(12) UK Patent Application (19) GB (11) 2 125 197 A

(21) Application No **8221271**
(22) Date of filing **22 Jul 1982**
(43) Application published **29 Feb 1984**
(51) **INT CL**[3] **G06F 3/023**
(52) Domestic classification **G4H** 13D KU **U1S** 2125 2126 2245 2247 G4H
(56) Documents cited **None**
(58) Field of search **G4H**
(71) Applicants **Pingyi Zhi, Long Jiang Lu 225, Shanghai, People's Republic of China, Shu Mei Wang, Long Jiang Lu 225, Shanghai, People's Republic of China, Kam Fu Wong, 22nd Floor, Wu Sang House, 655 Nathan Road, Kowloon, Hong Kong**
(72) Inventor **Pingyi Zhi**
(74) Agent and/or address for service **Lloyd Wise Tregear & Co., Norman House, 105—109 Strand, London, WC2R 0AE**

(54) **Encoding chinese characters**

(57) A method of encoding a Chinese character comprises dissembling the character into four constituent radicals each of which represents a pronounciable sound, as many radicals as possible, subject to the formation of a total of four radicals, being combined into the first radical and, again subject to the formation of four radicals, a succeeding radical is incorporated into a preceding radical, and in the event that there are less than four resulting constituent parts, after the constituent parts, the final stroke of the character as normally written is used as one of the constituent parts and in the event that there are still not enough total constituent parts to complete the four required, the whole character is repeated as a final constituent part, and thereafter representing the four constituent parts by the initial four Roman alphabet letters of the transliterations of each constituent part if the part is a character or its link character which the part closely resembles, the code for the Chinese character having therefore the resulting four Roman letter code.

FIG. 1.

CHARACTER	CONSTITUENT PARTS						CODE
路	口 K(ou)	止 Z(hi)	文 W(en)	口 K(ou)			KZWK
兢	克 K(e)	十 Z(uo)	口 K(ou)	儿 E(r)			KZKE
侍		人 R(en)	土 T(u)	寸 C(un)	丶 D(ian)		RTCD
余		人 R(en)	于 Y(u)	八 B(a)	丶 D(ian)		RYBD
吴			口 K(ou)	天 T(ian)	乀 N(a)	吴 W(u)	KTNW
天			二 E(r)	人 R(en)	乀 N(a)	天 T(ian)	ERNT

GB 2 125 197 A

The drawings originally filed were informal and the print here reproduced is taken from a later filed formal copy.
Chinese characters appearing in the printed specification were submitted after the date of filing, those originally submitted being incapable of being satisfactorily reproduced.

FIG. 1.

CHARACTER	CONSTITUENT PARTS						CODE
路	口 K(ou)	止 Z(hi)	文 W(en)	口 K(ou)			KZWK
兢	克 K(e)	十 Z(uo)	口 K(ou)	儿 E(r)			KZKE
侍		人 R(en)	土 T(u)	寸 C(un)	丶 D(ian)		RTCD
余		人 R(en)	于 Y(ii)	八 B(a)	丶 D(ian)		RYBD
吴			口 K(ou)	天 T(ian)	乀 N(a)	吴 W(u)	KTNW
天			二 E(r)	人 R(en)	乀 N(a)	天 T(ian)	ERNT

FIG. 2.

CHARACTER	CODE	FREQUENCY	CHARACTER	CODE	FREQUENCY
地	A	17	年	N	42
不	B	6	口	O	366
产	C	25	表	P	87
的	D	1	起	Q	96
而	E	62	人	R	13
分	F	33	是	S	3
个	G	18	他	T	29
和	H	7	水	U	89
义	I	26	这	V	10
进	J	52	为	W	15
可	K	48	学	X	37
了	L	14	有	Y	8
们	M	16	在	Z	4

FIG. 3.

CHARACTER	CODE	CHARACTER	CODE
一	1	六	6
二	2	七	7
三	3	八	8
四	4	九	9
五	5	十	0

FIG. 4.

CHARACTER - COMPOUND	CODE
但是 Dan Shi	DS
干部 Gan Bu	GB
无线电 Wu Xi Dian	WXD
工程师 Gong Cheng Shi	GCS
无产阶级 Wu Chan Jie Ji	WCJJ
科学技术 Ke Xue Jieshu	KXJS

FIG. 5.

CHARACTER - COMPOUND	CODE
一月……十二月	1Y……12Y
一日……三十一日	1R……31R
零点……二十四点	0D……24D
星期一……星期日	XQ1……XQR

SPECIFICATION

Improvements in the encoding of Chinese characters

This invention relates to the encoding of Chinese and other idiomatic character into a form in which they can readily be entered into a computer, typewriter or the like via a simple keyboard.

Chinese characters are not made up in the same way as English words from a relatively small number of Roman letters and since there are a very large number of characters not all of which are admittedly in wide-spread use, it is virtually impossible to design a simple keyboard where a single key is provided for each character. It is therefore desirable to try to work out some method of encoding each individual character into a unique code which can be entered through a keyboard of reasonable size and preferably a keyboard which is of conventional size and construction for use with Roman letters and numbers.

For ease of usage any such encoding method must have relatively simple and straightforward rules which are easy to learn and understand and which are preferably based on common knowledge and everyday usage by Chinese people. In addition, the rules should be relatively simple to remember and should provide an encoding method by means of which characters can be entered in a relatively easy and straightforward fashion through a simple keyboard such as a western languarge keyboard.

The invention has therefore been made with these points in mind and aims to provide a method of encoding Chinese characters which meets these requirements.

According to the invention, there is provided a method of encoding Chinese characters into a form in which they can be entered into a computer or the like in which the character is dissembled into four constituent radicals each of which represents a pronounciable sound, or many radicals are possible, subject to the formation of a total of four radicals, are combined into the first radical and, again subject to the formation of four radicals, a succeeding radical is incorporated into a preceding radical, and in the event that there are less than four resulting constituent parts, after the constituent parts, the final stroke of the character as normally written is used as one of the constituent parts and in the event that there are still not enough total constituent parts to complete the four required, the whole character is repeated as a final constituent part, and thereafter the four constituent parts are represented by the initial four Roman alphabet letters of the transliterations of each constituent part if the part is a character or its link character whilst the part clearly resembles, the code for the Chinese character having therefore the resulting four Roman letter code.

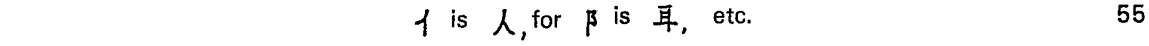
Such as system is relatively straightforward to learn, understand and operate. In particular, the final code will be in the form of four Roman letters and so it is possible to use a western style keyboard for the entry of this code corresponding to the Chinese character into a computer, typewriter or the like and provided the computer, typewriter or the like has been suitably programmed to recognise the codes of each particular Chinese character the appropriate character will be input or printed.

In addition, the chance of ambiguity as between two different Chinese characters having the same code is largely eliminated. Thus, for each of the four Roman letters, there are of course 26 possibilities and practice has shown that the chance of two different Chinese characters ending up with the same alphabet code is extremely low indeed. In addition, a very large number of codes are possible since each of the four Roman letters can be made up of 26 different possibilities and so, in theory, it is possible to encode the order of 450,000 different characters which is well in excess of the number of Chinese characters in use.

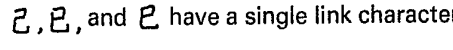
Amongst the advantages of the invention are that the rule for encoding characters are simple, clear and easy to learn and in the breaking up of character into radicals based upon the common knowledge of people who understand Chinese writing. Therefore regular reference to code books is not required. Further codes can be entered whilst touch typing and the keyboard can be programmed to give immediate warning of a mis-typing for a code which does not correspond to a code recognised by the keyboard.

For convenience, the transliteration used for the constituent parts is Pinyin spelling which is the modern form and so the first letter of the Pinyin spelling of each part is preferably used to obtain the code fed into the keyboard.

In the event that any of the constituent parts is not itself a pronounciable character, the popular custom of using a link character which resembles the part is followed, Thus, for exmaple, the link character for



If this cannot be done, the "external configuration" of the character is examined to find a real Chinese character which resembles the part. For example, the three parts



This is the general principle from which all further example are derived.

Thus, the following rules can be adopted:

1. The relative character for an element as a stroke is the name of the stroke. For example, the relative character of

"㇏" is "捺" (Na); that of "丨" is "直" (Zhi);

that of "丿" is "撇" is (Pie); that of "丶" is "点," (Dian) and so forth.

2. The selection is based on customary designations, such as the radical

"亻" having "人" (Ren) as its relative character; the radical "阝" with "耳" (Er).

As the radical "犭" is customarily known as a reverse "犬", its relative character is "犬" (Quan).

"虍" is commonly called the radical of "虎", so it has the relative character "虎" (Hu).

The others are on the analogy of these.

3. The elements in similar forms are grouped into one class and the relative character of the one most in use is used as that of the whole class. For example,

"直" (Zhi) is used as the common relative character of "丨", "刂", "刂", "刂", "儿", and so on;

"雪" (Xue) as that of "⺕", "⺕", "⺕", "⺕". "⺕", etc.;

"羊" (Yang) as that of "⺶", "⺷", "⺶", and so forth.

4. For a few elements for which we are actually unable to select relative characters, we use "O" as their symbol, such as the element

"西" formed by way of "assembling" in the character "德", which is disassembled into "彳, 西, 一, 心".

However, "O" cannot be used in case that the element does not allow further disassembly. "O" cannot be used for the element which can be further disassembled.

To give some examples, the character

"张" (Zhang) can be readily split into the two characters: 弓 and 长

and so the four parts chosen for the code, since these two characters cannot be further split would be:

1	2	3	the last to be written stroke of the character Zhang)	4 (the whole character)
弓	长	丶		张

Therefore these parts can be coded as follows:

	1	2	3	4
	弓	长	丶	张
	Gong	Chang	Na	Zhang
Code:	G	C	N	Z

Other characters such as the character "串" (Chuan)

cannot be divided geometrically but can be separated according to their writing order, e.g.;

口 Kou and 中 Zhong

Therefore the character "串" can be coded as follows:

	1	2	3	4
	口	中	丶	串
	Kou	Zhong	Na	Chuan
Code:	K	Z	N	C

To take another example, this time the character "㝵" (Bin), this is coded as follows:

	1	2	3	4
	亠	回	示	八
		Hui	Zhi	Ba
Code:		H	Z	B

In this example, although the character Hui could be divided into "口" and "口", this could result in encoding the four parts required.

Some further examples are shown in Figure 1 of the accompanying drawings.

The general principles which emerge from Table 1 are:

1. Parts of four strokes or less generally speaking are not dismantled further. For example, see the character "天" (Tian) within the character "吴" (Wu) or the "止" (Zhi) within the character "路" (Lu).

Of course, if the character "天" (Tian) itself becomes a subject for encoding in its own right, then it must be taken apart.

2. Complicated characters must still be reduced to four parts. In order to limit the ambiguities of dismantling, as much as possible is included in the first part of the character and the rest is dismantled. For example, in the character

"兢", the first "克"

is placed entirely into one part whereas the second one is dismantled into several.

3. In the case of "余", we dismantle it into "人十子十八" instead of into "人十 二十小" because the part "二" and the single stroke which follows it "亅" can be assembled together to form a new character "予".

Hence, whenever a part can, by the addition of the succeeding stroke, be transformed into yet another character, that transformation must be adopted, until such time as either four strokes or no new additional parts can be formed by adding the incremental stroke.

The method of the invention is easy to learn and remember and the steps of dismantling characters derives its ideas from popular modes of oral communication which are well known to the masses. This also lessens the burden of study. As to even the codes, though they are taken from the reading pronounciation, they need only be accurate in their first letter.

By limiting the codes to four letters it is easier to transform the codes into machine codes on entry into the computer and for the codes to be processed internally. This regularity and shortness contributes to raising the rapidity of data entry.

The selection of link characters does not depend on the radical-phonetic distinction, but rather proceeds directly from the form of the characters themselves. This raises the speed of distinguishing the sounds and it expands the breadth of a single constituent part's characteristics.

In order to increase the speed of typing a test, reducing the number of key strokes is necessary. Therefore a "High Speed Code" can be used for the most commonly used Chinese characters and character-compounds. The basic concept of this special code is one Chinese character one key stroke. This is, however, limited to 26 Roman letters. By way of example, the frequently used characters shown in Figure 2 of the accompanying drawings can be represented by the single letters shown in Figure 2. Of course for these characters the keyboard can still be programmed to recognise the four letter code as well.

As can be seen from Figure 2, the relative frequencies of these characters is given and they are marked roughly to the relative frequencies of the Roman letters on a keyboard which is itself arranged so that the most frequently used keys are the easiest and quickest to operate.

To further simply and increase the coding, Chinese numbers can simply be given the number of the corresponding Arabic numbers as shown in Figure 3 to create 10 more codes.

The method of the invention can also be expanded to include character expression codes. Such a character-compound can be thought of a Chinese character in the broad sense, with the characters composing the compound treated as the constituent parts of these characters broadly conceived. In this way, all the above principles of code construction can be applied to longer expressions. Examples are shown in Figure 4.

There is a very large number of Chinese character-compounds in the Chinese modern language, and, of course, for all of them codes can be created in this way. Normally however only the most

commonly used character-compounds will be treated this way, e.g. about 140, but it is possible to set up say an additionally 200 such characters for use in specialist applications.

Other high speed codes for character-compounds can be created by combining Arabia numeral codes with the Roman alphabet code as shown in Figure 5 for times and dates. These correspond with the Chinese STC, the Standard Chinese Telegraph Code.

With these "High Speed Codes" we can achieve a fourfold increase of the speed of typing compared with the Normal Code. If we can flexibly combine Chinese character codes with high speed Chinese character and character-compound codes, we can obtain in practice a typing speed of 100 to 120 characters per minute (that is 2.5 key strokes on the average per one Chinese character).

When it becomes possible for Chinese characters to be entered directly as machine codes into a computer, then the possibility arises of Chinese becoming a working language of computers in much the same way as English is presently used as a working language. These codes are manipulated and processed within the machine, but at output time are restored to their original form and printed, e.g. with an ink jet printer. For this to be done in Chinese, it will require a character generator for Chinese which has stored an image of the Chinese character, as well as a display device which can display characters (soft copy) or which is capable of printing characters (hard copy). To combine this type of equipment with the keyboard that can handle on sight encoding is to create an intelligent terminal for Chinese characters. As the typist enters his code he will see displayed the Chinese character on the screen. If necessary, he can have contents transmitted to a main computer connected to the terminal for more complicated processing from which a hard copy may be produced. Or he can use the terminal as a word processor and typewriter.

Examples of the uses of the invention includes:

1. Automated typesetting and editing. Today, advanced countries have all introduced computerized phototypesetting in order to eliminate the hand labour of traditional typesetting art and to reduce the work load of the workers. For Chinese, the difficult point has always been data entry. Our method is solution.

An intelligent Chinese terminal can also greatly reduce the workload of editing of newspapers and magazines. Editorial workers can sit at the terminal, make changes through the keyboard, do proofreading, and even set pages of type using the main computer. They can even exchange texts with other places through computer networks. A printer can be asked at any point to type a clear copy.

2. Mechanized translation, preparation of news indexes. Mechanical translation using the computer, whether from Chinese into foreign languages or vice versa will depend on input and output of Chinese characters. English can be translated mechanically into Chinese. At present Chinese into English requires the use of Pinyin spelling and symbols for the tones. Human beings have to be called upon to write out the characters one by one. If the invention is used to give an internal code then it need only be linked to the addresses of the shapes of Chinese characters in the machine.

Systems of indexing first compress their data into compact form, whether it be manuscripts, archives or other material. After establishing coding categories, the character-compound codes can be used through the computer terminal directly to read material or to ask direct questions using Chinese with the answers to be displayed on a screen.

3. A Chinese language computer, a truly Chinese computer, would be able to have Chinese as its working language with the capability of using Chinese in its systems programs. For example, you could use Chinese to write a mathematical programming language.

4. Management of enterprises and projects with a nation-wide network of computers, production statistics and other important economic indexes can be reported back to the leading offices. This too would require a Chinese language terminal.

5. Applications in the social science. The field of social science now everywhere is using computers with particular success in linguistic research. For example, someone has done very thorough and penetrating research on Shakespeare's works, their style and contents, using the computer, and has produced some proof in the solution of the mystery of who Shakespeare was. Thus computers of today's advanced science are not completely out of touch with literary masterpieces of hundreds of years ago. China's enormous literary tradition, the corpus of commentaries on ancient books, investigations of historical reality, dictionaries and editions, the verifications of authors and so on, all these are areas in which the computer can be of help.

Claims (Filed on 22 July 83)

1. A method of encoding Chinese characters into a form in which they can be entered into a computer or the like in which the character is dissembled into four constituent radicals each of which represents a pronounciable sound, as many radicals as possible, subject to the formation of a total of four radicals, are combined into the first radical and, again subject to the formation of four radicals, a succeeding radical is incorporated into a preceding radical, and in the event that there are less than four resulting constituent parts, after the constituent parts, the final stroke of the character as normally written is used as one of the constituent parts and in the event that there are still not enough total constituent parts to complete the four required, the whole character is repeated as a final constituent part, and thereafter the four constituent parts are represented by the intial four Roman alphabet letters

of the transliterations of each constituent part if the part is a character or its link character which the part closely resembles, the code for the Chinese character having therefore the resulting four Roman letter code.

2. A method as claimed in Claim 1 in which the transliteration used for the constituent parts is Pinyin spelling.

3. A method as claimed in Claim 1 or Claim 2 in which the "external configuration" of the character is examined to find a real Chinese character which resembles the part in the event that a link character resembling the part cannot be found.

4. A method as claimed in any preceding claim in which, in order to increase the speed of typing a text, a special code of one letter for one Chinese character is used for frequently used characters.

5. A method as claimed in Claim 4 in which the frequently used characters shown in Figure 2 of the accompanying drawings are represented by the single letters shown in Figure 2.

6. A method as claimed in any preceding claim in which to simplify and increase the coding, Chinese numbers are given the number of the corresponding Arabic numbers as shown in Figure 3 of the accompanying drawings.

7. A method as claimed in any preceding claim which is applied to character expression codes which can be thought of a Chinese character, with the characters composing the compound treated as the constituent parts of these characters.

8. A computer which has been programmed to accept the entry of a Chinese character coded by a method as claimed in any preceding claims, in which the codes are manipulated and processed within the machine, but at output time are restored to their original form and printed.

9. A typewriter which has been programmed to accept the entry of a Chinese character coded by a method as claimed in any preceding claim in which the character is input on the keyboard in the coded form and output as the character.

Printed for Her Majesty's Stationery Office by the Courier Press, Leamington Spa, 1984. Published by the Patent Office, 25 Southampton Buildings, London, WC2A 1AY, from which copies may be obtained.