them. If, in the I-rep, a concatenation specified by an operator agrees with this "natural" concatenation, we may omit the operator from the code-string. An abbreviated code-string will be called the A-rep in distinction from the I-rep [1]. A set of rules is provided to insert appropriate operators at the pertinent positions in the A-rep and generate the corresponding I-rep. When the concatenation of a pair of strokes differs from that predicted by these rules an operator should be specified explicitly in the A-rep.

predecessor \ successor	α	μ	ω
α	S		
μ	C	X	A
ω	P	E	T

Table II: Concatenator defined in terms of coincidence of the functional points of the successor and the predecessor to be specified by the affinity convention.

2. Rules for operator insertion

A pair of strokes qualified for concatenation is said to have an "affinity" between them. The first member of the pair is called the "predecessor" stroke and the second the "successor". In an A-rep, every constituent stroke except for the initial stroke has a predecessor stroke. Except for the case of a "pseudoconcatenation" (see infra), the predecessor of a given stroke is selected as the last preceding stroke that belongs to a different subset of strokes [2]. They are concatenated with each other by one of the concatenators. For a pseudoconcatenator[3], the predecessor is

---

1) An I-rep is bounded by "/" (slash) and an A-rep by "<" (angled bracket).

2) Strokes are divided into two subsets: the odd-numbered class and the even-numbered class. Graphically, the former is composed of horizontal strokes and the latter of vertical ones.

3) The "-" superposes the "μ" of the successor on the "μ" of the predecessor (degeneracy). In the actualization process, the degeneracy is resolved and the degenerated strokes are separated into parallel positions, their spacings being determined by rule (See Fujimura & Kagaya, 1969 and also 1972). Another pseudoconcatenator is the "∅", which functions as well as the "-". Also it introduces an infinitesimal stroke(s) between the degenerated strokes. After the degeneracy is resolved, the infinitesimal stroke(s)

the stroke immediately preceding the pseudoconcatenator and the successor the immediately following one. In this case, we should always specify the pseudoconcatenator in an A-rep.

When, in an A-rep, a concatenator is not specified for a pair of strokes with an affinity between them, it is determined as follows:

(1-a) For the predecessor, the functional point to be occupied by the successor is selected in the preference order of α , ω and μ . If one or two functional points of the predecessor are already occupied by other stroke(s) preceding the successor, the next point in the preference order is selected for the concatenation. This preference order in the selection of the functional point is the same as that in "zeroing" (see Fujimura & Kagaya, ibid).

(1-b) For the successor, the functional point to be occupied by the predecessor is selected as the same (kind of) functional point as that of the "cardinal stroke" through which the "cardinal stroke" is concatenated to the predecessor. Here, the "cardinal stroke" is defined as the stroke last preceding the successor that has an affinity with the predecessor.

By means of a pair of rules (1-a) and (1-b), we can transform an A-rep into a corresponding I-rep. The rule described above is applied to the A-rep in the following way. The A-rep is processed from left to right. When a stroke not immediately preceded by an operator is detected, the rule is applied to determine the appropriate operator. When an appropriate operator is inserted, the following strokes are further processed. The above operation is cyclically performed till the end of the A-rep is reached. For example, consider the A-rep<21S112111 > .
In this A-rep, both a concatenator between the second stroke 11 and the third stroke 21 and a concatenator between the third stroke 21 and the last stroke 11 are not written.
The first stretch to be processed is " < 21S1121". α of the second stroke 11 (the predecessor of this pair of strokes) is occupied by the first stroke 21. Then the functional point of the second stroke 11 to be occupied by the third stroke 21 (the successor) is determined as ω according to rule (1-a). Also the functional point of the third stroke 21 to be occupied by the second stroke 11 is determined as α according to (1-b), since the first stroke 21, which stands as the cardinal stroke for this pair of strokes, is concatenated at α to the second stroke 11. Thus the concatenator between them is determined as "P". Secondly, "< 21S11P2111" is the stretch to be processed.

In this pair, the predecessor is the third stroke 21 and the sucessor the fourth stroke 11. The cardinal stroke is the second stroke 21. The concatenator between the third stroke and the fourth stroke is determined as "T" in a way similar to that shown in the above. Thus the A-rep " < 21S11 2 1 1 1 1 > " is transformed into its corresponding I-rep "/21S11P21T11/", which identifies the unit " □ ".

---

become(s) stretched and this bridges the two strokes. The kind of the infinitesimal stroke(s), i. e., vertical stroke(s) or horizontal one(s) and a pair of functional points of the degenerated strokes to be concatenated by the infinitesimal stroke(s) is determined by the rule of this system (see Fujimura & Kagaya, ibid).

The rule is effective for a pair of strokes which is not at the beginning of an A-rep. For an initial pair of strokes, the functional point of the successor (i. e., the second stroke in an A-rep) cannot be determined by the rule. We should always specify an operator for the initial pair of strokes in an A-rep.

If an operator between the two strokes differs from that determined by the rule, a pertinent operator must be specified in the A-rep, much in the same way as the feature specification in the systematic phonology according to the theory of markedness (cf. Chomsky & Halle, 1968).

For example, in the I-rep /11P21T11S64/ which identifies the unit "己", the operator "S" between the third stroke 11 and the fourth stroke 21 cannot be determined by the above rule, and we must specify the concatenator "S" between them in the A-rep. On the other hand, the concatenator between the second stroke 21 and the third stroke 11 is determined as "T" by the rule and we need not specify this operator in the A-rep. Thus, the A-rep for the unit"己" is encoded as <11P2111S64> .

Graphic pattern of unit	I-rep	A-rep
冂	/21S11P25/	<21S1125>
几	/23S11P65/	<23S1165>
刀	/11P25C23/	<11P2523>
口	/21S11P21T11/	<21S112111>
夂	/47C11P42X32/	<47C1142X32>
工	/11C21E11/	<11C2111>
己	/11P21T11S64/	<11S2111S64>
弓	/11P21T11S21P11P25/	<11P2111S2111P25>
曰	/21S11P21T11A11/	<21S11211111>
月	/23S11P25A11-11/	<23S1125A11-11>
用	/23S11P25C21X11-11/	<23S112521X11-11>
田	/21S11P21C21E11X11/	<21S1121211111>
目	/21S11P21T11A11-11/	<21S11211111-11>
皿	/21S11P21C21-21E11/	<21S112121-2111>

Table III: Examples of A-reps.

It may be considered that a graphic pattern represented by an I-rep into which an A-rep is transformed by applications of the above rules is more natural (or more unmarked) than those represented by I-reps which are generated by a random insertion of operators between each pair of the given stroke-sequence. For example, when the graphic pattern "冂" is

given and the horizontal stroke 11 is added to the "  ", the generated patterns "", "", "", "", "" and "" are marked while the pattern "" is unmarked. In fact, among the marked patterns, the first pattern is graphically nonsense, because the two horizontal strokes are completely superposed onto each other. Also, the third pattern to the last pattern are not acceptable as units or their parts. (The last pattern is really a part of the pattern "", but the "" is not considered as a unit. It is simply a pictographic sympol used in Chinese and Japanese.) The second pattern appears as a part of units such as "", but it itself is not considered as a unit. Only the unmarked pattern "" is considered as a unit; it appears in the compounded characters such as "". 4)

Furthermore, it may be considered that an A-rep with a greater number of operators is less natural (or more marked) for the identification of a unit. Generally, a unit can be identified by more than two I-reps. For example, the unit "" is identified by the I-rep /11P25C23/ or /11C23P25/.

The I-reps are transformed from the A-reps < 112523 > and < 11C23P25 >, respectively. Then we can consider the first A-rep to be more natural than the second. In fact, the stroke-order is the same as that of the traditional writing order of the character. 5) Thus, we can consider that some of the regularities of the traditional writing order are reflected in the rule.

Acknowledgements

We wish to thank Prof. Shigeru Kiritani, Institute of Logopedics and Phoniatrics, University of Tokyo, and Dr. Sumiko Sasanuma, Tokyo Metropolitan Institute of Gerontology, Section of Communication Research, for their penetrating discussion and improvements of this manuscript.

4) A larger (complex) unit can be "compounded" in terms of spatial arrangements of two or more units. The type of arrangement is specified by one of three "compounders" defined in our system (see Fujimura & Kagaya, ibid).

5) The traditional writing order is not completely systematic, but the following principles are generally observed: write an upper stroke before a lower one and write a left stroke before a right one. A central stroke is generally written last except for a few cases. Some of these principles are reflected in the concatenator table in that each conjugate of concatenators is excluded from the table.

References

Chomsky, N. & Halle, M. (1968). The Sound Pattern of English. Harper & Row, Publishers.

Fujimura, O. & Kagaya, R. (1969). Structural Patterns of Chinese Characters. Annual Bulletin (Research Institute of Logopedics and Phoniatrics, University of Tokyo), No. 3, 131-148. Also the paper was presented at the International Conference on Computational Linguistics, Sånga-Säby, Sweden, September 1-4, 1969.

Fujimura, O. & Kagaya, R. (1972). A Pattern-Structural Code for Kanji. First USA-JAPAN Computer Conference, 287-290.

Fujimura, O. & Kagaya, R. (1973). Coding of Chinese Characters and an Experiment on Character Generation. Reprints for Patan Ninshiki to Gakushukenkyu-kai (Special Interest Group on Pattern Recognition and Learning, The Institute of Electronics and Communication Engineers of Japan). No. PRL73-11.

Fujimura, O. (1973). Kanji no Kozo (Structures of Chines Characters). Gengo [Languages], Vol. 2, No. 7, 546-555. (in Japanese).

Kagaya, R. and Fujimura, O. (1969). Automatic Display of Chinese Characters Based on the Structural Description. Reprints for the Study on Automata, No. A69-60. (in Japanese).