

# Coding Design of Oracle Bone Inscriptions Input Method Based on “ZhongHuaZiKu” Database

Dongxin Hu

College of Liberal Arts , Capital Normal University , 10089 Beijing , China  
602201042@qq.com

## Abstract

Based on the oracle bone glyph data in the “ZhongHuaZiKu” database, this paper designs a new input method coding scheme which is easy to search in the database, and provides a feasible scheme for the design of oracle bone glyph input method software in the future. The coding scheme in this paper is based on the experience of the past oracle bone inscriptions input method design. In view of the particularity of oracle bone inscriptions, the difference factors such as component combination, phonetic code and shape code ( letter ) are added, and the coding format is designed as follows : The single component characters in the identified characters are arranged according to the format of “ **structural code + pronunciation full spelling code + tone code** ” ; the multi-component characters in the identified characters are arranged according to the format of “ **structure code + split component pronunciation full spelling code + overall glyph pronunciation full spelling code** ” ; unidentified characters are arranged according to the format of “ **y + identified component pronunciation full spelling + unidentified component shape code ( letter )** ” . Among them, the identified component code and the unidentified component shape code are input in turn according to the specific glyph from left to right, from top to bottom, and from outside to inside. Encoding through these coding formats, the heavy code rate is low, and the input habits of most people are also taken into account.

## 1 Previous design and inspiration of oracle bone inscriptions input method

In the past, some scholars designed the input method of oracle bone inscriptions from the perspective of shape code in coding. For example, Mr. Xu Song of Central China Normal University developed a method in 1995, which applied 26 English letters and 9 Arabic numerals to correspond to more than 500 characters in oracle

bone inscriptions, and realized the input of oracle bone inscriptions by keyboard input characters. By 2012, researchers such as Li Qingsheng of Anyang Normal University jointly developed an input method of oracle bone inscriptions based on the dynamic description library of oracle bone inscriptions. On the basis of the coding and writing specifications of modern Chinese characters, the input side uses the dynamic description method to describe the oracle bone inscriptions with directed strokes and strokes, and combines the extended coding area with the external description character library. It is more effective to solve the input problem of variant characters and unliteracy in oracle bones.


Some scholars have developed image method, visual input method and handwritten input method from the perspective of non-coding to solve the input problem of oracle bone inscriptions. In 1990, Zhou Demin et al. of Henan University first developed the Calculator Oracle Information Processing System ( CJPS ), which laid an important foundation for the subsequent research and development of related input methods. In 2004, Mr. Liu Yongge and Li Qingsheng of Anyang Normal University developed a visual oracle bone inscriptions input method. The principle of the input method is to provide the input person with a table of oracle bone inscriptions. The input person selects the corresponding radicals contained according to the oracle bone inscriptions that he wants to input. The program presents the results containing these radicals to the input person in the form of candidates. The input person clicks on the glyph he wants to input to complete the input. After that, in 2020, Mr. Liu Yongge, Mr. Li Qiang and others from the Key Laboratory of Oracle Bone Inscription Information Processing of Anyang Normal University jointly developed a new oracle bone script handwriting input method. Based on the latest research results of artificial intelligence


deep learning and convolutional neural network, the oracle bone script recognition network and recognition module were developed. The method of using this input method is to operate the mouse to write the oracle bone script that you want to input to the virtual handwriting board to complete the recognition of oracle bone script, and then generate the glyph candidate, and then click the candidate glyph to complete the input of oracle bone script.


From the above content, there are two main problems in the design of oracle bone inscriptions input method in the past. On the one hand, computer professionals only design the input method of oracle bone inscriptions from the two directions of shape code and non-coding, and do not use phonetic code to participate in coding. The reason is that phonetic code can not encode the unidentified characters in oracle bone inscriptions.<sup>1</sup> Whether it is from the perspective of speech or keyboard input, most of us are used to associating phonetic symbols with text. Considering that people whose mother tongue is Chinese or people whose mother tongue is not Chinese, the first thing they learn when learning Chinese is the Chinese pinyin scheme, we think that setting phonetic codes in the input method coding is very convenient. From the perspective of ancient Chinese characters, since oracle bone inscriptions are already identified, they must have a clear pronunciation. The commonly used characters in oracle bone inscriptions are basically identified glyphs, it is precisely these identified glyphs that we often use when inputting characters. According to these, we should not abandon the phonetic code when designing input method encoding.

On the other hand, the shortcomings of using only shape code to encode are generally not convenient for input learners to learn and use. Some are used to input the roots to retrieve alternative characters. This method is similar to the five-stroke input method to split today's regular script characters, but its drawbacks are reflected in the fact that the keyboard is as inconvenient for users to master as the five-stroke input method. It is not in line with the character theory for some

oracle bone inscriptions, and it is not convenient to distinguish the large number of variant characters in oracle bone inscriptions. Some search for alternative glyphs according to the method of stroke input (Nie Yanzhao and Liu Yongge . 2010.) , but most of the modern so-called strokes are applicable to the glyph decomposition of Li and Kai characters, while many of the more pictorial characters in oracle bone inscriptions cannot be described by the concept of strokes. For example, the relatively representative glyphs of oracle bone

inscriptions , etc., strokes cannot truthfully describe the shape at the top of the glyph; the 车

characters of oracle bone inscriptions are  and

.<sup>5</sup> This special glyph of the record segment and the nuances between the glyphs cannot be combined and split simply by strokes. Some combine the similar four-corner number retrieval method with the configuration codes such as closed curve stroke and its extension line structure, cross stroke structure, discrete stroke structure, etc (Liu Yongge and Li Qiang . 2020.) . The **“Oracle Bone Inscription Six-digit Code Search Font Library”** is based on these three aspects as the basis for coding, but this search font library does not contain as many glyphs as ours. The most difficult thing for users is to learn this coding rule, which does not meet our requirements in simplicity and efficiency. Some use the method of dynamic description, based on the coding and writing norms of modern Chinese characters, using concepts and techniques such as directed strokes and pen elements to describe oracle bone inscriptions. (Li Qingsheng, Wu Qinxia, Wang Lei . 2012.) . The premise of this method is that the input must have a deep understanding of oracle bone glyphs, and the writing norms of modern Chinese characters are a kind of rules with strong regularity and serious symbolization, which is not very suitable for oracle bone glyphs with strong realism.

Since oracle bone inscriptions have a high degree of pictography, from the professional point of ancient Chinese characters' view, we hope to provide the academia with a coding scheme that

<sup>1</sup> Although “Yin Qi Wen Yuan” Data Platform (<http://jgw.aynu.edu.cn>) has provided this input method, it does not provide input method software that can be used away from the website, so its coding principle is not clear.

<sup>2</sup> *Jia Gu Wen He Ji* 6816

<sup>3</sup> *Jia Gu Wen He Ji* 27888

<sup>4</sup> *Jia Gu Wen He Ji* 584 front side

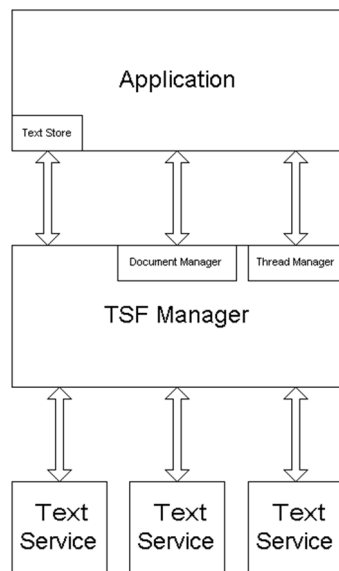
<sup>5</sup> *Jia Gu Wen He Ji* 10405 front side

conforms to the professional cognition of ancient Chinese characters. We urgently need a set of input method coding scheme that can be easily accepted by professionals to the greatest extent, faithful to the correct description of oracle bone inscriptions as much as possible, and convenient for users to learn and use.

## 2 A new design scheme of oracle bone inscriptions input method

### 2.1 Technical route of oracle bone inscriptions input method design

The Oracle Bone Inscriptions Input method is designed using Microsoft 's Text Service Framework ( TSF : Text Service Framework ). It is a COM-based input method framework that does not depend on specific input devices and can support multiple languages. It provides a simple and scalable technology for implementing text input and natural language processing technology. The text service framework includes three main components : application, TSF manager and text service. The architecture is shown in the following figure.



Picture 1 : TSF architecture

“**Application**” refers to the application software that supports and has adopted TSF, such as Microsoft 's MS Office, Notepad and other word processing programs. The application accesses text by implementing a COM server that supports a specific interface, and communicates with TSF by using an interface exposed by the TSF manager. Applications that support TSF do not need to consider the specific details of the input method, and can receive text input from the “text service” to achieve a series of operations such as displaying, editing, and storing text.

“**Text service**” refers to the text input processor, which can be keyboard input, handwritten recognition input or speech recognition input and other input programs. After registering with TSF, users can use language bar or keyboard shortcuts to interact with the text service. The text service can obtain text from the application or write text to the application. Text services can also associate data and attributes with text blocks. The oracle bone inscriptions input method implemented in this paper is a text service that inputs oracle bone inscriptions characters through the keyboard.

“**TSF Manager**” is an intermediary between an application and one or more text services, implemented by the operating system to enable applications and text services to share text. The text service does not interact directly with the application, and all communications are performed through the TSF manager.

Oracle Bone Inscriptions Input method implements the basic elements of TSF, such as Thread Manager, Client Identifiers, Document Manager, Edit Context, Ranges, Compartment, Properties and Composition.

The “**Thread Manager**” is responsible for completing the task of connecting the application and the text service. These tasks include activating or suspending the TSF text service, creating the document manager, and maintaining the correct association between the document and the input focus.

The “**client identifier**” is an identifier assigned by the thread manager that is received and must be maintained by clients such as applications and text services. The client needs to provide its own identifier when calling various TSF methods.

The continuous text stream created by the “**edit context**” through the interface can be created by the application and provided to the text service. In some cases, the text service can also create an edit context as needed.

The “**document manager**” is responsible for maintaining the last-in-first-out buffer, and the content stack stores the list of edited content managed by the document manager.

An “**input combination**” is a temporary input state that enables the text service to keep the application and user input text in a state of constant change. The application can obtain the display attribute information of the input combination and use this information to display the input combination state to the user. The application

determines how to display the text and what kind of operation to the text according to whether there is an input combination.

## 2.2 Coding scheme designed by oracle bone inscriptions input method

Professor Huang Tianshu of Tsinghua University presided over the press and publication of major scientific and technological projects of the Chinese characters project “中华字库- Oracle bone inscriptions collection and collation” ( 0610-1041BJNF2328-03 ), its network platform has been sorted out 12685 Oracle bone inscriptions, the number of this glyph is the past Oracle Bone Inscriptions Input method can not be compared. We divide the 12685 oracle glyphs into two categories : literate and unilliterate, and encode them separately. Among them, literate is divided into two categories : single component and multi-component. Unilliteracy is divided into three categories : components that are identified but not identified as a whole, some components can be identified but the rest are not identified, and components are not identified at all.

### 2.2.1 Coding scheme of identified glyphs

#### 2.2.1.1 Coding scheme of single component character

On the basis of “ natural classification ” , Mr.Huang Tianshu summarized and sorted out the radicals of oracle bone inscriptions into four categories :“象物” “象人” “象工” 和 “other” (Huang Tianshu. 2020.) . The “象物” refers to all non-living and living things in nature, “象人” refers to the shape of people and their five senses and limbs, “象工” refers to the products of human wisdom, “other” refers to parts that cannot be classified .For the input method itself, a constraint condition is added to the limited coding position, which can greatly reduce the repetition rate. Therefore, we roughly divide the structure of single-component characters into four categories : 象物, 象人, 象工, and others. According to the full spelling of the first letter of the character “物”, and at the same time, in order to distinguish it from “合文” in the following text, “v” is used to refer to “象物” . According to this setting method, the other three types of codes can be set as follows: “象人” corresponds to“r”,“象工” corresponds to“g”, and “other” corresponds to“t” ( “other” in Chinese is “其他”, based on the full spelling of the first letter of the character “他” is “t”) . It may

be the first time in the history of input method development to classify and encode single component characters by natural classification.

We set this type of coding in the first place of the input order. Because the identified characters in oracle bone inscriptions generally have corresponding interpretation opinions, there will be corresponding pronunciations of regular script characters in later generations. This pronunciation also belongs to the important coding attribute of single component characters in oracle bone inscriptions, so we set the corresponding Chinese pinyin spelling in the second place of the input order. From the perspective of reducing the repetition rate, under the constraints of the first two codes, sometimes there may be situations where the accuracy of the alternative characters is not enough. For example, under the premise that the expected input phonetic code “you” is added, the input situation is divided into vyou , ryou , gyou , tyou. The corresponding vyou codes are “柚” “困” etc., and the corresponding ryou codes are “又” “尤” etc. Corresponding to gyou codes, there are “酉” “卣” etc., corresponding to tyou codes, there are “缶” “由” “猷” etc., and it can be seen that the repetition rate is already very low, but there are many variant characters of the same character in oracle bone inscriptions. After entering the coding, the number of options becomes the number of variant characters of one character plus the number of all variant characters of another or two characters. The number is still a lot. In the input method software or patents that have appeared, there is no design for encoding tones. The tone symbols and corresponding codes we designed are listed below :

Intonation	Coding
High-level tone (first tone)	y
rising tone (second tone)	p
falling-rising tone (third tone)	s
falling tone (fourth tone)	q

Table 1 : Tones table

If we add the attribute difference of tone on this basis. Coding becomes vyouy, vyoup, cvyouy, cvyouq, ryouy, cryoup, ryous, ryouq, gyoyy, gyoup, gyous, gyouq, tyoyy, tyoup, tyous, tyouq and so on. In this way, in the previous coding, except that the characters “酉” and “卣” under the “gyou”

code are not distinguished, the characters under the other three codes can be distinguished. Therefore, the coding format of “**structure code + pronunciation full spelling code + tone code**” is completely feasible.

In the following, we show some practical examples of single component character coding.

Intermediate code of “中华字库”	Oracle glyphs	Input coding
0E9C7B		gzheny
0E9C86		trenp
0E9C88		vyueq
0E9C79		tdingy
0E9C75		gzus
0E79E0		vyangp
0E79D1		rhuangp
0E86DA		vzhiq
0E86E9		vlaip
0E86D6		rruoq

Table 2 : Single component characters coding table

### 2.2.1.2 Coding scheme of multi-component characters

The classification of multi-component characters is the most detailed. We divide the structure of multi-component characters in oracle bone glyphs into 14 categories, and according to the general shape of the structure, it is coded with English letters with similar shapes : the left and right structure correspond to the position of “”, and the corresponding code is “**h**” ; the corresponding position of the upper and lower structure is “”, and the corresponding code is “**z**” ; the corresponding position of the full inclusion structure is “”, and the corresponding code is “**o**” ; the corresponding position of the upper three-inclusion structure is “”, and the corresponding code is “**n**” ; the corresponding position of the lower three-inclusion structure is “”, and the corresponding code is “**u**” ; the corresponding position of the left three-inclusion structure is “”, and the corresponding code is “**c**” ; the right three-inclusion structure corresponds to the position of “”, and the corresponding code is “**b**” ; the corresponding position of the upper left contains structure is “”, and the corresponding code is “**p**” ; the corresponding position of the upper right contains structure is “”, and the corresponding code is “**q**” ; the corresponding position of the lower left inclusion structure

is “”, and the corresponding code is “**l**” ; the corresponding position of the lower right inclusion structure is “”, and the corresponding code is “**j**” ; the corresponding position of the covering structure is “”, and the corresponding code is “**f**” ; The corresponding position of the upper, middle and lower structure is “”, and the corresponding code is “**e**” ; the corresponding position of the left, middle and right structure is “”, and the corresponding code is “**m**”. The above-mentioned glyph structure and the corresponding coded letters are set up on the basis of the principle that the structural form of the first-level component is as close as possible to the letters .

We believe that the full-spelling syllable coding, which conforms to the typing and recognition habits of modern Chinese people, is very suitable for encoding multi-component characters, but it is also a problem to match the full-spelling coding with what coding. Tone coding is very suitable for distinguishing single-component characters. If multi-component character coding is designed to use tone coding on the basis of full pinyin syllables, the distinguishing ability of this coding will be greatly reduced. For example, in oracle bone inscriptions, the multi-component characters with pronunciation of fú are supported, 扶, 符 and 𠂔 etc., the three-character components are completely different but cannot be distinguished. In addition, some of the components in the multi-component characters have been deformed and voiced during the evolution of the glyph. The changes in this component can not be distinguished by the pronunciation of the characters. For example, 𠂔 characters are generally from the same 𠂔 deformed into 𠂔 ; 彳 left component is deformed into 𠂔 (簞) 𠂔, and the overall evolution is 𠂔. The above two aspects of coding problems, “**full spelling + tone**” method is not able to solve.

Therefore, we add two coding items, the whole glyph pronunciation and the disassembly component pronunciation, to the structural coding, that is, the coding format of “**structural code + disassembly component pronunciation full spelling + whole glyph pronunciation full spelling**” . This three-stage coding design scheme for multi-component characters of Chinese characters may be original.

In the following, we show some practical examples of multi-component character coding.

Intermediate code of “中华字库”	Oracle glyphs	Input coding
0E9C83	𠄎	zrourouduo
0E79D3	𠄎	zzhiwangwang
0E79DC	𠄎	zshengmuxing
0E79D9	𠄎	nmianwanbin
0E86E4	𠄎	hyiliyi
0E9C42	𠄎	nmianshizong
0E860E	𠄎	mchiwujieyu
0E7342	𠄎	ezhiweizhiwei
0E95E6	𠄎	ozhubei
0E9965	𠄎	uzhikanchu
0E8765	𠄎	pyanziyou
0E94C5	𠄎	fmeigemie
0E86AF	𠄎	fjiannvyan
0E7BD2	𠄎	broushitun
0E75F0	𠄎	oweichuangzang

Table 3 : Multi-component characters coding table

However, there are some of the multi-component characters that can be analyzed for structure and components, but they cannot identify the overall pronunciation of the characters. For the convenience of coding, we classify all these characters into the category of “**unidentified characters**” during coding, such as 𠄎 can be analyzed as 犬 and 大, 𠄎 can be analyzed as 爻 and 米, 𠄎 can be analyzed as 目 and 口, but these are not the exact overall pronunciation, we for the convenience of coding, this kind of multi-component characters into the category of literacy.

There are a large number of “**合文**” in oracle bone inscriptions. This kind of glyph refers to the phenomenon of combining several original independent glyphs into one glyph. For “**合文**”, although the identity of each part of the new glyph is a character rather than a component, the combination of “**合文**” is actually similar to “**multi-component character**” in terms of structure. In order to take into account the independent and common characteristics of “**合文**”, we regard “**合文**” as an input method structure category that can be independently classified, and the subsequent coding writes the pronunciation of each part according to the reading order of “**合文**”, and the coding format is roughly “**w + split component full spelling**”,

The coding order of each character in “**合文**” is arranged according to the order of reading.

When encoding the “**合文**”, we need to pay attention to the following aspects : some combinations of “**合文**” are connected or even have overlapping parts, such as 大 and 丁 is 𠄎, 柚 and 京 is 𠄎, 上 and 甲 is 𠄎, 三 and 牛 is 𠄎, and so on . Some combination methods are similar to the “**借笔**” in the combination characters, such as 大 and 甲 is 𠄎, 五 and 璧 is 𠄎, 五 and 牢 is 𠄎, 妣 and 丙 is 𠄎 and so on. Although some two characters are separated, the “**character spacing**” is slightly closer than the normal character spacing, such as 大 and 庚 is 𠄎, 母 and 癸 is 𠄎, 祖 and 己 is 𠄎 and so on , this is also the most common “**合文**” . In some combination forms, one of the components is separated, and even the separated components are quite close to the other character. This kind of combination is not easy to identify, such as 武 and 乙 is 𠄎, 三 and 牡 is 𠄎, 龐 and 母 is 𠄎, 武 and 丁 is 𠄎 and so on. Although some of the combined texts are separated from each other, one of the characters is simple, such as 多 and 子 is 𠄎. In addition, in order to facilitate the explanation of the rules of the input method, we also incorporate the situation of “**重文**” into the category of compound characters. At present, only one phenomenon of “**重文**” is found in oracle bone inscriptions, that is, 有 and 佑 duplicate characters is 𠄎.

In the following, we show some practical examples of “**合文**” coding.

Intermediate code of “中华字库”	Oracle glyphs	Input coding
0E8848	𠄎	wzuding
0E8B2E	𠄎	wshangjia
0E7D0E	𠄎	wshiyiyue
0E9A2E	𠄎	wbaoyi
0E9A20	𠄎	wwubi
0E7D90	𠄎	wxiaogao
0E8F1F	𠄎	wyoujing
0E9A82	𠄎	wfuding
0E977E	𠄎	wxiaolao
0E94E3	𠄎	wyouyou
0E8128	𠄎	wsanniu
0E91FB	𠄎	wliuyue
0E79C1	𠄎	wwushi
0E7DCD	𠄎	wduozi

Table 4 : “**合文**”characters coding table

## 2.2.2 Unidentified glyph coding scheme

### 2.2.2.1 All components are identified but the whole character does not identified

This part is relatively simple, although we do not identify the pronunciation of the glyphs, do not identify which glyphs they correspond to later generations, but each component in the glyphs is identified, so this part of the code can be coded according to the “y + component” format, and the pinyin of the component can be written in the order from left to right, from top to bottom, and from outside to inside.

In the following, we show some practical examples of coding.

Intermediate code of “中华字库”	Oracle glyphs	Input coding
0E8618		yyanripu
0E7786		yyanripu
0E7966		yriyan
0E7D7E		yrishi
0E79A2		yriyuan
0E8F54		yyuhuo
0E906C		yyuji
0E7F03		ybaohuo
0E93D9		yzhuilihuo
0E9157		yzhebuhuo

Table 5 : All components are identified characters coding table

### 2.2.2.2 Some components can be identified but the rest do not.

In addition to the glyphs that are clearly fit structures, we forcibly separate the uncharacterized glyphs that may be part of the single body into several parts for the convenience of the input method design. The coding order of this part is written in accordance with “y + identified components + unidentified components”. The order of identified components and unidentified components is also arranged in the order from left to right, from top to bottom, and from outside to inside. The unidentified components are represented by the selection of 26 letters similar to their shapes according to the specific glyphs.

In the following, we show some practical examples of coding.

Intermediate code of “中华字库”	Oracle glyphs	Input coding
0E7D86		ytianoo
0E734D		ymunn

0E75E6		ycai jie
0E9D74		yhzhui
0E8858		yyux
0E79CB		yyangmumin
0E9FF4/0E9118		yiikou
0E7353/0E73A7		yyykou
0E8109		yrioda
0E735D		yzuoyou
0E7E94		yodao
0E8FB0		yochu
0E967A		ypanren
0E9683		yshuio
0E82FA		yxjie
0E73BF		ymjiewang
0E7D2D		yygan
0E8675		ywxing
0E7F13		yfuy
0E933C		ymuyx
0E95B2		ykoukoux
0E74F1		ywda

Table 6 : Some components can be identified characters coding table

The middle part of the font of 0E7D86 is like “田”, and the closed semicircle on both sides is replaced by two “o”.

The lower part of the glyph of 0E734D is “目”, and the upper part looks like two upward raised curved pens, so it is replaced by two “n”.

The left side of the font of 0E75E6 is “才”, and the right side is both an undetermined and inseparable component. The component on the right side is composed of the head of “鷹” and “卩”. In this case, if the coding is designed according to this splitting, the coding will become very long, so we choose the “卩” which is easier to identify to replace the component on the right side.

The lower part of the glyph of 0E9D74 is “隹”, and the upper part looks like “H”.

The upper part of the glyph of 0E8858 is “雨”, and the lower part does not know what animal it refers to. Because there are cross strokes in the lower part of this component, and the repetition rate of coding “yyux” is very low, we use “x” to replace the following components.

The glyph periphery of 0E79CB may be “皿”. Although the inner component does not know what it is, it can be forcibly disassembled into two parts : “羊” and “木”.

The left side of the glyph of 0E967A is “另”, and the right side of the component is not “人” but still like the shape of human.

The glyph of 0E73BF has half part of the “王”, and the left half is like the kneeling figure. The upper side of the left component that cannot be split is like “m”, so it can be forced to split into “m+β”.

Part of the glyph of 0E8675 is an obvious “行”. the upper side of the non-separable part that almost encloses the “行” is like “w”, the lower side is like “I”.

0E933C may be a single font. The top is “目” and “巾”, and the lower part can no longer be split. But the part like “巾” is not “巾”, so we use the approximate trident “Y” instead. Because the repetition rate of encoding “ymuyx” is also very low, the lower part is replaced by “x”.

### 2.2.2.3 Components completely unidentified

In this case, we can only split these glyphs into several parts, and arrange the letters similar to each part according to their order in the glyph. In order to make the input of this part of the glyph easier, we try to arrange the parts with roughly similar shapes in the same letter as much as possible.

In the following, we show some practical examples of coding.

Intermediate code of “中华字库”	Oracle glyphs	Input coding
0E774A		yooox
0E94F7		youx
0E8AB9/0E97C5/0E72FC		youuy
0E7E8E		ym
0E8482		yy
0E999E		yi
0E7BB3		yii
0E7180		yam
0E87B9		yyooa
0E8B29		yooy
0E8B2D		yuooy
0E7EDC		yox
0E9A2D/0E8275/0E96CF		yox
0E778B		yummy
0E832D		yco
0E97B3		yhh
0E717C		yh
0E8283		yi
0E813C		yo
0E899D		yww
0E9740		yuo
0E9822		yyooo
0E7E23		yoooy

<sup>6</sup> For the convenience of writing, we use the intermediate code to replace the original character, and the corresponding character can refer to the above table.

0E850A		yiiiix
0E7687		yuu
0E8326		ys
0E8327		ys
0E97F8		yk
0E8E6F		yk
0E9172		yx
0E9096		yy
0E78F8		yl
0E78FD		yoo
0E8C8A		yl
0E9F9B		yy
0E9FC3		yh
0E98B2		yui
0E7814		yeo
0E8680		yto
0E7E64		yox
0E7EAA		yi
0E7E3D		ym
0E8100		yuu
0E761F		yl
0E80F6		yj
0E7F29		yi
0E818E		yf
0E89F5		ycj
0E8A23		ym
0E8E21		yomx
0E8415		yoy

Table 7 : Components completely unidentified characters coding table

In order to facilitate the reader to understand our ideas, we split the description of complex characters . Due to space constraints, we list some of the more special examples to illustrate.

According to the strokes, we can see that the upper two curved pens of 0E774A<sup>6</sup> form three rings, and only “o” is a ring in the 26 letters. Therefore, we have compiled three “o”, and there are two crossed strokes. The image of “x” is more consistent, so this character is coded as “yooox”.

From top to bottom, 0E94F7 is an image of a ring, a “L” shape, two eyes, and a combination of a person’s upper limb and a frog’s lower limb. This glyph has many and obvious distinguishing features. For the convenience of input, the part of the eye shape that can not be encoded. In other aspects, the “L” shape can be encoded by “u” , and the rest of the glyph is similar to “☆”, but the 26 letters are not similar to it, so the “x” with cross



stroke features is used to encode. So this character is coded as “youx”.

From top to bottom, 0E8AB9 is a ring, two “□” shapes, and the rest can be seen as an inverted “Y” shape, so the character is encoded as “youuy”. The glyphs 0E97C5 and 0E72FC look similar to 0E8AB9, so we think these three characters can use the same code.

The shape of 0E8482 is similar to that of oracle bone inscriptions “步”, but it should not be the same character. The whole shape of this character is three-line intersection, so it is encoded by the letter “y”.

The shape of 0E7180 is originally a single body, but in order to facilitate input and avoid coding repetition, we divide this font into two parts. The top tip shape is like “A”, so we use “a” to encode it. The lower part is like a bird spreading its wings, which is similar to “M”, so we use “m” to encode it.

The shape of 0E8B2D from top to bottom is approximately “□”, two circles, inverted triangle, and other strokes. A vertical stroke is connected under the triangle below, and they are combined together to be similar to the “Y” shape, so the word is encoded as “yuooy”. Similarly, the coding of 0E8B29 is “yooy”.

The glyph of 0E7EDC can be divided into two parts. The periphery is a circle, and the inside is three lines that intersect at the same point. The intersecting lines can still be encoded by “x”. 0E9A2D, 0E8275 and 0E96CF all have similar characteristics, like the larger version of “田”. The periphery of the three is basically closed and can be coded with “o”. There are many dry cross lines at the center of the font, which can be coded with only one “x”. We use the same coding on the two types of glyphs, which may lead to high repetition rate, but there are few cases similar to 0E7EDC glyphs, and there are not many uncharacterized glyphs similar to 0E9A2D, 0E8275 and 0E96CF. Therefore, these two types of glyphs are easy to distinguish in the input process and will not affect the efficiency of input.

Both 0E8326 and 0E8327 are on the 合集 22507, and it remains to be further investigated whether they are glyphs or characterization symbols. However, the shape and composition of the two are very strange. Like today’s one-stroke, it is not common in oracle bone inscriptions. Because we use “s” to encode the curved linear components of

the rope shapes in other glyphs, we also use “s” to encode here.

### 3 Conclusion

According to our internal test program, the above coding design is indeed feasible, and the coding repetition rate is very low, which is conducive to accurate search. The number of glyphs involved in the input method coding scheme we designed is unprecedented, so our coding design will be closer to the real situation of oracle bone inscriptions than previous coding designs. We try to provide a coding scheme for the input method in line with the professional cognition of ancient Chinese characters. Therefore, we are different from the previous design: the coding design is carried out for different types of Oracle glyphs, and the concept of “natural classification” is added to the coding of single component characters. For the unidentified glyphs, we also imitate the multi-component characters as much as possible to carry out the separation in line with the cognition of ancient Chinese characters to encode, and use the English letters to refer to the unidentified parts with similar shapes. Not only that, we have also implemented the coding form of “shape code + phonic code” that has not been tried in the past.

The remaining number of variants that are not often used is not a lot of unidentified glyphs. Although some of these are not encoded according to the knowledge of ancient Chinese philology, there are still general rules to follow. For example, we use “o” to refer to the closed form component, “x” to refer to the cross part of the two lines, “y” to refer to the trident part, “x” to refer to the unidentified component that cannot be split without increasing the repetition rate, and so on. The design of o, x and y is similar to Oracle Bone Inscriptions Six-digit Code Retrieval Font, but we refer to it from the perspective of “component” of ancient Chinese characters, not using the concept of “stroke” of subsequent of Chinese characters. Nevertheless, it still takes a lot of effort to form a regular coding design for the unidentified glyphs.

At present, the learning manual matching the input method formed on the basis of this coding design is still in preparation, and the preparation of the learning manual is the subject we will study next. In the following research, we will improve the part of the above coding design that is not convenient for fast input, and try to fit the coding rules with strong regularity as much as possible for the unidentified glyph part.

## Acknowledgments

Thanks to Mo Bofeng, Liu Ying, and Li Aihui for their guidance on this research. Thanks to Mr. Deng Jian for his guidance in computer technology. Thanks to the anonymous reviewers for their valuable comments on this paper.

## References

- Huang Tianshu. 2020. *A Preliminary Study on the Side and Head of the Oracle Bone*. *Collection of Huang Tianshu 's Oracle Bone Science*. Zhonghua Book Company. page172-181.(in Chinese)
- Li Qingsheng, Wu Qinxia, Wang Lei . 2012 *Oracle Bone Inscription Input Method Based on Oracle Bone Inscription Character Dynamic Description Library*. *Chinese Journal of Information Issue* 4. (in Chinese)
- Liu Yongge and Li Qingsheng . 2004. *Design and Implementation of Visual Oracle Bone Inscription Input Method* , *Computer Engineering and Application Issue* 17. (in Chinese)
- Liu Yongge and Li Qiang . 2020. *Summary of Oracle Bone Inscription Input Method* . *Yindu Journal* . Issue 3. (in Chinese)
- Liu Zhixiang and Liu Xiaorong . 2019. *Oracle Bone Inscriptions Six-digit Code Retrieval Font*. Sichuan Dictionary Publishing House. (in Chinese)
- Nie Yanzhao and Liu Yongge . 2010. *Free Stroke Input Method of Oracle Bone Inscriptions* , *Chinese Journal of Information*. Issue 6.(in Chinese)
- Xu Song and Hu Jinzhu . 1995. *Realization of Oracle Image Code Input Method* , *Journal of Central China Normal University ( Natural Science Edition )*. Issue 3. (in Chinese)
- Zhou Demin, Wang Guoan, Zheng Tongbin, Su Yue, Li Feng . 1990. *Design and Implementation of Computer Oracle Bone Inscriptions Information Processing System ( CJPS )*. Henan Science and Technology. Issue 1. (in Chinese)