PROGRAMMING OF CHINESE CHARACTERS FOR MECHANICAL TRANSLATION

by

NATHAN SIVIN

Submitted in Partial Fulfillment

of the Requirements for the

Degree of Bachelor of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June, 1958

Signature redacted

Signature of Author . . . . . .

Department of Humanities, June 3, 1958

Signature redacted

Certified by . . . . . .

Thesis Supervisor

Signature redacted

Accepted by . . . . . .

Chairman, Departmental Committee on Theses

PROGRAMMING OF CHINESE CHARACTERS FOR MECHANICAL TRANSLATION

by Nathan Sivin

ABSTRACT

In order to develop an operational method for the input of non-alphabetic Chinese characters to a digital computer, it is necessary to find a means of deriving unique numerical sequences from the structure of the characters themselves. An examination of the structural characteristics of printed Chinese characters shows that the most satisfactory concepts from the standpoint of automatic recognition are the stroke and stroke order, provided that the latter is redefined as a scanning procedure. Modification of the recently developed Chinese Photocomposition Machine, which now uses strokes and traditional stroke order as the basis of binary code input by a Chinese operator, is a practicable solution to the problem. Adaptation of the Chinese Photocomposition Machine to mechanical translation computer input is discussed, and characteristics of a computer-scanner for completely automatic input are outlined. The proposed method may also be used for computer input of Chinese characters as they occur in Japanese and Korean texts, and for military applications of the Chinese Photocomposition Machine.

PREFACE

" . . . a language reputedly invented by the devil to prevent the spread of the Gospel in the Middle Kingdom."

--Achilles Fang

In order to concentrate his reprehensible calligraphy in one location, the author has used the unfortunately ubiquitous Wade-Giles romanisation within the body of this paper. The occurrence in the text of a number between slashes (e. g., /54/) is an indication to the reader to turn to Appendix A, where the characters referred to will be found next to the corresponding number.

The author wishes to express his gratitude to the following persons:

To his thesis supervisors, Professor Victor H. Yngve, Modern Languages Department, MIT, and Professor William D. Stahlman, Humanities Department, MIT, for their stimulating advice and criticism;

To Professor Samuel H. Caldwell, Electrical Engineering Department, MIT, and Robert G. Crockett, Graphic Arts Research Foundation, Inc., for their patient introduction to the intricacies of the Chinese Photocomposition Machine;

To Professor E. Neal Hartley, Humanities Department, MIT, who originally proposed the possibility of a thesis topic within the field of mechanical translation;

To Dr. Alfred K'ai-Ming Ch'iu, Librarian, Harvard-Yenching Institute, for valuable bibliographical guidance;

To Achilles Fang and Wu Pei-Yi for data on Chinese lexicography;

To Dr. T. Su for information on Chinese stroke order;

And to Professor William N. Locke, Modern Languages Department, MIT, Professor Erwin Reifler, University of Washington, and Professor P. A. Boodberg, University of California, for their help and encouragement.

All data in this study concerning the Chinese Photocomposition Machine have been verified by Professor Caldwell, and are released with his permission.

TABLE OF CONTENTS

CHAPTER PAGE

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The art of applying automation to the translation of languages is still in its incubatory stage. Many questions, foreseeable easily enough in theory, have not yet been touched upon in practice. This study deals with one of them.

The Problem

The purpose of this paper is to establish an operational method by which Chinese characters may be translated into uniquely determined successions of binary digits, which may be fed into a digital computer for such purposes as mechanical translation. The projected product of such an operational method is (1) a set of simple instructions so phrased that a person having no prior acquaintance with the Chinese language can, by following them to the end, derive a unique and reproducible binary code for each character, which may be typed into the computer, or (2) an analogous electronic program which will allow a correctly designed computer-scanner device to act as input, thus obviating any human intermediary between the printed Chinese text and the finished translation in the output language. Only the method itself will be derived here.

Qualifications of Terms

It is advisable to indicate immediately the delimitations of usage of certain words in this study:

Translation. In the word "meaning" lurks many a trap for the unwary; its use will be eschewed for the most part except in the most mundane sense.

Translation is fundamentally the transfer of information from one language to another.[1] It is a highly selective process, in which a large part of the information which the original text is capable of yielding is sacrificed for the sake of that part which the translator requires, or is at any rate satisfied with. "How may we compare what a sentence in English may mean," asks I. A. Richards, "with what a sentence in Chinese may mean?"[2] If one restricts oneself to denotation, to the range of "dictionary meaning," the sentence is usually sufficient context for understanding of a word.[3] But for connotation, the context is the original writer's whole knowledge and understanding of his own culture.[4]

Translation is criticism, perhaps the most rigorous mode of criticism. Every translator must make his peace with what he considers worth bringing over, and with what he can bring over. The former is almost invariably conditioned by the latter.

---

[1] Anthony G. Oettinger, "The Design of an Automatic Russian-English Technical Dictionary," Machine Translation of Languages, A. D. Booth and W. N. Locke, editors (Cambridge: The Technology Press, 1955), pp. 49-50.

[2] I. A. Richards, "Toward a Theory of Translating," Studies in Chinese Thought, Arthur F. Wright, editor (Comparative Studies in Cultures and Civilisations, No. 1; Chicago: The University of Chicago Press, 1953), p. 149.

[3] V. H. Yngve, "Syntax and the Problem of Multiple Meaning," Machine Translation of Languages, A. D. Booth and W. N. Locke, editors (Cambridge: The Technology Press, 1955), p. 209. This is not true in Chinese, where tense, number, etc., must often be inferred from previous context.

[4] N. Sivin, "A Primer of Chinese Poetry" (Cambridge: Humanities Department, MIT, 1958), pp. 30-36, passim. (Mimeographed).

That is to say, there are degrees of translation, from the masterpiece of scholarship and empathy to the beginning student's crib, from Chapman's Homer to an anonymous English rendering of a Russian technical paper. These degrees depend not only on the translator's ability to use the dictionary, but on his success in handling the problems of connotation and style. The measure is in large degree subjective and qualitative, and does not lend itself to quantitative expression except in a very general way.[5]

Mechanical translation. Just as the range of subjects that a human translator can capably handle is limited by the extent of his knowledge and by his particular abilities, so the capabilities of a computer for translation are limited by inherent factors. Its function is that of a coding device, which, given a word (or a sentence) in one language, will supply its equivalent in another. Its memory is large but limited by engineering factors. It can translate a word only insofar as meanings, and instructions for looking them up and spewing them out in the proper order, are supplied to it beforehand.

The simplest application of the computer to translation is its use as an automatic dictionary, where its only function is that of looking up words. A sentence is typed at the input, and is converted into a sequence of numbers. The machine, following a program of instructions, selects the sequence of numbers which represents the corresponding word or words in the

---

[5] G. A. Miller and J. G. Beebe-Center, "Some Psychological Methods for Evaluating the Quality of Translations," Mechanical Translation, III (December 1956), p. 80.

output language. At the output, that numerical sequence is converted to a printed sequence of letters, a literal, word-by-word translation. Where a word in the input language has more than one possible translation, either the most general equivalent, or preferably a list of enough equivalents to ensure semantic coverage, is provided. This method is capable of crude, fairly understandable translations which may be used for rough screening of foreign texts.

Even this step is still far from realisation. The labor required to construct an automatic dictionary is immense; limitations in the size of present-day computer memories require extreme care to cover the range of meaning with a minimum number of output language equivalents for each input language word. Dictionary projects are underway in several of the major languages at the moment (Russian and German are receiving most attention in this country) but, except for a few demonstrations of very restricted scope, mechanical translation is not yet an actuality, nor is it likely to be for a matter of a decade or more.

When word-by-word translation becomes practicable, two further problems which affect comprehension must be solved in order to increase the value of computer application: multiple meaning and word order. Both of these are presently under attack; their solution, which is indispensable for adequate readability, involves syntactic analysis of languages, a new linguistics based on operational rather than descriptive grammar. The value of such an investigation in its own right is obvious.[6]

---

[6]Anthony G. Oettinger, "Mechanical Translation--Fact and Fancy" (lecture

Discrimination in matters of connotation and style depend on more than mere memory. It is possible that, when developed far enough, the computer will be able to perform analogous discriminatory functions, but such functions would involve a memory of astronomical proportions. It is for this reason that present-day workers have limited themselves to preparing for the translation of subject matter in which connotation and style play the smallest possible part--scientific and engineering papers.[7]

Concentration on technical publications is worth while on other counts. There is a shortage of competent professional translators, and their services come high. The demand for translations is becoming more acute as the function of research in industry becomes more important; this demand justifies the large investment required for the development of mechanical translation. The relatively crude translations available during early stages of the process will still be valuable for rough screening of foreign publications. Further, limitations in the size of present-day computer memories may be mitigated by the use of "micro-glossaries" of technical vocabulary peculiar to the field with which the material to be translated is concerned. Such "micro-glossaries" may be inserted into and withdrawn from the machine's memory at will. They allow adequate vocabulary coverage with a minimal glossary, and simplify the multiple-meaning problem; e. g., only the equivalent "hyperbola" need be provided for the French

---

given at Harvard University, February 27, 1958).

[7]Warren Weaver, "Translation," Machine Translation of Languages, A. D. Booth and W. N. Locke, editors (Cambridge: The Technology Press, 1955), p. 20.

"hyperbole" in a mathematical "micro-glossary."[8]

Digital computer. While work in mechanical translation has, to date, been influenced by the concrete existence of general-purpose high-speed electronic computers, such as the American IBM's and the Russian BESM, such machines do not fully take into account the special requirements of translation. A machine designed for that particular purpose will be forthcoming only when research in the field has developed to a point which justifies the expenditure involved. It would match sequences of digits by direct comparison rather by comparatively involved computations, and would be better equipped for dealing with sequences of different length, and for quick search of a greatly expanded memory.[9]

For the purpose of this paper, the term "digital computer" will be taken to include both the machines now in use and those which may be developed in the future to fit the specific needs of mechanical translation. It is unlikely that any changes made will materially affect the applicability of this study.

Operational definition, etc. The word "operational" will be used here to indicate validity in terms of machine operations.

More generally speaking, "operational" refers to the series of actions

---

[8]Victor A. Oswald, Jr., "Microsemantics" (paper read at the first MIT Conference on Mechanical Translation, June 17-20, 1952).

[9]Paraphrased from Victor H. Yngve, "The Technical Feasibility of Translating Languages by Machine" (paper read at the A. I. E. E. Fall General Meeting, Chicago, Illinois, October 1-5, 1956), p. 3.

involved in doing something. An operational grammar of a language would categorise the speaker's choices in moving from intention to expression, rather than concerning itself with more abstract general description. An operational grammar would be of value to mechanical translation insofar as, rephrased in terms of actual machine operations, it would enable the computer to recast the information in the input language sentence into a "good" comprehensible sentence in the output language.

An operational definition, then, is restricted to mechanical parameters. "A character is the basic discrete unit of meaning in the Chinese language" is not an operational definition. The following is a far from perfect operational definition of a character: "A character is an area the average density of which is greater than that of the paper on which it is printed, which may be inscribed in a rectangle each side of which touches an area of black, and the exterior of each side of which is bounded by a margin of the density of paper. It is separated from all other areas of like density by a clear border of at least x microns in a horizontal direction and y microns in a vertical direction (depending upon type face). Exceptions: Erh /1/ and ch'uan /2/ are single characters, and punctuation marks (examples would have to be given) are not characters."

A set of operational instructions clearly defines the actions to be taken in a certain process. Analogous to the program of a computer, it (1) is given in a series of imperative sentences, (2) covers every possible exigency, and (3) is so phrased that the desired process will be completed when it is followed to the end. The following set of instructions, for instance, might be used to separate a bagful of mixed marbles and candy

drops, of which only the marbles need be saved:

I. If there are no objects in the bag, stop. If there are, take one object from the bag, put it into your mouth, and bite on it.

II. If the object cracks, chew it up and repeat step I. If not, go on to step III.

III. If the object remains in your mouth, remove it to a receptacle and repeat step I. If not, repeat step I.

The Chinese language. Even ignoring the multitude of territorial dialects, there are, so to speak, three Chinese languages, each of which, by our standards, has very little to do with the others. These are the spoken language, pai-hua /3/, predominantly polysyllabic and highly redundant; the classical literary style, wen-yen /4/, mainly monosyllabic, rhetorical rather than grammatical, and extremely condensed; and the modern writing style, pai-hua-wen /5/, a development of the post-revolutionary period, which is an amalgam of the traditional writing style with the vernacular. The latter is used in all modern publications, and it is with its translation that this study will be concerned.[10]

Some comparison of the three styles is advisable in this paper, and will be made in apposite detail in Chapter II. All references to the spoken language there, and to pronunciations throughout this study, will take into account only the Peking dialect, the purest form of the so-called

---

[10]Quotations from earlier documents in the classical style, which would pose horrifying complications, appear often in general writings, but very seldom in scientific publications.

National Dialect, kuo-yü /6/.

Value of the Study

The economy which justifies mechanical translation depends significantly upon speed and maximum economy at the input stage, best attainable by a scanner input.[11] But scanner inputs for this purpose have yet to be developed. If they are still not available by the time mechanical translation of Chinese becomes feasible, the services of a human operator will be required. The method outlined in this paper will effect savings even in that case, since it will enable any typist of average intelligence to perform the input function with little or no special training. By taking advantage of the applicability of the already developed Chinese Photocomposition Machine, which is likely to become standard in its field, the solution proposed here can be put to use with a minimum of new design expense.

Two functions are indispensable to successful translation: intelligence and memory. The device most suitable for mechanical translation, the digital computer, incorporates various mechanical or electronic systems which may be said to remember. It is not intelligent, but by remembering and applying a program of instructions devised by intelligent humans, it performs functions which correspond, within rigid limitations, to the operation of the intellect.

The digital computer performs its tasks by juggling numbers. In

---

[11] A. D. Booth and R. H. Richens, "Some Methods of Mechanized Translation," Machine Translation of Languages, A. D. Booth and W. N. Locke, editors (Cambridge: The Technology Press, 1955), p. 26.

order to translate from one language to another, it is necessary that the information in both the input and output texts be expressible as a sequence of numbers.

One obvious solution to the problem of programming (i. e., numbering) is to simply number the words in the computer's vocabulary in alphabetical, or even random, order. Each word in the text is assigned its corresponding number, which can then be typed into the machine. With a vocabulary of significant size, however, the time required for looking up the numbers is prohibitive, and this method is useless.

A better solution is available for alphabetic languages. If the letter be taken as the basic encoding unit rather than the word, no look-up operation is necessary. The operator can type the text out on a typewriter keyboard, each key of which transmits a characteristic sequence of binary pulses to the machine. A scanner performing the same function would have to make correspondingly fewer comparisons between the units being scanned and those in the machine's glossary. While this method requires considerably more storage space in the computer's memory than the former,[12] it is superior practically and theoretically; practically

---

[12] The number of binary units (bits) required to represent a unit of input is equal to the logarithm to the base 2 of the number of possible different units of input. If the result is not a whole number, the next larger integer is taken. For a vocabulary of ten thousand words, fourteen bits are required for each word. Thus, if the word be taken as the unit of input, 140,000 bits are required for the input glossary.

Five bits are needed for each letter of the English alphabet. With an average word length of seven letters, 350,000 bits are required for the input glossary if the words be indexed by their spelling. The disparity

because a high input rate is required for a high output rate, and theoretically because the internal ordering of the word (i. e., its spelling) is preserved in its numerical code.

But Chinese characters are complex visual entities; they are not susceptible of reduction to combinations of a small number of recurring, linearly ordered graphic elements. In order to take advantage of the potential economies of mechanical translation, some structurally determined breakdown, analogous to the breakdown of words into letters, is indispensable. This study, by a consideration of the structural properties of Chinese characters, proposes such a breakdown, compatible with both human and electronic programming techniques.

Two objections to the importance of this study will likely occur to anyone even slightly familiar with Chinese affairs:

1. The dearth of original scientific and technological publications in Chinese does not seem to justify the large investment required for mechanical translation from that language.

2. The campaign which aims to introduce an alphabetic written language threatens to make the character obsolete.

The first objection is a shortsighted one. It is only a matter of time, certainly less than a generation, until the breadth of fundamental and

---

becomes less as the size of the vocabulary increases, but the inverse difference in input rates becomes greater.

This discussion ignores storage requirements for interword spacing, punctuation, and for syntactical programming and definitions in the output language which would accompany entries in the input language glossary.

applied research which is concomitant with industrial growth results in a substantial and important scientific and engineering literature in Chinese. In terms of the enormous labor required to compile an automatic dictionary and construct the syntactical grammar required for unambiguous, readable sentence-by-sentence translation, a Chinese-English project started today would be far from premature. It is unlikely, however, that such a project will be commenced until a time when the finished product is already needed. Most projects in this country are concerned with the Russian language, because it is the language in which the military is most willing to invest its research funds. The Russian Academy of Sciences, however, is not so constrained by the bonds of exigency.[13]

The theoretical objections to official alphabetisation of Chinese need not be recapitulated here; suffice it to say that the practical problems involved in the replacement of one writing system by another so different are sufficiently formidable that a carefully planned program extending over a period of at least one generation is indispensable. Romanisation is not, after all, a new idea; it was first proposed by Fang I-Chih /7/ about 1650.[14]

The most commonly expressed objections to the use of characters are:

1. Since each character must be memorised, the achievement of literacy is unnecessarily onerous.

---

[13] D. Panov, "On the Problem of Mechanical Translation" (trans. Friedman and Halle), Mechanical Translation, III (November 1956), pp. 42-43.

[14] L. Carrington Goodrich, A Short History of the Chinese People (New York: Harper Brothers, 1951), p. 228.

2. The written language in its present form is poorly adapted to the communication of scientific ideas.

3. Characters are old-fashioned.

The first two objections are most frequently entertained by Westerners, and are probably responsible for the prevalent misunderstanding of reports published in 1956 which stated that Communist China had adopted an official romanisation system.[15]

The former low rate of literacy was due partly to economic conditions and partly to the peculiar demands of the traditional classical education. The classical education is no more; the success of widespread adult education in China is evidence that learning to read takes no longer for a Chinese than for anyone else.

The limitations of characters as a mode of conceptual expression are simply limitations of the language and its philosophical background in general.

The third objection, puerile as it sounds, is probably the real motive for the Chinese government's experiments with alphabetisation; in China all traditional ways are open to comparison with more modern, pragmatic methods. The view that that experiment does not amount to a "campaign" is supported by a recently released official Chinese communication, which

---

[15] "Revolution in Language," Time, LXVIII (July 9, 1956), pp. 25-26; and Henry R. Lieberman, "China's Linguistic Revolution," The New York Times Magazine, January 8, 1956, p. 78. A grossly inaccurate and misleading article also appeared in the Saturday Review in 1956. It attempted to prove the curious hypothesis that the purpose of "the romanisation campaign" was to stamp out individualism in the Chinese populace.

announces the adoption of the 26 letters of the Latin alphabet, and states that this alphabet will be used for transliteration, annotation of characters in dictionaries, telegraphy, indexing, and popularisation and standardisation of the Peking pronunciation. " . . . Chou En-lai confirms this statement as to the extent to which the alphabet will be used." No mention whatever is made of its being taught to the populace in general, nor of the prospect of its replacing characters.[16]

Limitations of the Study

In formulating a solution to what is fundamentally an engineering problem, one is confronted with a choice between using existing mechanical devices, originally developed for a different purpose, or designing machinery specifically for the task at hand. The latter possibility is beyond the author's capabilities and resources, but the former is amply justified in this case, since the Chinese Photocomposition Machine, presently under construction, can be simply and economically adapted to the purpose. Mechanical details of the necessary adaptation are not within the scope of this paper, but the principles are set forth in sufficient detail to serve as a guide to their working out.

Since the problem was approached with the idea of formulating a solution that would be applicable to a scanner input, the method proposed, while compatible with that end, seems cumbersome for use by a human operator

---

[16]The Chinese document is quoted in the China Bulletin, Far Eastern Office, Division of Foreign Missions, National Council of Christian Churches, USA, VIII (February 3, 1958). None of the listed uses is new; the import of the statement is that several Cyrillic letters are being discarded from the previously announced romanisation system.

performing the input function. That problem may never arise, since it is quite probable that a scanner input will be perfected by the time mechanical translation of Chinese becomes feasible. The principles of a practical, speedy input by Chinese operators have already been worked out in connection with the invention of the Chinese Photocomposition Machine. There is no dearth of Chinese who can easily be trained to perform this function. The method proposed here for input by a human operator will be of primary value in certain operational military applications where use of an operator trained in Chinese is unavoidable.

A complication arises in that structural characteristics of Chinese characters--the types of strokes and their relative positions--vary with different type faces and even type sizes. The same problem arises with other languages, and must be dealt with partly by accomodations in scanner design (or a short period of familiarisation on the part of a human operator), and partly by having a different program for each type face or size. In Chinese scientific journals which the author has seen, however, a uniform type face is used.

Previous Approach

There exist two published treatments of the problem, a practical one in Russian and a speculative one in English.[17] In both, the Chinese telegraph code is mentioned as the basis for input. The telegraph code is simply a

---

[17]Panov, op. cit., p. 43; and Erwin Reifler, "Report on the First Conference on Mechanical Translation," Mechanical Translation, I (August 1954), p. 32.

loosely ordered list of characters numbered from 1 to 9999; the time and skill involved in looking up numbers makes that method unwieldy for any except experimental use. Reifler reports that a device has been made which provides characters when their numbers are fed into it, and suggests the possibility of inventing a machine to perform the inverse function. There is, however, nothing in the Chinese character, taken by itself, from which its telegraph code can be derived. With that qualification, this paper may be regarded as carrying out Reifler's suggestion.

Organisation of the Remainder of the Thesis

To one concerned with turning Chinese characters into numbers, the multiplicity of characters dictates that the numbers be determined on the basis of the structure of the characters themselves.

The body of the thesis is divided into two chapters. The first is a consideration of the Chinese character as a structural entity. The evolution of its form is briefly noted; its relation to the spoken language, and the relation of printed to written characters, are examined with a view toward explaining the extent of the differences, much greater than in Western languages, and establishing structural considerations which are peculiar to the printed character. Structural characteristics--strokes, stroke order, and "entities" composed of recurring combinations of particular strokes--are defined and discussed.

The second chapter is an examination of several possible methods by which a unique numerical sequence may be derived from a Chinese character. The central thesis is then set forth as it arises from the principle upon

which the recently invented Chinese Photocomposition Machine is based; viz., if the strokes occurring in Chinese characters are divided into a limited number of characteristic types and these types are numbered, by writing down the number corresponding to each stroke in the character in the same order as that in which the strokes are written, one derives a numerical sequence which, except for a very few ambiguities, uniquely defines that character.

The use of this principle in the Chinese Photocomposition Machine is explained briefly, and other features of the machine noted. The elements of the machine which are applicable to computer programming are explained.

It has been found that the traditional stroke order used in writing characters is sufficiently ad hoc that it cannot be completely formalised by uncomplicated operational means, and cannot be quickly and accurately used by anyone without a considerable Chinese education. The use of a simple and rigorous synthetic stroke order, which may be used by a Western operator or an electronic scanner, is proposed. Necessary modifications to the Chinese Photocomposition Machine are discussed, as are the characteristics required of an appropriate computer-scanner. The results of an analysis of five hundred characters to determine the average size of the minimal number sequences resulting from the use of a synthetic stroke order, and the frequency of ambiguities which occur (see Appendix B), are evaluated.

The final chapter contains a summary of the proposed method, a recapitulation of its limitations and suggestions for their amelioration, and a note on other applications.

Appendices include a key to Chinese characters referred to or represented by romanisation in the body of the paper; a description of the

method by which the application of the synthetic stroke order was experimentally tested on a large number of frequently encountered characters; and a table of stroke types used in the Revised Four-Corner System, one of the entity systems discussed in Chapter III.

CHAPTER II

THE CHINESE CHARACTER AND ITS STRUCTURE

Chinese is the only remaining logographic language of any wide currency. It evolved from crude pictographic and ideographic inscriptions to its present etymologically complex brush-written form. Each character is composed of as many as thirty-five brush strokes, of a number of distinctive types, written in a traditionally prescribed order. Certain graphic entities composed of two or more strokes constantly occur within characters. This chapter, by an examination of these structural characteristics, lays the groundwork for a structurally determined program for the input of characters to a digital computer.

The Evolution of the Chinese Character[1]

The most satisfactory explanation of the origin of Chinese characters is that they developed from signs used in divination. The oldest existent characters are engraved on tortoise shells and on pieces of bone; they are pictographic and ideographic on a fairly literal level. As written communication developed, the use of representational drawings, and combinations of them to express abstractions, was found to be too limited, and other means of character formation became important. Already existing characters,

---

[1]Primary references for this section are Ch'en Kung-Che /8/, Ch'u-Hsüeh Shu Fa /9/ (Elementary Calligraphy) (Taipei: Hua Kuo Publication Society, 1952), pp. 4-20; Chiang Yee /10/, Chinese Calligraphy (Cambridge: Harvard University Press, 1938), pp. 18-36; and Wan Wing-Sum/11/, Lynx-Eyed Calligraphist, Specimen of Chinese Calligraphy (Vancouver, B. C., 1954).

for instance, were borrowed to represent homophones and gradually lost their original meaning.[2] The most prevalent type of character came to be a combination of a usually pictographic classifier part (usually called the radical), which denotes the class of meaning to which the character was assigned, and a phonetic part, a homophone of irrelevant meaning which indicates the pronunciation. Kou /13/, the character for "dog," is made up of the classifier ch'uan /14/, which is used for certain small animals (and for adjectives like "beastly") and the phonetic chü /15/, which occurs in many other characters of the same, or similar, pronunciation. As in this case, a character which is not pronounced like its phonetic is no uncommon phenomenon; pronunciations have changed greatly in the more than two thousand years since the characters were formed.

The chief reasons for the shift away from pure pictographs and ideographs are clear:

1. The radical system was a characteristic manifestation of the Chinese tendency toward categorisation.

2. The phonetic system established an ordered relation to the spoken language that would be absent in a purely logographic written language.

3. The new means were simple and practically inexhaustible. Use of recurring entities simplified the stylisation of characters, supplying an at least partly logical basis for learning.

---

[2] The character ch'i /12/, as an example, was originally a pictograph for "basket." It was taken to represent the homophonous pronoun "his, her, its," and is no longer used in the former sense.

The currently accepted system of radical classification, used for arranging characters in dictionaries, includes 214 radicals. Earlier systems established as many as 560. The radical concept is cumbersome in practice, and does not hold up as an objective means of classification. Its use is gradually being abandoned.

Phonetics have never been systematised by Chinese, but tables of phonetics have been prepared by W. Simon and W. E. Soothill. Both are useful but arbitrary.[3]

The greatest stylistic revolution in writing was the ascendancy of the brush as the main writing implement (Chou dynasty, B. C. eleventh-third centuries). Use of the brush tended to limit the lines of which characters are composed to a relatively few specific types. Each type of stroke was formalised, and tended to be visually separable from the others to which it was joined, at least in the less abbreviated writing styles. The number of strokes in a given character became fixed, as did the order of their writing. Even to the present day, although the Western pen has in large measure supplanted the brush, the basic writing techniques have not changed. The only recent development of major importance has been the large scale introduction of simplified characters to take the place of the most complex common characters.[4]

---

[3] W. Simon, Chinese Radicals and Phonetics (London: Lund Humphries and Company, Ltd., 1944), pp. 109-257; and Henry C. Fenn, The Five Thousand Dictionary (Cambridge: Harvard University Press, 1942), pp. xiv-xix. Simon uses 545 phonetics, Soothill 880. Neither system purports to be exhaustive.

[4] "Spisok Uproshchennykh Kitaiskikh Ieroglifov" (List of Simplified Chinese Characters), Narodnyi Kitai (People's China), VI (1956), supplement, pp. 3-6.

Relation of the Written Language to the Spoken Language

The Chinese written language is separated from the spoken language by a stylistic gap which has no counterpart in Western languages. Since the character does not contain within it a fixed indicator of phonetic value, its meaning has not been greatly affected by the evolution of the vernacular. The emphasis on tradition in education resulted in a generally unchanging literary syntax very much removed from the more grammatical, predominantly polysyllabic colloquial dialects. The written language is more monosyllabic, because the visual specificity of the character obviates the redundancy seemingly occasioned by its narrow range of possible pronunciations.[5] The modern literary style is often referred to as colloquial, but it is so only by contrast with the old classical style--a uniquely economical, poetic language, rhetorical rather than grammatical.[6]

The old wen-yen /4/ literary style was, in its inception, formally correlative with a spoken language of considerably more phonetic complexity than the present-day Mandarin.[7] But the gradual simplification of pronunciation, and consequent grammatical sophistication, of the vernacular did not result in corresponding changes in the written language. The logical frame-

---

[5]Marcel Granet, La Pensée Chinoise (Paris: Éditions Albin Michel, 1950), p. 54. There is a discussion of the syllabic basis of the written language in H. A. Gleason, Jr., An Introduction to Descriptive Linguistics (New York: Henry Holt, 1955), pp. 305-306.

[6]Achilles Fang, "Some Reflections on the Difficulty of Translation," Studies in Chinese Thought, Arthur F. Wright, editor (Comparative Studies in Cultures and Civilizations, No. 1; Chicago: The University of Chicago Press, 1953), pp. 282-83.

[7]L. Carrington Goodrich, A Short History of the Chinese People (New

work of the written language was more or less fixed on a not very flexible level, but the accessibility of the Confucian classics to the average educated man was assured. The perpetuation of the classical style was largely due to the Imperial examination system (abolished in 1904), which tested mastery of the classics and the art of elegant composition as prerequisites to official service.[8]

The wen-yen style survived many revolutions and modifications; its end came only with the demise of the traditional, humanistic Confucian system. In confronting the West, China, already debilitated by 270 years of Manchu domination, was forced to the realisation that if it continued to rely on the old order it would become a third-rate nation. Liberal sentiment was for Western ideas, and favored radical reforms which, it was hoped, would allow China to take its place among the Great Powers. It was in this spirit that Dr. Hu Shih /16/, a disciple of John Dewey, advocated in 1916 a new written style patterned on colloquial Mandarin--easily learned by any Chinese, and suited to the introduction of foreign ideas.[9] His pai-hua-wen /5/ was soon almost universally adopted as the coin of written expression. While quotations from older wen-yen works occur frequently, especially in literary and philosophical works, for

---

York: Harper Brothers, 1951), p. 14.

[8]The profession of bureaucrat, it is necessary to note, was reserved for the scholarly class, and was the most highly honored in China.

[9]Hu Shih /16/, "Wen-Hsüeh Ko-Ming Yün-Tung" /17/ (The Literary Revolution), Current Chinese Readings, Wang Chi-Chen /18/, editor (New York: Bookman Associates, 1950), pp. 208-213.

the modern reader they are like Latin quotations in English--one has learnt them or one hasn't.

But the modern written language is still far removed from the colloquial. Its formal resources are based on those of wen-yen, and it is considerably more compressed and less redundant than the vernacular. A Chinese might be unable to understand a newspaper headline read aloud to him unless certain substitutions and interpolations were made, a process which amounts to translation.

For this reason, a Chinese input based on transliteration of the text would be at best uneconomical and at worst impracticable. The semantic basis of the written language is graphemic, not phonemic; transliteration would occasion an intermediate translation process in order to resolve otherwise insurmountable ambiguities.

Chinese seems, however, to be uniquely suited as the input language for experiments in mechanical translation of speech.[10] The narrow range of syllable pronunciations, while increasing the semantic redundancy of each morpheme, lessens the occurrence of near homophones. More important, elision and slurring are almost absent even in fast speech.

It is possible that the input problem for spoken Chinese could be most simply treated by recourse to a Braille-type transcription, the

---

[10]For general background of mechanical speech translation, see William N. Locke, "Speech Input," Machine Translation of Languages, A. D. Booth and W. N. Locke, editors (Cambridge: The Technology Press, 1955), pp. 104-118.

principles of which were worked out by W. H. Murray.[11]

Relation of the Printed to the Handwritten Character

While printed and handwritten characters are superficially very similar, an operational treatment of their structural characteristics must take into account the very different circumstances of their production. For this reason, only a small part of our knowledge of the structure of script characters may be applied to the problem of programming printed texts.

In comparing the fully written out, or k'ai-shu /19/ character with the ordinary printed character, impressed from movable metal type, it is obvious that the two are cognate, although occasional disparities show that the latter is derived from an earlier version of the former. The individual strokes of which each is composed are very similar in form, and certain entities may be identified which recur from character to character. In the case of written characters, these components may easily be defined--the stroke operationally, the entity statistically. A stroke is that deposit of ink laid between the time that the brush (or pen) is applied to the paper and the time that it is next lifted from

---

[11] S. M. Russell, "An Explanation of Mr. Murray's System for Teaching Illiterate Sighted Chinese," Constance F. Gordon-Cumming, The Inventor of the Numeral-Type for China by the Use of which Illiterate Chinese Both Blind and Sighted Can Very Quickly Be Taught to Read and Write Fluently (London: Downey and Company, Ltd., 1899), pp. 147 ff.

the paper.[12] An entity is a combination of successive strokes which occurs in the total glossary with a frequency greater than a certain prescribed minimum. Within the printed character, however, the idea of successive strokes is meaningless, since at least a page of characters is imprinted at one time.

In both the script and printed forms, the character may be visualised as inscribed within a square. In both forms, a great deal of attention is paid to an aesthetically balanced composition of the component strokes and entities, and to perfectly formed individual strokes.[13] This does not mean that each stroke of a type is shaped exactly like all others of the same type. In written characters, one of the marks of a good calligrapher is that no two specimens of a character in close proximity are exactly alike. In printed characters, that would be impossible in view of the typesetter's burden, but a given entity is sometimes differently designed when it appears as a large part, than when it appears as a small part, of a character. For instance, the radical p'u, which appears on the right or bottom of many characters, takes one form when it is a large part of the character /21/ and another when it is a small part /22/. The design of a character also sometimes varies with type size. In the most common type face, the character

---

[12] This defintion holds only for the K'ai-shu /19/ character. In a more abstract style, such as T'sao-shu /20/ ("grass-writing"), the whole character may be written in one application of the writing implement. The characters in Appendix A are written in K'ai-shu style.

[13] "Balance within the square" is one of the basic principles of composition in calligraphy.

wai takes one form in large type /23/ and another in small type /24/. It would be difficult for an unsophisticated machine to identify /21/ and /22/, or /23/ and /24/, as alternate forms of the same entity.

The concept of stroke order is a very important one in calligraphy. Tradition has decreed a fixed order in the writing of the strokes which make up each character.[14] Certain variations in the work of various calligraphists have made unanimous agreement on one correct system of stroke order impossible. While any treatise on calligraphy will give a few loose guides ("Start at the upper left and work toward the lower right . . ."), no set of rules has ever been enunciated which even approaches a complete description of even the most popular system. Unsettling though it may be to the rigorous Western mind, the rationale of stroke order is in the last analysis an aesthetic one and, except where general trends hold, the stroke order of each character must be individually learned. Systematisation of the traditional stroke order would be too involved a procedure to yield a useful product.[15]

The process of learning stroke order is only partly inductive; there are too many cases where memorisation is required. The only criterion of correct stroke order given by Chinese teachers, in the author's experience,

---

[14] The number of strokes per character ranges from one to thirty-five (or more), although the maximum for the most common characters is about twenty-five.

[15] Statement by Achilles Fang, personal interview, November 23, 1957.

is "It looks right." Experimentation has failed to yield more operational criteria of any generality.[16]

Structural Characteristics of the Written Character: a Recapitulation

There are several concepts which are useful in an analysis of the structure of written Chinese characters:

Stroke. The stroke is the basic operational unit, formed by one application of writing implement to paper. Chinese authors disagree as to the number of different stroke types.[17] In any enumeration, a few of the types differ only in relatively minor proportions.

Entity. An entity is a combination of strokes which may be seen to occur as a part of many characters. The concept is extremely broad, not to say vague, as will be seen in the next chapter. In order to be useful as a basis for systematisation, it must be limited by specifying (1) a minimum frequency of occurrence in the total collection of characters, and (2) a minimum and maximum size (in terms of number of strokes) of the entity itself. A simpler procedure, and the one ordinarily used by

---

[16] A detailed analysis of the Chinese Character Exercise Book (Monterey, Calif.: Chinese-Mandarin Dept., U. S. Army Language School, 1955), a workbook giving the stroke order of 600 characters, was attempted. The author was early forced to conclude that a "grammar of stroke order" would be too complicated to be useful; the analysis was abandoned. It was found, however, that one of the bases of stroke order is the flow of strokes in the more abstract writing styles.

[17] Ch'en, op. cit., pp. 51-53, notes 32 stroke types in four classifications. Chiang, op. cit., pp. 151-62, favors eight basic strokes and sixty subtypes. He also cites previous systems based on 14 and 72 basic stroke types.

researchers, is to define the entities to be considered rather than allowing them to define themselves. This is not a matter of lack of rigor; the investigator is invariably caught between the Scylla of an unmanageable multiplicity of entities and the Charybdis of a selection too limited to be of much use.

The entity is not an operational concept in the sense that the stroke and the character are.[18] In writing the character, the completion of a stroke is signalised by the raising of the writing implement from the paper. The completion of a character is signalised by the shifting of the implement, raised from the paper, for a certain distance (depending on the size of the character) in a certain direction (depending on whether the characters are written in columns or horizontally). No such definitive action characterises the writing of an entity. It is a descriptive concept only. Additional evidence for this view is seen in the fact that no two independent investigators have ever agreed on a list of entities (see Chapter III).

Stroke Order. The order in which the strokes are written to form a given character is traditionally fixed on the basis of aesthetic rather than logically systematic criteria. Stroke order must be learned. The process of learning involves (1) memorisation of the stroke order of a

---

[18] Strictly speaking, the stroke and the character are entities in their own right. The word "entity" is, however, used here to denote a graphic constant intermediate between stroke and character. A very common entity is the "mouth," k'ou /25/, which appears twice in the character yü /26/, "speech," and once in the character kuo /27/, "country." In each case the entity is written as a whole with no other strokes intervening. Its entity is only graphic; its meaning, unless it is also a radical, is irrelevant.

large number of basic characters, (2) application to new characters of a sense of correct stroke order gained by induction from the examples already learned,[19] and (3) memorisation of exceptions as they occur in further learning. Teaching stroke order to a computer operator, or to a scanner, would be nearly equivalent to teaching a complete course in written Chinese.

Structural Characteristics of the Printed Character

The success of the Chinese Photocomposition Machine, in which an operator trained in Chinese derives a numerical code for each character by typing out strokes, numbered by type, in the traditional stroke order, suggest a line of attack for the present problem. In order to set up a theoretical basis for the automatic input of printed texts, it is advisable to explore the structural characteristics of the printed character. The concepts applied to the written character must be redefined so that they may be economically applied to the printed character. The definitions must be in operational terms, capable of being expressed in simple rules rather than in vague generalisations.

Stroke. An operational definition of the stroke must take into account the nature of the input, human or scanner, that will be making the decisions in practice. Since the number of basic stroke types is small, it is expected that the decision process will be based on comparison.

---

[19] The stage at which this process of generalisation takes over varies, of course, with the intelligence of the individual and the aptness of his teachers. In the author's case, application of (not wholly conscious) rules began to play a really useful role only after the stroke order of more than a hundred characters was memorised.

Therefore, a stroke may be defined as "a deposit of ink, part of a character, of the same shape, but not necessarily of the same size, as one of the examples on the keyboard (or in the scanner glossary)." A further condition is set that, if a form may be read as one stroke or a combination of two or more others, it be read as one.

There are substantial difficulties in the use of this definition, particularly in the meaning of the word "shape." In a compound stroke such as the feng-kou /28/, for example,[20] the proportion of the horizontal to the curved part varies in certain characters. The na /29/ stroke is often longitudinally compressed to the point where it is difficult to grossly distinguish it from the tien /30/ stroke.[21] Moreover, imperfections in printing may make one stroke look like two. These problems are not particularly troublesome to the human operator, but they require taking into account when an input scanner is designed. There is no reason to believe that their solution will be beyond the resources of the designers, so long as they recognise them as problems. In the light of current progress in the field, it would be useless to attempt to make definitive statements about the capabilities of scanners ten years from now.

Entity. In the case of the written character, the entity has been seen to be a descriptive rather than an operational concept. The word "entity" describes a compound of a small number of strokes, written as a

---

[20] See stroke type "T," Figure 1, Appendix B.

[21] See stroke types "V" and "E," Figure 1, Appendix B.

whole, which occurs frequently in characters.

For the printed character, the qualification "written as a whole" is meaningless. Identification of an entity for programming purposes involves, as in the case of the stroke, comparison with a list prepared in advance. But, because any comprehensive list would include hundreds of entities, the time such a comparison process would take renders it useless. Even if a limited selection of high frequency entities were used, the identification problem would be unduly complicated for a scanner. For these reasons, it is unlikely that the entity would be a useful concept for application to printed text input.

Stroke Order. The printed character has no stroke order, because it is printed all at once. In connection with the use of a scanner, however, an analogous concept suggests itself--a synthetic stroke order, so to speak, which is determined not by the order of writing strokes, but by the order in which strokes are encountered when the printed character is scanned. This concept will prove to be of utmost importance, because it derives from the structure of the character itself an ordering relation for the numerable strokes.

In scanning, the scanner (or input operator) starts at a fixed point on the character, e. g., the upper right corner.[22] It sweeps across the width of the character, noting and identifying each stroke as its upper extremity is encountered. After one sweep, it returns to the right side

---

[22]This description applies to a horizontal scan. The merits of various starting points and scanning directions are discussed in Chapter III.

of the character, moves down a short distance, and scans across again. The process is repeated until the rest of the character has been covered, and all strokes identified. Identifying codes are fed, stroke by stroke, to the input of the computer.

In order to assure a reproducible stroke order, the following quantities must be specified:

1. Type size and style. Since the relative alignment of strokes in a given character occasionally varies with type size and style, it will be necessary to have a separate computer program for each variation. Since the range of variation in technical journals is very small, this will not be a major problem.

2. Number of sweeps per character. Since all strokes newly encountered on one sweep will be considered as starting on the same level, the order in which the strokes are reported will depend on the sweep frequency. Sweeps per character is a more useful parameter than sweeps per inch, since it allows direct comparison between different type sizes, and simplifies the task of reprogramming.

CHAPTER III

A SOLUTION TO THE PROBLEM

The structural characteristics of the written Chinese character have been investigated from a predominantly phenomenological point of view to determine which concepts are most useful for the operational identification of characters. Analogous concepts have been developed for printed characters, and problems involved in the application of each discussed. The groundwork has been laid for application of the concepts of stroke, entity, and stroke order to the development of a specific method for input of Chinese characters to a digital computer.

What is required is a simple intrinsic basis for numbering characters. The numbers derived must be of minimal length, since redundancy means waste of computer storage capacity, always at a premium.[1] There must be few (or preferably no) cases of two or more characters yielding the same number. If ambiguities are involved, the means of their resolution must not be unduly complicated.

In this chapter, several numbering systems already in existence are examined for applicability to computer input. The most suitable system, that used in the Chinese Photocomposition Machine, is discussed in some detail. A method is proposed for adapting it to automatic programming of

---

[1] Y. Bar-Hillel, "Can Translation be Mechanized" (author's abstract), Mechanical Translation, III (November 1956), p. 67. Original article published in the Hebrew journal Mada'.

printed texts. The results of an experimental analysis of five hundred common characters, using the proposed method, are evaluated.

The Revised Four-Corner System

The fact that Chinese characters are not spelled prompts an obvious question: How are they arranged in dictionaries? Since the first step in mechanical translation is the construction of an automatic dictionary, this question has considerable relevance to the subject of this study.

Because the pronunciation of a character varies greatly in different dialects, an arrangement according to pronunciation would be valueless for a dictionary of more than regional value. For this reason, traditional practice has been to arrange characters according to their radicals. Even as early as the first century A. D., Shuo Wen Chieh Tzu /31/, the great etymological dictionary, used a system of 560 radicals for classification.[2] The tabulation which has replaced all earlier ones includes 214 radicals. The radicals themselves are ordered according to the number of strokes they contain, from one to seventeen. The order of radicals composed of the same number of strokes has been arbitrarily fixed.

The radical system is gradually being abandoned, because it is artificial and cumbersome. As noted before, not all characters, etymologically speaking, have radicals. In the case of those that have, the radical is

---

[2]This remarkable work, compiled by Hsü Chen /32/, has been superseded (in part) only by recent research, and is still available in many editions, including one published by the Commercial Press, Shanghai, in 1914. The title may be translated as "A Discussion of Simple Characters and an Explanation of Compound Characters."

often singularly difficult to identify, particularly because it may appear in any part of the character.[3] All traditional Chinese dictionaries contain an "Index of Characters Having Obscure Radicals," in which a large number of characters, including many of the common ones, is grouped according to number of strokes.[4] Even knowing the radical, it is difficult to locate the section devoted to that radical in the dictionary, since thumb indexes are not used. Characters are arranged under each radical according to number of strokes (excluding those in the radical). Having come thus far, the frustrated Chinese has determined the character's location to within a few pages. the last step is to glance over those pages until the character is found. The whole process can easily consume five minutes, even when performed by a person who has had considerable practice.

A simple and speedily used indexing system was announced in 1928 by

---

[3]Knowing the meaning and pronunciation of a character, there are three indications a Chinese uses to locate the radical:

A. One part of the character obviously indicates the class of meaning. E. g., in yü /26/, "discourse," the left half of the character is the radical yen, which is used for characters having to do with speech.

B. One part of the character is obviously the phonetic, so the remainder is probably the radical. E. g., in chih /33/, the upper half of the character is pronounced chih when it stands alone, so it is likely that the lower half is the radical.

C. Certain forms, when they appear in characters, are almost always the radical, e. g., jen, shui, mu /34/.

There are, unfortunately, many cases where none of these indications are useful.

[4]The table in the popular Mathews' Chinese-English Dictionary (Cambridge: Harvard University Press, 1943) contains over a thousand characters, including variant forms. The dictionary defines only 7773 characters.

the Commercial Press of Shanghai, one of the world's largest publishing houses. It was originally worked out and revised by Wang Yün-Wu /35/, the firm's editor in chief. The Revised Four-Corner System derives a four-digit number from the strokes, or stroke combinations, appearing in the four corners of each character. The types are numbered 0 to 9; the reader, having memorised the table of types, can find the four-digit number corresponding to a character by merely looking at its four corners in turn.[5] This, at least, was the intent of the original system; in the revised system, however, a number of additional rules are appended. These rules reduce ambiguity by establishing a means for deriving two additional digits for each character.[6]

A similar method, the Five-Stroke System, was worked out by the Chung-Hua Bookstore /37/, a competitor of the Commercial Press in the dictionary field.[7] Because the latter system lacks the convenience of Wang's, and because it requires a knowledge of stroke order, it has not attained the popularity of the Four-Corner System. In the Chinese publishing

---

[5] See Appendix C for Wang's table of types. Note that numbers 1, 2, 3, and 7 each corresponds to several stroke types; that 0, 4, 6, 8, and 9 each corresponds to a combination or combinations of strokes; and that 5 corresponds to a vertical which has been crossed at least twice by other strokes. For simplicity, "stroke type" is used in the text to denote any of these strokes or combinations.

[6] A complete outline in English of the revised system may be found in Y. W. Wong, Wong's System for Arranging Chinese Characters (Shanghai: Commercial Press, Ltd., 1928), pp. 39-44. A Chinese exposition is available in any dictionary which uses the system, e. g., Wang Yün-Wu Ta T'zu-Tien /36/ (Shanghai: Commercial Press, Ltd., 1933), pp. i-ii.

[7] T'zu Hai /38/ (Shanghai: Chung-Hua Bookstore, Ltd. /37/, 1937), Vol. II, pp. i-ii.

world, popularity is best measured by the frequency with which one's work is pirated.

At first glance, Wang's system seems to promise an almost ready-made basis for computer programming. It does, indeed, provide numbers derived from the structure of characters. The author's expectation that the system could be regularised was, however, disappointed, for the following reasons:

1. The system does not always work on the actual corners of the character. In the character wen /39/, for instance, only two stroke types, 0 and 4 /40/, are needed to cover both the top and bottom. In a case of this sort, one ascribes the first number to the upper left and the second to the lower left, assigning 0's to the other two corners. The four-digit code for wen is thus 0040. The character shih /41/is completely described by stroke type 4 /42/. This number is assigned to the upper left corner and 0's used for the other corners. The four-digit code for shih is thus 4000. In neither of these two characters, which are common types, are the corners of primary importance for numbering purposes.

2. The additional rules of the revised system are often arbitrary. In the character kuo /27/, the outside "frame," which includes all four corners, is equivalent to stroke type 6 /43/. One would thus (assigning this number to the upper left corner) expect kuo to have the four-digit code 6000. But one of the subsidiary rules states that in characters having a "frame," the two digits assigned to the lower corners are taken from the strokes inside the "frame." In this case, the two "lower-corner" stroke types inside the frame are 1 and 5 /44/, so the code for kuo is 6015. This rule holds only when the "framed" character stands alone, not when it is the

phonetic of another character. The rule, therefore, does not hold for the character kuo /45/, in which the actual upper right and lower right corners determine the corresponding numbers.

Many of the other subsidiary rules are equally complicated and equally arbitrary. The system was designed for the use of Chinese; it would be difficult to teach to a Western operator in a short time and impossible, the author believes, to economically program for a computer-scanner.

3. The system does not successfully cope with ambiguity--cases of two or more characters with the same numerical code. In the T'zu Yüan /46/, a large modern dictionary which defines over 11,000 characters, as many as 36 characters are listed under the same four-digit code. The average is about ten per code.[8] This poses no particular problem for a Chinese, who needs only a moment to locate the one character he wants in a list of 36 characters. For operational purposes, however, a further principle of order is requisite. The revised system includes rules for deriving two additional digits:

(Fifth digit) The first stroke above, but not touching, the stroke in the lower right corner, is counted according to the table of stroke types.

(Sixth digit) The total number of horizontal strokes (stroke type 1 /47/) is counted.

These two rules do not altogether do away with ambiguity, but they

---

[8] T'zu Yüan /46/ (Combined edition; Hong Kong: Hua T'ung Company /64/, 1955), pp. iii-cxviii.

reduce it to small dimensions. At the same time, the sixth-digit rule adds a third operation (counting) to the two already required (locating and comparing). Three different types of operation in a total of six numerating steps makes for an unduly complicated scanner program. Where the ideal solution would be to derive a code from one continuous scan of each character, here many scans would be necessary. The Revised Four-Corner System does not meet the criterion of a simple operation for the derivation of each digit in the code.

Entities

In the last chapter, the meaninglessness of the concept "entity" from an operational viewpoint was discussed. It will be worth while here to consider the quantitative aspects of several systems in which various entities are used for complete descriptions of characters.

Radicals alone. Taking the 214 radicals themselves as entities, it is possible to break down every character into a combination of entities, as P. A. Boodberg has shown.[9] The character kuo /27/, for instance, may be broken down into four entities /48/, all of which are radicals.[10] While 214 is, compared with the number of entities in other systems, a small figure, the necessity for a scanner's comparing each part of each character

---

[9] P. A. Boodberg, Cedules from a Berkeley Workshop in Asiatic Philology (1955). (Mimeographed). Boodberg's system was worked out for philological applications.

[10] That is, all four are listed in the standard radical table. Only the first is the radical of the character kuo itself.

with 214 glossary entries, makes this method prohibitively cumbersome for application to the problem of this study. Another drawback is that of ambiguity. The same character, kuo /27/, could also be broken down into a total of seven radicals /49/, or into several intermediate combinations of five or six. Ambiguity in other characters would extend over an even greater range of possibilities; a very complicated program for minimisation would have to be devised. Again, a large number of complete scans would be required for the input of each character, creating a temporal bottleneck in the mechanical translation process.

Radicals and phonetics. The problem of ambiguity would be less troublesome if each character were broken down into its radical and phonetic. This is the method used in dictionaries compiled by W. Simon and H. C. Fenn.[11] Both dictionaries list the standard radicals; in addition, they number the remainder of the character (character minus radical) according to a list of phonetics. Simon uses his own compilation of 545 phonetics. Fenn uses W. E. Soothill's table of 880 phonetics.[12]

If these schemes were not arbitrary, if they could be recast simply and operationally, a radical-phonetic system might be of value for our purpose, since there would be little or no ambiguity of the sort that results from several possible breakdowns for the same character. As stated in Chapter II,

---

[11] Henry C. Fenn, The Five Thousand Dictionary (Cambridge: Harvard University Press, 1942); and W. Simon, A Beginner's Chinese-English Dictionary (London: Lund Humphries and Company, Ltd., 1947).

[12] Fenn, op. cit., pp. xiv-xix; and W. Simon, Chinese Radicals and Phonetics (London: Lund Humphries and Company, Ltd., 1944), pp. 109-257.

the fact is that not every character was formed from a radical and a phonetic; even with those that were, the radical is often extremely difficult to find. It may be on the left, right, top, bottom; it may surround the phonetic or be surrounded by it; it may be divided into two parts which sandwich the phonetic. There are, in fact, ideographs which do not contain their radicals. The radical of wang /50/ is yü /51/, which contains one more stroke than wang. The radical of wei is chua /52/, which appears in the printed form of wei /53/ but not in its written form /54/. A radical-phonetic system, then, although it employs 759 (Simon) or 1094 (Soothill) entities, is no less ambiguous than Boodberg's system of 214 entities.

Subradicals. P. Polletti proposed a system of entities called subradicals, derived by isolating the entities of which radicals are composed.[13] According to Wang Yün-Wu /35/, "I have personally made a careful study of that part and found that the number of subradicals for this purpose would be no less than 584."[14] This system deserves only passing mention, because it has all the disadvantages of Boodberg's radical system without even the virtue of a small number of entities. To illustrate the ambiguity inherent in the system, Wang cites the case of the character chung /55/, which may be decomposed into three /56/, four /57/, or five /58/ subradicals.[15]

Strokes and Stroke Order--The Proposed Method

Recent research done by the Graphic Arts Research Foundation, Inc., Cambridge, Massachusetts, points the way to a satisfactory method for

---

[13]Wong, op. cit., p. 6. [14]Ibid. [15]Ibid.

automatic input of printed Chinese characters to a computer. This research resulted in the development of the Chinese Photocomposition Machine, by means of which a Chinese operator can prepare material for offset printing at a much greater speed than is possible with old-fashioned methods.[16,17] The input function of the CPM is not entirely operational, since a knowledge of traditional stroke order is required. In order to adapt the CPM for computer input use, it is necessary to substitute an operational analogue of stroke order, and to make other changes to accomodate it to working from printed texts.

Working Principle of the Chinese Photocomposition Machine. The CPM uses the concepts of stroke and stroke order to derive numerical sequences from the structure of Chinese characters. Twenty-one basic stroke types are represented on a keyboard. Striking a key inputs a binary code signal--corresponding to a stroke type--to the machine. By striking the keys corresponding to the strokes of a character, in the order dictated by the stroke order of that character, a numerical sequence is derived which is unique, or nearly so, to that character. There are only 24 cases of ambiguity in a

---

[16]Patent application for the Chinese Photocomposition Machine (hereafter abbreviated CPM) has only recently been made. Because no descriptive material has yet been published, all facts cited here have been obtained by personal interview with S. H. Caldwell of MIT or R. G. Crockett of the Graphic Arts Research Foundation, Inc., and are released with Caldwell's permission. Caldwell is the inventor of the CPM, and Crockett the project administrator.

[17]Typesetting of Chinese publications is conventionally done entirely by hand. A Chinese newspaper must keep cast-metal types for up to ten thousand characters on hand. It is the unenviable job of the compositor to locate (and replace) each piece of type quickly and accurately.

glossary of 2332 characters; provision is made for resolution of these ambiguities by the operator.[18]

The obvious advantage of using the stroke as the basis of numbering is that the stroke, unlike the entity, is capable of operational definition. Further, a particular stroke may be identified by comparison with a small glossary of stroke types, while any usefully comprehensive collection of entities would be numbered in the hundreds. Coding of characters is according to a simple, self-consistent rule rather than, as in the case of the Revised Four-Corner System, a number of relatively complicated and sometimes arbitrary rules.

Operation of the Chinese Photocomposition Machine. The glossary of the CPM is contained on a glass matrix five inches square. The matrix is divided into 2400 (50 x 48) equal areas. Negative images of 2332 common characters are represented on the matrix.[19] The fifty-two characters which are not uniquely defined by their codes appear twice. In addition, there are fifteen punctuation marks and a focussing target.

The operator of the CPM, by operating a keyboard of forty-three keys, inputs a binary code signal for each character in the input text, including

---

[18]There are four cases of three characters having the same code, and twenty cases of two characters having the same code; the total of characters of which a unique numerical sequence is not derived is thus fifty-two.

[19]In choosing the characters for the machine's glossary, several frequency lists were employed, among them an analysis of a speech by Mao Tse-Tung /59/ and a list prepared by L. S. Yang, Harvard University. The final selection was checked with a Standard Kanji List.

punctuation.[20] The binary impulses, by actuating a relay-operated character selector, position the matrix so that the character chosen is moved into juxtaposition with an aperture.[21] The character is then projected from the matrix onto a ground-glass viewing screen. If the correct character appears (i. e., if the input was correctly performed), the operator presses a key which instructs the machine to print the character. A flash lamp then projects the image of the character from the matrix through a lens (the focal length of which may be adjusted to determine any one of four type sizes) onto a film. By repeating the input process, a strip of film is obtained from which an offset plate is made and the copy printed.

In order to reduce the binary storage requirements of the machine, an analysis of its vocabulary was made during the initial programming process. In this analysis, each stroke type was represented by a letter. A card was prepared for each character, giving its spelling (letter code). The cards were arranged in alphabetical order according to letter codes. By comparing each card with the one before and after it, it was possible to determine how much of the letter code was required for differentiation, and what portion was redundant. The non-essential part of the letter code was

---

[20] The keyboard consists of 23 keys for stroke types, including two keys each for the common vertical and horizontal strokes, so that they may be struck with either hand; three for the Chinese numerals 1, 2, and 3; ten for a total of twenty-four common entities; six for punctuation marks; and one for the "termination" key (see below).

[21] Characters are arranged on the matrix so that the distance from the upper left corner is a function of frequency. Since that corner corresponds to the rest position of the matrix, the distance the matrix must be shifted, and thus access time, are functions of character frequency.

then cut off,[22] and the remaining code (the minimum spelling) converted to its binary equivalent and programmed.[23]

In practice, the operator "spells out" a character on the keyboard, and each binary pulse actuates relays until, when the input of the minimum spelling is complete, the character is located. The keyboard then locks to cut off further input, since the operator is not expected to know the minimum spelling. The image of the character appears on the viewing screen. The operator issues instructions to print or cancel, at which time the screen clears and the keyboard unlocks.

In cases where the full spelling of a character is part of the spelling of a more complex character, the keyboard does not lock after all the strokes in the former character have been punched onto the keyboard. It is necessary for the operator to first press the "termination" key, which signals the machine that input has been completed.[24]

---

[22] In this series of characters, for instance, it will be seen that the second is distinguished from the first and third by its minimum spelling, and that the last six letters of its full spelling are thus redundant:

CHARACTER	FULL SPELLING	MINIMUM SPELLING
Tsou /60/	BDBDBGV	
Ch'ü /61/	BDBDBGVBDBBBDJV	BDBDGBVBD
Yüeh /62/	BDBDBGVBMNGE	

[23] In cases where the full spelling of one character is the first part of the spelling of a more complex character, the letter "A," which represents the pulse from the "termination" key, was added to the spelling of the shorter character. Minimum spelling of tsou /60/ in footnote 22 is BDBDBGVA.

[24] The "termination" key is not used in any other case, since the character is normally projected onto the viewing screen, and the keyboard locked, as soon as the character selector positions the matrix.

Resolution of ambiguous codes, each of which corresponds to more than one character, is semi-automatic. When an ambiguous code is fed to the machine, the keyboard does not lock, and no image appears on the viewing screen. Since this is ordinarily an indication that the "termination" key is to be pressed, the operator will do so, unless he knows from experience that he has encountered a case of ambiguity. After the termination key is pressed, the keyboard lock and viewing screen still do not operate.[25] This is a clear indication that the operator is to look up the character he wants on an auxiliary list, where a number--1, 2, or 3--appears opposite each ambiguous character. He presses the key bearing the appropriate Chinese numeral, and the selected character appears on the viewing screen, ready for printing.

A short-cut for additional input speed is provided by the inclusion of twenty-four entities on the keyboard. These were chosen on the basis of frequency, and are grouped on a total of ten keys.[26] When an entity key is pressed, a signal equivalent to that for its constituent strokes is fed into the machine. Use of the entity keys is optional, and depends in the operator's skill.

---

[25] In order to make sure that the character is correctly located whether or not the "termination" key is pressed, each ambiguous character appears in two positions on the matrix.

[26] Each key, as on an ordinary typewriter, has a shift position, giving a total of 20 entity positions. In addition, there are two groups of three entities which have the same spelling (are written with the same strokes in the same order, e. g., /63/, code BDB) and thus require only one position for each group.

Adaptation of the Chinese Photocomposition Machine to Automatic or Semi-Automatic Input of Printed Texts. The CPM offers a very attractive basis for solution of the problem of this study. Its basic principles are sound and simple; just as important, it is, mechanically speaking, a fait accompli. As was shown in Chapter II, however, traditional stroke order is useless to an untrained operator who is working from printed texts.

By substituting a synthetic stroke order, a straightforward, operational scanning procedure, the input process may be adapted to automatic operation by a suitably designed computer-scanner or to semi-automatic operation by an operator with no Chinese training.

Application of a synthetic stroke order to the CPM was experimentally tested by deriving the minimum spelling of five hundred characters, using substantially the same procedure as that used in early stages of the CPM's original programming procedure. A detailed report is given in Appendix B. The scanning order chosen for the test was from right to left, with the first scan across the top of the character and succeeding scans proceeding toward the bottom. The general top-to-bottom direction is that best suited for characters printed in columns; a bottom-to-top direction would necessitate a scanner's changing direction twice for each character. The right-to-left sweep direction was chosen because a preliminary survey indicated that, using that direction, there would be a greater diversity of stroke types for the first stroke counted. It was believed that there is a relationship between diversity in the first stroke (over the whole glossary)

and the average minimum spelling length.[27] The optimal scanning order can be finally determined only in practice. If the present trend toward arranging the text of Chinese publications in left-to-right lines instead of in columns continues, a vertical rather than a horizontal sweep direction may prove preferable.

The results of the test, in terms of number of ambiguities, average minimum spelling length, etc., were successful, which is to say that no unsuspected flaws in the proposed method were revealed. It became clear that in many cases the spelling of a character depends on idiosyncrasies of orthography, so that reprogramming for different type sizes and type faces is essential. It was also noted that spelling varies greatly with the number of scans per character. That figure can be standardised for a scanner, but for a human operator further experiments are necessary to determine whether consistent results can be obtained by relying on the naked eye to determine which of two strokes begins at a higher level within a character. If not, it will be necessary to use a transparent grid, or similar device, to indicate scanning levels. All strokes originating between two horizontal grid lines would be counted as starting on the same level.

The CPM's procedure for resolving ambiguous codes may be modified to depend less on an auxiliary key in which the selected character must be looked

---

[27] This idea can be proved or disproved only through much more intensive testing. As a first step, another experiment using a left-to-right sweep direction was contemplated; the number of ambiguous codes for each of the sweep directions was to be compared. Time limitations precluded completion of the experiment.

up. In the experimental trial of the right-to-left scan, there were eleven characters (of five hundred tested) which were not uniquely determined by their spellings. For most of these characters, a different method of resolution suggested itself, which can be used even by a scanner. The alternate code for each of the ambiguous characters can be that obtained by rescanning in a new direction. Results of a left-to-right rescan are shown in Figure 4, Appendix B, where it is apparent that only two of the eleven ambiguous characters cannot be dealt with in this way. It is unlikely that a rescan in any direction will resolve all ambiguities. A human operator can still be provided with an auxiliary key, much shorter than is now the case. The best method for final resolution of ambiguities by a scanner will depend on the capabilities of the scanner finally designed.

The first step in the adaptation of the CPM will be a reprogramming for the synthetic stroke order. For this purpose, it will be necessary merely to rewire the character selector mechanism.[28] Since characters are arranged on the matrix according to frequency, a new matrix will not be required.

If the CPM is to be used for manual input to a computer, parts required will be its keyboard, with lock and viewing screen, and the later stages--relay network, character selector, matrix, and projector---needed to project the characters onto the viewing screen or a strip of facsimile paper as a visual check on accuracy of input to the mechanical translation computer. It may prove feasible to include the functions of the relay

---

[28] Statement made by R. G. Crockett, personal interview, 26 December 1957.

network and the character selector in the computer's program. The keyboard can be somewhat simplified, since the entity keys will be of no value.[29]

If the CPM is to be used with a scanner input, the keyboard with its auxiliary devices will no longer be needed as such, since the binary pulses will come directly from the scanner's computer.

It is impossible at this point to lay down specific design requirements for a computer-scanner; in this case invention must be the mother of necessity. It is possible, however, to outline the task which the scanner must perform. The following description assumes a horizontal sweep.

It will scan across the top of the character until it encounters the upper extremity of a stroke. It will then trace down the stroke, remembering the areas it passes over. When it reaches the end of the stroke, it will compare the shape of the area it has covered with standard shapes in its glossary. When the stroke has been identified, information as to the stroke type will be routed to the mechanical translation computer. The scanner will then return to the point where it first encountered the stroke and continue its scan, repeating the identification process for other strokes. When it has completed one sweep, it will move into position for the next sweep and start scanning at a lower level. By checking the coordinates of subsequently encountered ink areas with its memory, it will ignore strokes identified during an earlier sweep. The scanning and identification process will continue until the scanner receives a signal from the character selector,

---

[29] See Chapter II.

indicating that the minimum code has been received and the character located.[30] The scanner will then locate the next character and begin scanning it.

An exact definition of the word "stroke" in the above description must, again, depend on what turns out to be feasible when the scanner is finally designed. Whether a scanner can trace continuity of slope of a stroke's outline past an intersection with another stroke, is a question that cannot be answered now. It is unlikely that any unexpected answers cannot be taken into account in the computer-scanner's program.

---

[30] I. e., the corresponding entry in the automatic dictionary of the mechanical translation computer has been found. The signal will be analogous to the locking of the CPM's keyboard.

CHAPTER IV

SUMMARY

In order to establish a basis for the automatic or semi-automatic programming of printed texts for mechanical translation, it is necessary to find a method for deriving unambiguous numerical sequences from the structure of characters. The basic structural concepts of the Chinese character--stroke, entity, and stroke order--were defined. Since these are ordinarily defined in terms of manuscript characters, it was further necessary to examine each concept for its applicability to printed characters. It was found that the concept "stroke" had approximately the same meaning for both written and printed characters. The concept "entity" was shown to be of no operational value in a treatment of printed characters. The concept "stroke order" was replaced by an analogous "synthetic stroke order," a straightforward scanning procedure.

Several systems used for classifying characters according to structure were weighed for suitability. The Revised Four-Corner System of Wang Yün-Wu /35/ was discarded because of its complexity and arbitrariness. The entity systems of Boodberg, Simon, Fenn, and Polletti were shown to be unacceptable because of the large number of entities involved, and the difficulty of their identification in practice. The systems of Boodberg and Polletti, while employing a smaller selection of entities than the other two systems, suffer from a substantial ambiguity problem. While all of these systems are well suited to the purposes for which they were conceived, they are useless for the purpose of this study.

The Chinese Photocomposition Machine offers a workable basis for adaptation to mechanical translation input. Its working principle makes use of the concepts "stroke" and "stroke order," both of which are capable of redefinition for printed text input.

The experimental pre-program for five hundred characters indicated that the proposed method is feasible. At the same time, it emphasised certain limitations that must be taken into account when designing and programming a computer-scanner:

1. What a computer-scanner will be able to recognise as a stroke cannot be specified until the design of such a device is worked out, and its general capabilities definitely known.

2. While a glossary of twenty-one stroke types is neither too small nor too large, normal variations in the proportion of each stroke as it appears in different characters, and other variations caused by imperfect printing, must be recognised, and the strokes related to their respective stroke types.

3. Because of variations in individual character design, a given program can be expected to be effective for only one type face and one type size.

It was also noted that manual input using the proposed method is much slower than regular input to the CPM by a Chinese operator. Use of the proposed method by a human operator is indicated only as an interim step, pending design of a computer-scanner. Since a number of years is likely to elapse before the problem of Chinese input for mechanical translation becomes pressing in practice, it is impossible to say whether manual

input will be used at all except in experimental work.

Other applications

According to S. H. Caldwell, the proposed method will be of value for military applications of the Chinese Photocomposition Machine itself.[1] For security reasons, it is desirable under certain conditions that the CPM be operated by military personnel who have had no training in Chinese. It was first contemplated that Chinese "programmers" would be used to reduce Chinese texts to letter code, with actual operation of the CPM carried out by untrained personnel. Since in this case there would be an absence of feedback in the checking process, the proposed method offers a more desirable solution. Mechanical alterations required of the CPM would be minor in scope.

While the proposed method can cope with Chinese characters as they appear in Korean and Japanese, a large number of simple phonetic symbols is used in both languages. The actual mechanical translation input method finally adopted should be a combination of the proposed method for characters and, for the phonetic symbols, a comparison method similar to that used for alphabetic languages.

---

[1] Statement by S. H. Caldwell, personal interview, May 29, 1958.

APPENDIX A

CHINESE CHARACTERS TRANSLITERATED OR OTHERWISE REFERRED TO IN THE TEXT

/1/ 二

/2/ 川

/3/ 白話

/4/ 文言

/5/ 白話文

/6/ 國語

/7/ 方以智

/8/ 陳公哲

/9/ 初學書法

/10/ 蔣彝

/11/ 溫永琛

/12/ 其

/13/ 狗

/14/ 犭(=犬)

/15/ 句

/16/ 胡適

/17/ 文學革命運動

/18/ 王際真

/19/ 楷書

/20/ 草書

/21/ 𠂇

/22/ 夕

/23/ 外

/24/ 外

/25/ 口

/26/ 語

/27/ 國

/28/ 乙

/29/ 丶

/30/ 丶

/31/ 說文解字

/32/ 許慎

/33/ 智

/34/ 亻氵木

/35/ 王雲五

/36/ 王雲五大辭典

/37/ 中華書局

/38/ 辭海

/39/ 文

/40/ 亠乂

/41/ 十

/42/ 十

/43/ 口

/44/ 一戈

/45/ 幗國

/46/ 辭淵

/47/ 一

/48/ 口戈口一

/49/ 弋，冂冂一一一

/50/ 王

/51/ 玉

/52/ 灬

/53/ 爲

/54/ 爲

/55/ 鐘

/56/ 金立里

/57/ 人玉立里

/58/ 人玉立田土

/59/ 毛澤東

/60/ 走

/61/ 趣取

/62/ 越

/63/ 工士土

/64/ 華通公司

APPENDIX B

AN EXPERIMENTAL PRE-PROGRAMMING ANALYSIS OF FIVE HUNDRED CHINESE CHARACTERS

This experiment was designed to test, on a necessarily limited basis, application of a simple scanning procedure to the programming method of the Chinese Photocomposition Machine. It is identical with that part of the CPM programming procedure which precedes conversion of letter codes to binary codes (see Chapter III), except that a synthetic stroke order is used instead of traditional stroke order.

Experimental Conditions

Since a definition of the stroke in scanner parameters must await the design of a suitable scanner, the stroke was defined simply as any configuration appearing in the stroke type table (Figure 1). In order to avoid ambiguity, in cases where a given configuration within a character could be counted as one stroke or two, it was counted as one. For instance, a stroke of type "Z" was not counted as two strokes of types "D" and "Q" respectively. While the stroke table in Figure 1 is not the only possible stroke table, nor even necessarily the best for the purpose, its use allows direct comparison with the existing program of the CPM.

In order to avoid variation in relative alignment of strokes within characters of different type faces and type sizes, one common type face, large enough for detailed examination, was used throughout. The type face chosen was that used in character definition headings of the T'zu Yuan /46/,[1]

---

[1] T'zu Yüan /46/ (combined edition; Hong Kong: Hua T'ung Company /64/, 1955).

FIGURE 1.

TABLE OF STROKE TYPES MODIFIED FROM THAT USED IN THE CHINESE PHOTO-COMPOSITION MACHINE[1]

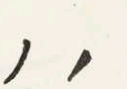
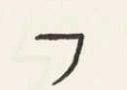
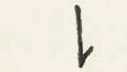
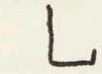
STROKE TYPE LETTER	CORRESPONDING STROKE TYPES
B	
D	
E	
G	
H	
J	
K	
L	
M	
N	
P	
Q	
R	

Figure 1 (continued)

STROKE TYPE LETTER	CORRESPONDING STROKE TYPES
S	
T	
U	
V	
W	
X	
Y	
Z	

[1] Adapted from chart, "Basic Keyboard Characters," file RC 8-16, Graphic Arts Research Foundation, Inc.

a comprehensive modern dictionary which features clear, conventional orthography.

In order to simulate the action of a scanner set for a large number of sweeps per character, a 25X magnifier was used to determine which of two strokes within a character originated at a higher level. Only those differences which were readily apparent through the magnifier were taken into account..

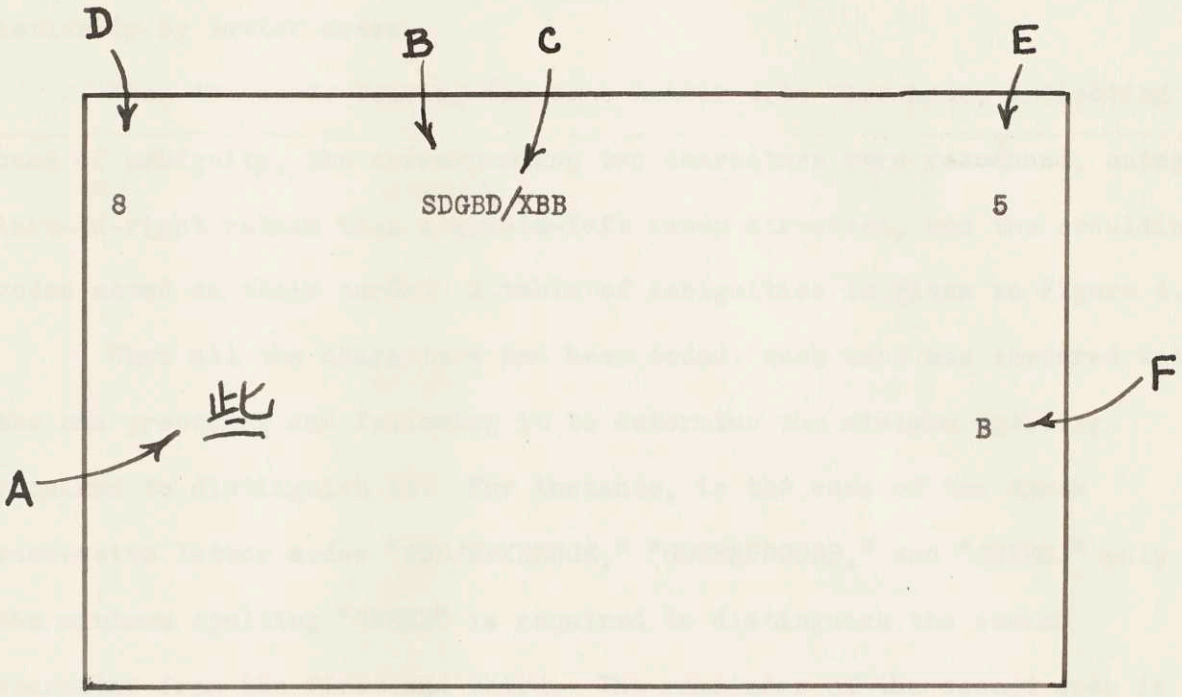
Experimental Procedure

The starting point for the experiment was an Ozalid print of the CPM matrix, furnished by the Graphic Arts Research Foundation, Inc. Five hundred characters were chosen for the experiment. In order to ensure a large selection of graphic types, 250 characters were chosen from the five hundred most frequent characters, and 250 were chosen from the remainder of the one thousand most frequent characters.

Since 2500 characters are contained in an area five inches square on the matrix, and since the graininess of the Ozalid process precludes close examination of the characters, it was necessary to locate each character in the T'zu Yüan for the actual scanning. The letter code for the stroke whose upper extremity was topmost was written on a file card (Figure 2).. In the case of two or more strokes the upper extremities of which were at the same level, the one furthest to the right was taken first, followed by the others proceeding from right to left. The remaining strokes were coded in like manner, in descending order, until the spelling of the entire character had been noted. The number of letters in the full spelling was

Figure 2.

SAMPLE CARD USED IN PRE-PROGRAMMING EXPERIMENT



KEY

A. Character.

B. Letter code, full spelling.

C. Slash indicating cutoff point for minimum spelling.

D. Length of letter code, full spelling.

E. Length of letter code, minimum spelling.

F. Code for location of radical within the character.

written in the upper left corner of the card for reference. The character was also written on the card, as was a code for the location of the radical. A similar card was prepared for each character; they were filed alphabetically by letter codes.

When two cards bearing the same letter code were made, indicating a case of ambiguity, the corresponding two characters were rescanned, using a left-to-right rathan than right-to-left sweep direction, and the resulting codes noted on their cards. A table of ambiguities is given in Figure 4.

When all the characters had been coded, each card was compared with the one preceding and following it to determine the minimum spelling required to distinguish it. For instance, in the case of the three successive letter codes "GDDJEBKEBBGE," "GDEEHPDBDBB," and "GDEEX," only the minimum spelling "GDEEH" is required to distinguish the second character from the first and third. The remainder of the second code is redundant. A slash was placed within the full spelling on each card to indicate the end of the minimum spelling, and the number of letters in the minimum spelling was written in the upper left corner of the card for reference. Where the full code for a character was equivalent to the beginning of the code for the next (and more complex) character, the letter "A" was added to the spelling of the first character. "A" represents the signal of the "termination" key on the CPM. An example of the full and minimum spellings for a group of successive characters is given in Figure 3.

Experimental Results

For the five hundred characters examined, the average full spelling

Figure 3.

TABLE OF FULL SPELLINGS AND MINIMUM SPELLINGS FOR THIRTEEN SUCCESSIVE CHARACTERS

CHARACTER	FULL SPELLING, LETTER CODE	MINIMUM SPELLING, LETTER CODE
些	SDGBDXBB	SDGBD
皆	SDGBXGPDBB	SDGBX
此	SDGDBX	SDGD
北	SDGXX	SDGX
化	SGGD	SGGDA
貨	SGGDPDBBBVG	SGGDP
跳	SGPDGEBDEXBDX	SGPDG
真	SGPDRBBBVG	SGPDR
飛	TDYGGTY	TD
設	TGBBBJVBPDB	TGBB
恐	TGBDEXESEE	TGBD
沿	TGEXEPDB	TGE
風	TGGDPDBEX	TGG

FIGURE 4.

AMBIGUOUS CODES ENCOUNTERED DURING THE COURSE OF THE PRE-PROGRAMMING EXPERIMENT

DESIGNATION	LETTER CODE	CHARACTER	LETTER CODE, L-R RESCAN
1	DGBBSG	老	DGBBSG
		先	GDBBGS
2	GBDB	午	GBDB
		仁	GDBB
3	GBDBDB	年	GBDBDB
		在	GBDDBB
4	PDBB	日	DPBB
		丑	PDBB
5	VG	人	GV
		八	VG
		八	VG

length was 9.6 letters; the average minimum spelling length was 4.5 letters. The former figure would tend to increase very slowly for larger collections of characters; the complexity of a character is only indirectly related to its frequency. The average full spelling length for all 2332 characters in the glossary of the CPM is approximately eleven letters. The minimum spelling length would increase more rapidly with a larger collection of characters, since a larger number of letters would be required to distinguish the code for each character from those of the characters preceding and succeeding it. A figure for average minimum spelling length for the 2332 characters in the CPM's glossary is unavailable, but the author would estimate it, on the basis of material at hand, at about seven letters. For a very large collection of characters, redundancy would tend to level off, probably at about twenty-five per cent.

It was also found that five, or approximately one per cent, of the codes were ambiguous. Four codes corresponded to two characters each; one corresponded to three characters. Only two characters could not be distinguished by their left-to-right rescan spellings. It is noteworthy that these two are among the simplest characters in the collection (see Figure 4).

For the 2332 characters in the CPM's glossary, there were twenty-four ambiguous codes for a total of fifty-two characters, using the original CPM programming procedure. The author's a priori impression that ambiguity would increase proportionately with the size of the glossary was shown to be incorrect. For the most frequent five hundred characters

in the CPM glossary, there are seven ambiguous codes for a total of sixteen characters, using traditional stroke order. This is evidence, but not conclusive evidence, that the proposed method gives fewer ambiguities.

Other general conclusions from the experiment are given in Chapter III.

APPENDIX C

TABLE OF STROKE TYPES, REVISED FOUR-CORNER SYSTEM OF WANG YÜN-WU[1]

STROKE TYPE NUMBER	CORRESPONDING STROKES AND/OR COMBINATIONS
0	一
1	一 丶 乚 乀
2	丨 丿 亅
3	丶 乀
4	十 又
5	丰
6	口
7	𠂇 𠃌 乚 𠂆 𠂉
8	八 丷 人 𠆢
9	小 灬 忄 木 扌

[1] T'zu Yüan /46/ (combined edition; Hong Kong: Hua T'ung Company /64/, 1955), p. i.