SCIENCE OF THE STROKE SEQUENCE OF KANJI

Takeshi SHIMOMURA
Technical College of Osaka Prefecture
Neyagawa-shi, Osaka 572, JAPAN

Summary

The stroke sequence of kanji has been investigated chiefly from the viewpoint of energy in the dynamics of writing together with the informational viewpoint in memorization in learning. Sample characters include all the members for daily use, and also katakana. The results show that the standard sequence fundamentally follows the law of energy minimization in writing movements. The degree of satisfaction is highest for the vertical succession among practical writing conditions. The standard sequence is considered to originate from the human adaptation for circumstances. As for the characters with more strokes, it is found that to add the contribution of easy memorization by compression of information content of the sequence, the law is satisfied by selection of the sequence of sub-systems in place of individual strokes. These results indicate that the stroke sequence greatly affects the human ability of kanji processing.

1. Introduction

In general the standard stroke sequence is traditionally designated for each kanji(and also its offspring letters as katakana), and both in calligraphy and in primary education, writing characters according to the standard is imposed. But the unknown origin prevents justifying the observance. And also recently the stroke sequece has applications in engineering as kanji-recognition and so on. Actually, however, some varieties in existence pose some problems.

In view of these situations, clarification of scientific basis of the standard sequence will be not a little valuable to the above fields and also to serve as an aid to explicate the intrinsic nature of the character in linguistics.

In this report the considerations from the viewpoint of energy in the dynamics of writing is extended, and the study of contribution to easy memorization, the possibility of which was already suggested by the author as an additional factor for the characters with more strokes[1],are developed, by aiming at the total research on the traditional standard sequence.

2. The Stroke Sequence on Hypothesis

In writing kanji, if the shape alone is enough, the direction and the sequence are various and therefore a great number of ways of writing produced by their combinations will be possible. What is the reason for a specified sequence selected as the standard ? By daily experiences, it seems to the author that according to the standard sequence "easy, rapid and beautiful" writing is possible to be accomplished. Now, putting aside the factor of beauty for the moment, it can be thought that "ease and rapidness" means the standard sequence is one requiring as little energy consumption as possible in writing movements. And therefore this study began with building up a hypothsis that the standard sequence satisfies the law of energy minimization and verifying by electronic computation for the kanji samples with fewer strokes.[1] As for kanjis with more strokes, in addition to the contribution to writing movements, the standard sequence is assumed to have an effect to ease the memorization of kanji, i.e. a contribution to reduce information content, which is studied by the use of the theory of information.

3. Modelling of the Writing Movements and the Rank Distances

If the directions of strokes are conventionally fixed, kanjis with n-strokes have as many possible sequences as n!. The energy consumed in writing depends on pen-path length, pen velocity, pen pressure and so on. For simplicity, assuming that velocity and pressure are constant and the effects of direction and others are negligible, the energy is reduced to a function of pen-path length alone. Pen-path consists of stroke vectors and stroke-with-stroke combining vectors for a single character, and in case of character succession, character-with-character combining vectors add to this.

Sample kanjis used number 1850 for daily use. The character form employed is of square-style and each stroke is approximated by a straight line. As circumstances for character, an appropriate field is assumed and the situations of singleness and character successions in three different directions are introduced by the boundary condition setting.

The process of verifying the hypothesis is to calculate the energy consumption in all the possible sequences for normalized standard kanji samples and to examine whether the energy consumption for the standard sequence is in the lowest.

The results of electronic computation show, as a whole, the hypothesis is fairly well satisfied, as an example of which the degree of satisfaction is shown in Tab. 1 by the rank distance D for samples of not more than 6 strokes, where D is defined as

$$D = [(k - 1)/(n! - 1)] \times 100 \ (\%)$$

when the standard one is in the k-th from the lowest in energy. Among these, the relationship of D and cumulative number of characters N (%) for 6-stroke characters, for example, is shown in Fig. 1. As for katakana, the perfect satisfaction D = 0 holds good for more than 60% of all the samples in the vertically downward succession condition and the anisotropy is quite small.

In Tab. 1, all the samples of not more than two strokes are independent of either singleness or succession, and of its direction, and completely optimized, which fact is suitable to the nature as the most fundamental constituent, together with katakana, in kanji system. Though with the number of strokes per character increasing, some small spreading of D and anisotropy appear, the whole trend can still be seen to support the hypothesis. Here it is noticeable that human ability of selection through the cumulative experiences is splendidly high: in spite of the number of possible sequences n! abruptly increasing with increment of strokes, perfect satisfaction is found in quite a few samples.(In case of 5-stroke samples, for instance, in which n! = 120, optimum holds good in more than 20% of them !)

Some difference in the degree of satisfaction by diffrent environments is also observed. The degree is highest for vertical succession among practical writing conditions, which corresponds well to the traditional kanji calligraphical modes in past China and Japan.

The history of the standard sequence has been rather stable, in connection with the past stable modes, in which,however, some examples changed exist. And the inspection of their D's indicates that the transitions are mostly towards the lowest. New phenomena observed now about the sequence are mostly related to the recent change in circumstances. These facts assure that the stadard sequence can be regarded as an example of human adaptation for circumstances.

These characteristics of the standard sequence agree with the general features of the natural language as a social custom. (Though the above-mentioned considerations were performed under the condition of the fixed directions of stroke vectors, the direction itself is clarified to follow the law of energy minimization by introducing direction dependence in energy consumption per path-length.)

4. The Structure of Kanji and Memorization

Some small D's spreading phenomena with increasing strokes per character suggest an additional factor existing. With this respect, in the previous report, possible relevancy to facilitation of memorization, i.e. the effect on reducing information content, was just point out.[1] Here this factor is examined, with correlation to the kanji structure, by the aid of the information theory.

Information for writing kanji is assumed to be input to/output from the memory device of the cerebrum as a symbol string of kanji-forming stroke vectors. In the following calculations, encoding only about direction is employed for simplicity, putting aside position and magnitude. Quantization is in 8 different directions according to the traditional calligraphy. (The case in which simplified to 5 directions is also considered.)

At the first stage, the amount of information by the statistcs of the direction occurrence frequency, multigrams, etc. about all the samples is calculated, and possibility of data compression is analysed by the theories of Markov process and encoding.

By the results, mean information content per stroke by the transition probability in Markov chain, is found to be only about 10% less than that by frequency statistics. Great reduction of information amount is, therefore, unpromising as far as a stroke is the string element even if the transition probability is learned.

Next consideration is on compressbility of data by deviding stroke string into sections, i.e. reduction by forming a kind of supersymbol,[2] viewed from transinformation between strokes. At the same time, as a graded structure is observed in the stroke symbol string, transinformation content between compound events and possibility of compression are also calculated by setting each previous section as an encoding element, in which a close coupling relation is found between them, such as the transinformation amount between compound elements of 3-strokes, is about ten times as much as in case of sectioning. These results suggest a possibilty of compression to about a half or less.

And then, therefore, the reduction of information content for all the samples by selecting suitable subsystems like the traditional radicals, etc. as compound element, is examined. With this respect, however, besides our calculation a similar study for other purpose has already been reported.[3] With some different viewpoints included, it is thought sufficient to cite here instead, for the estimation of this. By their data, the whole information amount can be reduced to as small as 40% of the amount in case that the individual stroke is the element.

In view of these facts on information,the satisfaction of the law of energy minimization in the dynamics of writing is examined again by unit of sub-system. That is, by expressing sub-system in terms of a cummulative vector, the similar calculations to the previous chapter are performed. And the results prove the law is well satisfied, for instance, for about 450 samples composed of two sub-systems, each of which is again the member of samples, more than 99% of all are optimized in the vertical succession condition.

The system of kanji is, by origin, of a graded structure, and most subsystems like radicals have important symbolic functions such as phonetic value, meaningful element, etc.. Therefore, as for characters compound in structure, the stroke sequence determined by selection of sequence of subsystems has an effect to fully functionate these symbolic actions, by which generating additional redundancy in human processor is expectable.

By the above considerations, it is clarified that for a simple character with fewer strokes, the standard stroke sequence is determined by the energy minimization in writing and that a compound character with more strokes are of multiplex structure, where the energy minimization is satisfied by selecting the sequence of memorization-facilitating sub-systems whose stroke sequences have previously been decided by the energy minimization.

5. Conclusion

This investigation leads to the following conclusion: the traditional standard stroke sequence of kanji is thought as a human experiential result toward the optimization of writing and memorization in learning. And the sequence, therefore, greatly affects human ability of linguistic activities using kanji.

6. Acknowledgements

The author would like to express his cordial thanks both to Emerit.Prof. E.Sugata and to Prof. Y.Inuishi of Osaka University for their useful advice and encouragements throughout this investigation. Sincere appreciation is also shown to Prof. S.Ijichi of our college for his support and facilities, and to my students for their experimental assistance.

References

1) T. Shimomura: A Scientific Approach to the Stroke Sequence of Chinese Characters, Trans.I.E.C.Japan, 58-D,12,756 (1975)

2) F. von Cube: Ueber ein Verfahren der mechanischen Didaktik, Gr.K.G. 2, 1 (1961)

3) T. Sakai, M. Nagao,and H. Terai: A Description of Chinese Characters Using Sub-patterns, Johoshori(Journ.I.P.S.Japan),10,5,285(1969)

Tab. 1 Average Rank Distance D (%)

number of strokes	number of samples	conditions			
		singleness	succession		
			diagonal	vertical	horizontal
1	1	0	0	0	0
2	5	0	0	0	0
3	21	16.2	13.3	15.2	17.1
4	34	18.5	11.0	15.1	20.2
5	58	12.5	7.5	10.1	13.8
6	73	10.4	5.3	8.7	11.7

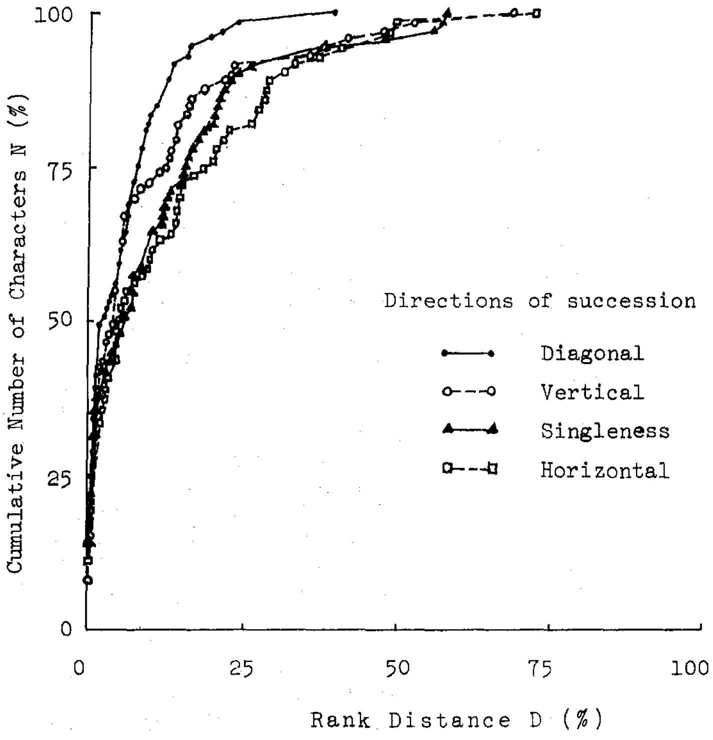

Fig. 1 Cumulative Number of Characters and Rank Distance for 6-stroke Characters