

STRUCTURAL PATTERNS OF CHINESE CHARACTERS

Osamu Fujimura

Research Institute of Logopedics and Phoniatics
Faculty of Medicine, University of Tokyo

and

Ryohei Kagaya

Institute for the Study of Languages and Cultures of Asia and Africa
Tokyo University of Foreign Studies

Chinese characters, as used in Chinese and Japanese for orthography,* bear inherent meanings as well as sound shapes. Apart from these aspects, the graphic patterns of the characters also vary in complex ways and they appear very different over a wide range. It is obvious to native users of these characters, however, that the graphic patterns are mostly composed of different but frequently used subunits, and regularity is observed in the structures of character patterns. Quite frequently, a character is clearly composed of more than one character (with some minor modifications in shape). We can intuitively identify some commonly used strokes such as vertical and horizontal lines as constituents of characters or their subparts.

These obvious structural regularities have not been studied thoroughly, in spite of accumulated knowledge concerning etymology and historical developments of the characters and interpretation of the sound-meaning association.** This paper describes a way of describing the regularities of the Chinese characters as graphic patterns, without any explicit reference to the sound or meaning. The description may be considered as given in a form of generative grammar¹⁾

* The authors are native Japanese, and we are primarily concerned with the Chinese characters used in the contemporary Japanese orthography. There is obvious difference between Chinese and Japanese in the collection of character patterns used, but difference in the structural regularity itself is not apparent and remains to be studied.

** See e. g. B. Karlgren's classical work "Word Families in Chinese" (B. M. F. E. S. Vol. 5, 1934) and Grammata Serica (1957), and a more recent study by A. Todo, Kanji-no Gogenkenkyu (Etymological Studies of Archaic Chinese) (Tokyo: Gakutosha, 1963, in Japanese).

the patterns, with an important deviation from the concept of generative grammar. Namely, the system of rules given here generates character-like patterns, but it does not define the actual set of Chinese characters, viz. the accepted vocabulary of either Chinese or Japanese, but rather the set of patterns each of which could represent a Chinese character as far as the structural characteristics of a pattern are concerned.* The regularity described here is thus more like phonological regularities of lexical items of a language than syntactic regularities of its sentences. Nevertheless, the formal similarity of this rule system with the transformational theory of syntax is rather interesting.

The abstract representation of a character according to the generative rules make it possible to specify the patterns of essentially Chinese characters completely in terms of elements (strokes) and operators (concatenators and compounders). In other words, we can code Chinese characters by use of strokes and operators based on this framework of graphic theory of Chinese characters. Practical applications are of great import and interest, but these points are not of our direct concern in this paper.

This descriptive system was in essence proposed by one of the present authors several years ago.²⁾ Some modifications and additions have been made and are still being made, and generation of actual character patterns by rules are being tried with use of a digital computer with an oscilloscope display as the output device and a keyboard typewriter as the input. Some results are presented in this paper, and a demonstration is in schedule for the meeting.

Formation of Units

A unit is a separable subpart of a character. In our rule system, it is represented by a string of alternating strokes and concatenators. The form generated in this way is an underlying form called i-representation, and it is interpreted into s-representation—a conversion process described below in order to obtain the graph-pattern. When no element of the string in i-representation is left uninterpreted, and if the derived s-representation does not violate certain restrictional criteria performing filtering functions (see fra), the string of alternating strokes and concatenators represents

In this analogy to generative grammar, we consider the i-representation to be formed by concatenation of any alternating strokes and operators. Restrictions may be treated by filtering function of transformational rules that interpret i-representation in terms of s-representation.

a simple unit. A unit can represent a character by itself, or it may be compounded with some other unit(s).

1.1. Strokes and Operators

A set of strokes is given in the Stroke-table (Table 1). Each stroke is identified by a two-place number called "stroke identifier," and is defined by a stroke representation pattern. The first numeral of the stroke identifier represents the class of the stroke and the second the variation within the class. For example, the stroke "21", which is the first variational stroke of class-two, is defined by the stroke pattern as shown in Fig. 1.

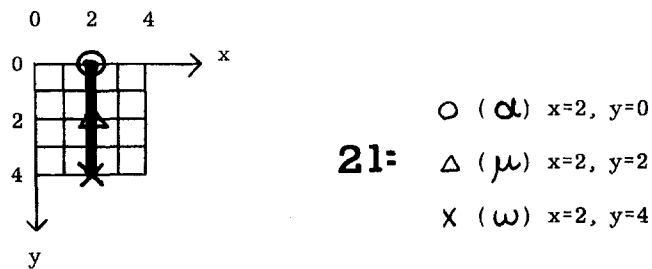


Fig. 1 - Stroke representation pattern

For each stroke, three functional points α , μ , ω are defined in terms of their x-y coordinate values in the stroke pattern field covering a range of integer values 0-4 for both x and y. A stroke with its three functional points can be represented by the following format (s-representation for the stroke):

[21; 20, 22, 24].

In this representation, the first number (21) is the stroke identifier, and the following three sets of numerals represent the x (on the left) and y (on the right) coordinates of the points α , μ , and ω , respectively.

Concatenators are listed in Table 2. A concatenator defines a particular positional interrelation between two strokes in terms of coincidence of a pair of the functional points. The set of strokes is divided into two functionally distinct groups, one for those with odd class numbers and the other for even class numbers, and a concate-

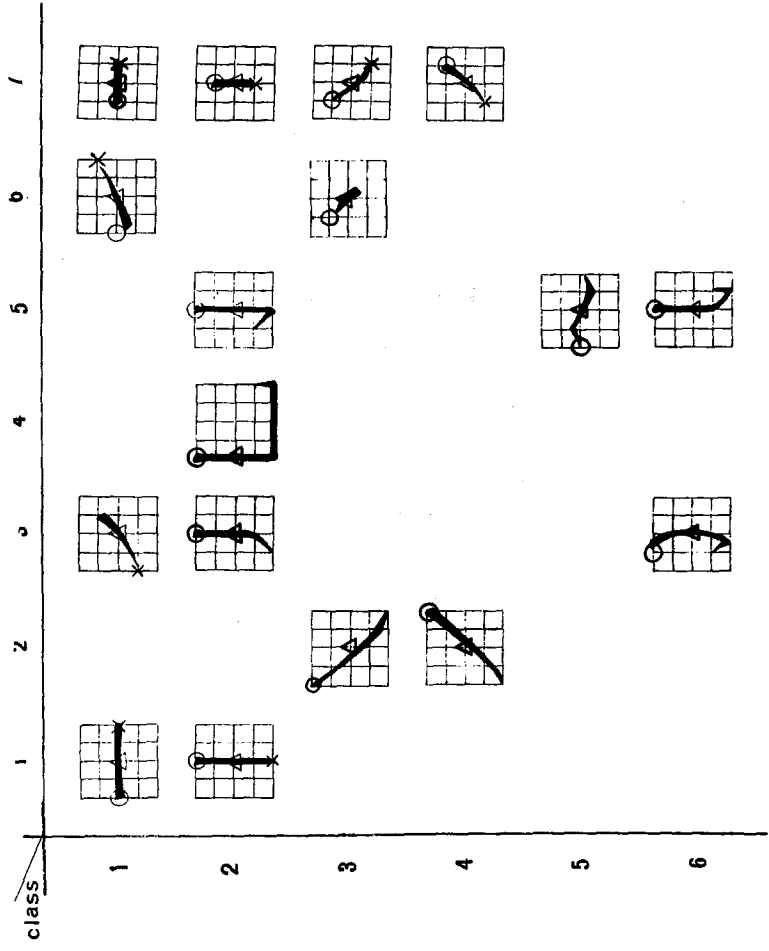


Table 1

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

Table 2 - Concatenators defined in terms of coincidence of the functional points of the preceding and succeeding strokes.

nator can combine only a pair of strokes of different groups. When a pair of strokes are qualified for concatenation, the pair of strokes are said to have "affinity" between them. For a given concatenator in i-representation, the pertinent pair of strokes with affinity is defined by a general convention of this rule system as the next stroke following the concatenator and the last preceding stroke that does not belong to the same group as the following stroke. The first member of this selected pair shall be called the "predecessor" of the concatenator and the second member the "successor." For example, in the string in i-representation /21S11P21C21X11E11/, the concatenator "C" operates on the fourth stroke 21 (successor) and the second stroke 11 (predecessor) skipping the more immediate stroke 21. Similarly, the last concatenator E concatenates the last stroke 11 to 21 rather than another 11.

1. 2. I-Representation vs. S-Representation

A string of alternating strokes and concatenators which shall be called "i-representation" (input representation) can represent an underlying form of a unit. The unit can be actualized as a character shape through executing some shape-adjusting rules and looking up the stroke table that stores the stroke representation patterns for all strokes. For example, the above mentioned string that represents a simple-unit character \boxplus is actualized as shown in Fig. 2.

The generated pattern can be represented by giving the x and y coordinate values of the three functional points belonging to all the constituent strokes. For the example above, the stroke positions are represented as:

[21; 00,02,04 11; 00,20,40 21; 40,42,42 21; 20,22,24
11; 02,22,42 11; 04,24,44].

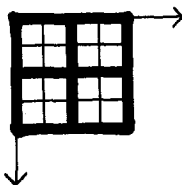


Fig. 2 - The pattern for
the character
/21S11P21C21X11E11/

This shall be called the "s-representation (stroke representation) of the unit," and it completely specifies the graphic pattern in terms of (abstract) functional interrelations of the constituent strokes.

1. 3. Degeneracy and Pseudoconcatenators


More than one stroke can coincide in position as specified in s-representation only when they are connected to each other through special operators, called pseudoconcatenators. The strokes thus interrelated are called degenerate strokes. There are two pseudoconcatenators, one designated by - (hyphen) and the other by \emptyset (zero).* A pseudoconcatenator always selects its nearest preceding and the next following strokes as the predecessor and the successor, respectively, and these strokes must belong to the same stroke-class. Any degenerate strokes must be concatenated (or compounded by a superconcatenator, see *infra*) to a stroke of the opposite group at the latter's μ in s-representation.** A string in i-representation that does not meet these conditions is blocked in generation and thus is rejected as a representation of a unit.


The pseudoconcatenators allot the same position (in terms of their μ (and often also α and ω consequently) coordinates in s-representation to a pair of identical or similar strokes. The degenerate strokes in s-representation are marked for the hyphen or the zero. The order of stroke occurrences is generally preserved in s-representation. Degeneracy of more than two strokes are not allowed. In the actualization process, as discussed later, degeneracy created by the hyphen is resolved and the degenerate strokes

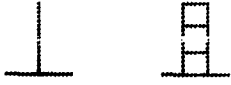
* In earlier reports of our study, we assumed the hyphen and a comma as pseudoconcatenators. The rule system is revised here.

** There are some further restrictions about the kind of strokes to be degenerate and those to be concatenated to degenerate strokes, and also about combination of these. Subclassification of strokes in this respect is still to be studied.

are separated into parallel positions, their spacings being determined by rule (see Fig. 3).


 a: /21S11C11-11P11T21/


 b: /21S11P11C11011/


 c: /2100021E11/



 d: /11C47P11S21C210021
 P21T11/

Fig. 3 - Examples of characters generated by use of pseudoconcatenators. (degenerate vs. resolved)

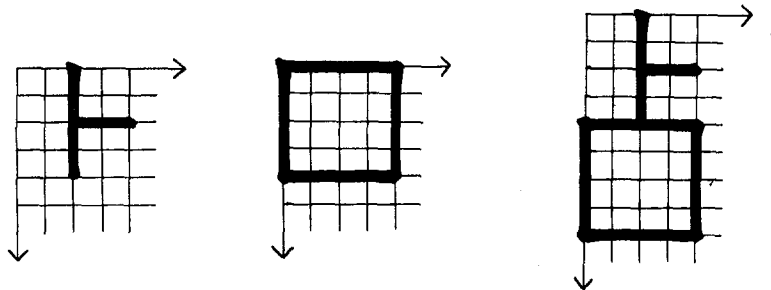
In the case of the zero, the strokes are separated in the same manner, but at the same time a special stroke of the opposite class (horizontal line for class-2 degenerate strokes and vertical line for class-1 degenerate strokes) is automatically introduced in s-representation. This additional stroke has an "infinitesimal length," and this bridges the pertinent two (degenerate) strokes. Where this bridge should be placed along the degenerate strokes is determined, after the unit has been completed in s-representation, according to a preference order that is given by convention of this rule system. The preference order for the selected point on the degenerate strokes is α , ω , and μ , but if a particular point shows coincidence with any other stroke(s), i. e., when the point is used as a junction in the pattern, this point is avoided and the point with the next degree of preference is selected for placing the infinitesimal stroke. The infinitesimal stroke becomes "stretched" when the degenerate strokes separate, giving an actual bridging between them. The infinitesimal stroke can be placed only at a point where the functional points (of the same kind) of the degenerate strokes coincide.

The zero can be used repeatedly in the same space between a pair of (degenerate) strokes in i-representation. Each symbol of zero inserts an infinitesimal stroke at the place of the highest preference that remains available. Examples for the use of degenerate strokes and the pseudoconcatenators are given in Fig. 3.

1. 4. The Dummy Concatenator "?"

A concatenator in general concatenates a stroke with another stroke. We introduce a dummy concatenator "?," so that we may concatenate a string with another string. The "?" in i-representation marks its immediately preceding stroke as the predecessor of

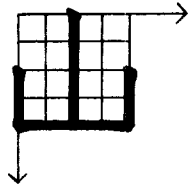
a concatenator that remains to be specified later in the string in conjunction with the selected successor stroke. In the string following this dummy concatenator, an extraneous concatenator must be found consecutively following another concatenator without a stroke identifier in between, and the second concatenator in the sequence selects the stroke marked previously by "?" as its predecessor stroke. The following stroke serves as the successor for both of the concatenators in pair, thus specifying a junction of two strings. For example, in the case of a unit represented by /21C17?21SE11P21T11/ (Fig. 4-c), the pattern /21C17/ (Fig. 4-a) is abutted to the second pattern /21S11P21T11/ (Fig. 4-b) through the concatenator E operating on the stroke 21 of the former and 11 of the latter. This operator "?" is convenient to form a unit according to the stroke order in the traditional handwriting.



a: /21C17/ b: /21S11P21T11/ c: /21C17?21SE11P21T11/

Fig. 4 - Concatenation of substrings by use of "?."

The example above could be generated by a string /21C17E11S21P21T11/ if only we disregard the tradition. Sometimes, however, the use of "?" is necessary for generating existing characters. The pattern of Fig. 5, for example, can be transcribed as /21?27PE11T27/, but there is no way to generate it without using the dummy concatenator, unless we define a new concatenator filling in the space in the concatenator table with so-to-speak a conjugate concatenator (in this case an E* that would select μ of the predecessor and ω of the successor for coincidence). Introduction of these conjugate concatenators is not desirable in consideration of the generalization of the rule system, because it expands the set of i-representations considerably without resulting in any additional acceptable patterns.



/21?27PE11T27/
 Fig. 5 - The use of "?."

The particular side of the diagonal in Table 2 is used in favor of the traditional stroke order.

1. 5. Restrictions on S-Representations

Some restrictions in terms of the generated s-representation have been stated in connection with the degeneracy and the infinitesimal stroke. There are some more restrictions of a general kind given in terms of the derived s-representation. These restrictions may be interpreted as filtering functions of the transformational process of actualization (see § 3).

One rather obvious restriction is that no strokes of the same class except degenerate ones can share the same set of coordinate values for any members of their functional points, whether they both are of the same kind (α , μ , or ω) or different. The convention of concatenation with the notion of affinity eliminates the possibility of generating two such strokes as a result of immediate succession of these in i-representation. A string, for example, like /11C21S11/, however, is permissible in i-representation but must be rejected by the criterion stated above.

Another possible restriction that may be imposed on an s-representation of a unit is in terms of the ratio of the largest dimension of the generated pattern to the number of strokes utilized. A threshold may be set and a pattern with a larger value of this ratio may be rejected, by use of an appropriate definition of length across a unit. This would exclude a long zig-zag of alternating 11 and 21, for example, from the set of acceptable characters.

A restriction of a more essential kind is probably in regard to the selection of a particular variation on the basis of contextual redundancy. It may well be the case that this kind of restriction is so strong that we can totally omit specifying the variation numbers of the strokes for the input transcription of any character. These points remain to be investigated.

2. Compounding of Units

More than one unit can be compounded to form a complex unit, which in turn as a unit can be compounded with another unit. The derivation of a character by a sequence of compounding can be

represented in i-representation by recursive use of pairs of parentheses, each surrounding a substring as a unit. A more illustrative representation may be given in a form of tree diagram, where the type of compounding is given by the compounder symbol at each node (see Fig. 6). For a unit to make a subpart of a character, it is

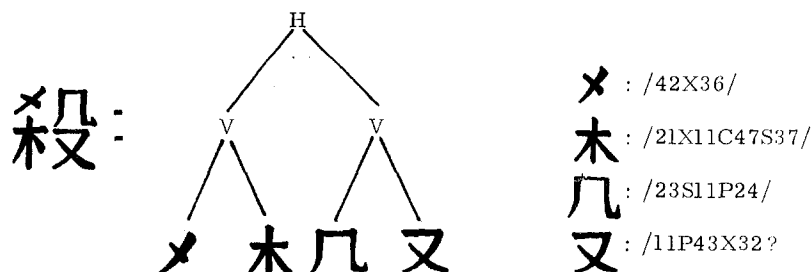


Fig. 6 - Complex compounding by use of appositional compounders.

in general necessary to go through a set of transformational rules that adjust the entire shape of the pattern to fit the context, as well as some special rules that makes minor changes in variation numbers of some strokes.

2. 1. Compounders

The compounder "H" can arrange more than one unit in a horizontal row, and the "V" can arrange some vertically. These two compounders form a class and may be called "appositional compounders." The character in Fig. 6 exemplifies a complex use of



Fig. 7 - Reduction of the last stroke.

the appositional compounders. There are many cases where the left subpart (hen) of the H compounding can be regarded as an affected form of a "free unit" whose "last stroke" is reduced in shape. Thus stroke 36 in Table 1 is a variation that serves as a reduced form of 32, and the stroke 11 becomes 16 in this context.* Fig. 7 gives a typical example.

As a special case, where the stroke 55 is identified as the last stroke of a unit used as the left constituent unit (hen) in H-compounding, this stroke undergoes a process of elongation, and the right constituent unit (tsukuri) is placed above the tail of this stroke (see Fig.



((11X21E11)V(21C17?47CE35))R(21C17)

Fig. 8 - Elongation of the last stroke in nyo.

8). Traditionally, the subparts (radicals) of this sort are called nyo.

In some cases similar to this, units serving as a subpart of a character cannot be identified as a transform of any "free unit," viz., a unit that can represent a character. Typical examples are those traditionally referred to as tare (the "appendants" or two-side



Fig. 9 - Examples of tare

Fig. 10 - Examples of kamae

surrounding radicals, see Fig. 9). These units, as well as the elongated unit nyo, have open space in which the other unit must be em-

* This is one of the phenomena that suggest redundancy of specifying a particular variation for a stroke class.

bedded. Another subclass of units that can embrace other units is called kamae (see Fig. 10).

These surrounding compounders are all represented by the symbol R in i-representation. The last stroke of the compounding unit (nyo, tare, or kamae) that follows the symbol R in i-representation tells where the preceding unit should be located in s-representation.

The third class of compounders consisting of X, C, E, S, and P is provided for cases where a stroke is superposed onto a unit in a special manner given by definition of the particular compounder.

Thus in the example given in Fig. 11, the compounder X places the vertical stroke 22 across the unit which itself is a V compound of two identical units, leaving the two ends of the vertical stroke sticking out.

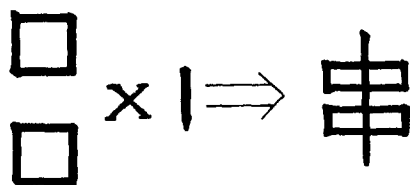
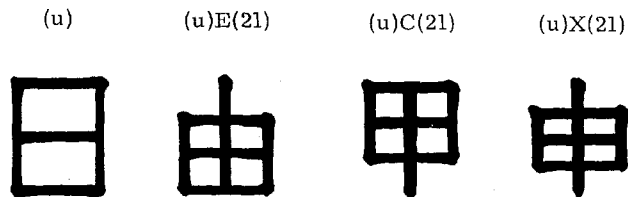


Fig. 11 - Superconcatenator X

The compounder C concatenates the point of the compounding vertical stroke (typically 21) at the point of the uppermost horizontal stroke, leaving the other end of the compounding

stroke sticking out of the lowest (most largevalued) y-coordinate of 's in the compounded unit. The compounder E, S, and P are de-



$$u = /21S11C11P11T21/$$

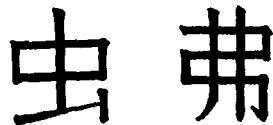
Fig. 12 - The stroke 21 with different superconcatenators.

finied in a similar manner reflecting the properties of the concatenators of the same names. Some examples are given in Fig. 12.

In this class of compounders, which may be called "superconcatenators," the compounding unit is typically a single stroke constituting a unit by itself. In some cases the succeeding unit is composed of more than one stroke, where only one of them can be desig-

nated as the "major stroke" that determines the manner of compounding. Variations 6 and 7 of all stroke classes and also all strokes in classes 4 and 5, and stroke 13 (Table 1) cannot serve as the major stroke. The major stroke can be degenerate. The superconcatenators act like concatenators in enabling the compounding (major

strokes to be degenerate in the case of C, X, and E (cf. 1. 3.). Thus the rejection of unconcatenated degenerate strokes has to be performed beyond the minimal unit, when the unit is preceded by a superconcatenator. Examples are given in Fig. 13.



a b

a: (21S11P21T11)X(21E16E27)
 b: (11P21T11S21P11P26)X(23-21)

Fig. 13 - The superconcatenator X with a compounding unit of more than one stroke.

In the compounding of the third class, the unit to be compounded may be collapsed in size in one dimension treated as though it were a degenerate group of strokes either horizontal or vertical. For example, in the unit (21S11C11P11T21) X (21), the compounded unit (21S11C11P11T21) could be regarded as a class-1 stroke. In this interpretation, it can be said that the superconcatenator in effect

acts as a concatenator of the same symbol. In a case like the unit (11) X (21C37S47), the pattern actually can be represented by a single unit 11X21C37S47, simply by removing the parentheses.

We may introduce another superconcatenator D, which is defined as a combination of C and E, namely a compounder that superposes a stroke which is "stretched" in such a way that both ends coincide with the two strokes, at the extreme positions in the compounded unit. This kind of compounded patterns can be generated in the rule system stated above by a succession of compoundings by use of the superconcatenators C and E.

2. 2. The Point Unit

The "point" designated by an apostrophe that follows a unit is an infinitesimal unit compounded to the preceding unit. It shows varied shapes in the actualized pattern, and a set of points is distributed in space in different prescribed manners depending on the context. Special rules are required for taking care of those seemingly varied phenomena, but technical details are still to be worked out. Typical examples are shown in Fig. 14. The examples are transcribed from

left to right as follows:

upper: (25)', (25)'', (25)''', (25)'''' ,
lower: (21S11P21T11)', ((21P11S11P63)X(11))',
((42X36)R(21S11P21T11))'

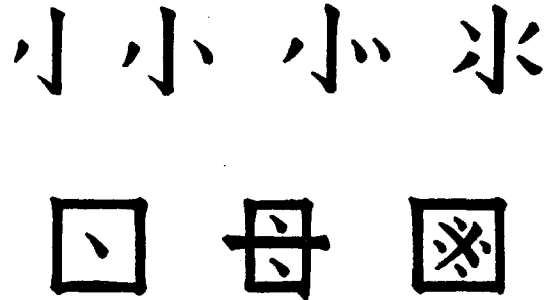


Fig. 14 - Actualizations of points in accordance with the context and the number of the points.

3. Actualization

A character is transcribed as a set of units combined through compounders in any depth of complexity. Each constituent unit is transcribed as a string in *i*-representation placed in parentheses. The *i*-representations of units determine their *s*-representations, specifying positioning of all occurrences of strokes in a frame of the pattern field. The frame is normalized and placed together according to the specification of the compounder to form a compound unit, and this process of normalization and abutting can be recursively repeated. The set of rules for normalization and stroke reduction (see supra) is thus cyclic in the sense of the cyclicity of phonological rules. The *s*-representation for a unit after the normalization no longer has the quantized coordinates. In the last stage of actualization of a character. Strokes shapes are called in form the stroke representation pattern into this generalized *s*-representation.

3. 1. Stroke Arrangement

It may be obvious intuitively that in the actualized form of any

character the constituent strokes are distributed in space somehow evenly. This fact can be accounted for by designing a later part of the actualization process to form a set of stroke distribution rules. As a general principle for this distribution of strokes in space, we may assume a potential field defined in the stroke pattern of each stroke surrounding the actualized shape of the stroke. We then may hypothesize that superposition of the potentials belonging to the distributed strokes in the finally actualized pattern results in a state of equilibrium by attaining the total potential energy minimum. In short, strokes exert repulsive force against each other, and the strokes can translate and be compressed within a given unit frame as long as the topological interconnections are not changed. The end points α and ω are always rigidly related to the actualized stroke shape, but the midpoint μ can shift along the line defined typically as a straight line connecting α and ω .

3. 2. Practical Approximation

A practical approximation for this principle of distribution may be devised as follows. Each stroke has a two-dimensional measure of spacial occupancy for x and y directions, defined in a stroke table. A "size normalization factor" of a unit is defined as the sum of these measures of occupancy of all the constituent strokes. The area which is occupied by each constituent unit in a complex unit is determined by the proportion in terms of the "size factor."

Within a unit, the actualized distribution of constituent strokes is attained in a similar manner, by allowing typically equal spaces between similar strokes in the direction perpendicular to the stroke line. An equally weighted space is allowed at the margin between the border of the frame and the outermost stroke. There are some details of the rules which will not be discussed here.

Some examples of characters are illustrated in Fig. 15. These were actually generated on an oscilloscope display of a computer by typing in the i-representations. The rules used for this practical approximation of the actualization process are only preliminary and some character patterns suggest necessary corrections of the program which can be mostly readily done.

4. Concluding Remarks

Many details are still to be worked out and some are simply not described here for brevity. It is obviously true that the same character can be generated by different i-representations, partly due to different stroke orders and partly due to different selection of variational shapes of strokes. Another sort of ambiguity is possible in

端	踊	陣	價
微	撤	燻	橫
潮	箱	賜	疑
彰	撤	濱	美
秒	紀	紅	胎
胃	青	要	貞
草	莊	香	勇
面	香	促	保
契	姿	峯	故
昨	昭	呈	枯
架	染	孟	柳
洋	活	祖	相
砂	研	佳	秋
卓	叔	和	供
卓	委	季	始
宝	彼	征	定
怪	拍	拒	徑
沼	放	明	易

some special cases depending on whether a compounder or a concatenator is used, as mentioned in 2.1. The use of degeneracy against compounding gives still another sort of ambiguity. Thus for example, the character can be generated either as a simple unit /110011X21/ or as a compounded form /(21S11P21T11)X(21)/.

These sorts of ambiguity have been to a large extent eliminated by some care taken in formulating the rule system, but some of them are interesting and seem to indicate the inherent problems concerning the nature of Chinese characters. The system we have given here is concrete and valid in fair details, but it is still subject to even major changes for improvement. The essential principle, however, seems to us convincingly effective for description of the graphical structures of the characters.

Fig. 15 - Oscilloscope display examples of computer-generated Chinese characters. All characters were generated by rule out of the input representation type in through ordinary keyboard.

SUMMARY

A system is proposed for specifying any one of the accepted patterns of Chinese characters, or similar patterns that could be used as Chinese characters. The system may be considered as a generative grammar of the set of character patterns. A unit is formed by concatenating strokes by operators. A set of strokes is given in a stroke table where three abstract functional points α , μ , ω , as well as a typical actualization form, are defined for each stroke. Concatenators and pseudoconcatenators are provided, each of them defining a particular positional interrelation between two strokes in terms of coincidence of the functional points. The set of strokes is divided into two functionally distinct groups, and a concatenator can combine only a pair of strokes of different groups, a pseudoconcatenator only those of the same group. Thus a string of alternating strokes and operators, which may be called the "i-representation," determines an underlying form of a unit, which can be actualized as a character shape through looking up the stroke table and executing some shape-adjusting rules. On the level of i-representation, more than one unit can be combined to form a more complex character pattern, by use of one or more of compounding operators that specify "transformational processes" to be executed before the shape adjustment process. Preliminary results of an on-line computer experiment will be shown where the actualizations of characters are displayed on an oscilloscope when characters are specified by typing in the i-representations.

REFERENCES

1. N. Chomsky: Syntactic Structures, The Hague: Mouton and Co., 1965.
2. O. Fujimura: "Some Remarks on the Character Recognition," Information and Control (Institute of Electrical Communication Engineers, Japan) No. 4, 2-7 (1963) (in Japanese).
_____: "Structure of Language and Coding of Chinese Characters," Kagaku (Science) 34, 321-324 (1964) (in Japanese).
"The University of Electro-Communications," Current Research and Development in Scientific Documentation (National Science Foundation Office of Science Information Service, U. S. A.) No. 14, 516 (1966).