2009 International Conference on Electronic Computer Technology

A Structure Character Modeling for Chinese Character Glyph Description

Shixiao Wu, Shijue Zheng
Department of computer science
Central China Normal University
Wuhan, China
e-mail: limr2011963@yahoo.com.cn, zhengsj@mail.ccnu.edu.cn

***Abstract*—The major problem existing in current Chinese character glyph descriptions is the standard character set can't cover all possible Chinese characters, some particular variants of characters that are "unified" also can't describe and display. Character Description Language (CDL) can help resolve both of the difficulties just mentioned. This paper presents the key features and syntax of the language, and discusses some of its applications, especially to character encoding standards work and Chinese writing teaching.**

Keywords-component; Chinese character glyph description ; CDL; Character code; Chinese writing teaching

I. INTRODUCTION

For many overseas students, the most significant snag to learning Chinese is Chinese writing[1]. The number of Chinese characters contained in the Kangxi dictionary is approximately 47,035, although a large number of these are rarely used variants accumulated throughout history[2]. The quality of Chinese character is numerous, and also its structure is complicated. As one of the research interest to Chinese information processing, Chinese character glyph description offers a great convenience to overseas students who want to learn Chinese character writing.

Speaking of Chinese character glyph description, it occured us to chracter code. Character code is a mapping, often presented in tabular form, which defines a one-to-one correspondence between characters in a character repertoire and a set of nonnegative integers. Unicode & ISO 10646 is one of the well-known character codes in the world, it is a standard by the Unicode Consortium. Chinese character "一" by Unicode is U+4e00.

Character code treated the Han character as a whole, but that losed sight of its internal composition. In fact, we can decompose Chinese characters into components and strokes, that has been good for overseas students to find out the structure of Han character. Methods about Chinese character glyph description have a lot, but all in all there are five different kinds. We will elaborate that in section 2.

The problem we want to work out in Chinese charater description is the standard character set can't cover all possible Chinese characters, some particular variants of characters that are "unified" also can't describe and display. Character Description Language co-created by Tom Bishop and Richard Cook can resolve these questions. As a database language based on XML, a set of 39 strokes allow to construct a set of 1,000 components which allow to construct tens of thousands characters' descriptions. A change in the shape of one of the 39 basic strokes will be instantly visible on all other character including it[3].

we proposed the existing problem in section 1. The next part we will talk about work on Chinese character glyph description. Section 3 we expound the key feature and syntax of CDL and discuss its appliction. A Chinese character case depicited through tools called wenlindemo341 will be given in secton 4. Conclusion about our work is in part 5. References are given in section 6.

II. RELATED WORK

Work on decomposing Chinese character into components and strokes is currently led by five different ways as follows.

Han Character Information Dictionary[4], which is published by the Han character encoding group of Shanghai Jiao Tong University. It encodes essential information, includeing stroke count, stroke type, stroke order, component analysis, radicals and residual strokes, and coordinates of strokes.

Components Specification of Han character[5], providing 560 basic components as encoding units for 20902 Han characters to regroup.

IDS(Ideographic Description sequence), which describes how a Han character looks like. That is, if a specific Han character is not encoded in Unicode yet, you could use IDS to describe how the character looks li- ke. IDS consists of description characters and Han characters.

CDPL (Character Document Processing Language), a description language developed for ancient books classification, Taiwan Academia sinica proposed[6]. More than 1000 etymons is defined, over 4000 components is included.

Besides, Sun Xingming, YIN Jian-Ping, etc, proposed a novel mathemetical method to describe a Chinese character. Using this method, 505 components and 6 well-defined

978-0-7695-3559-3/09 $25.00 © 2009 IEEE
DOI 10.1109/ICECT.2009.131

operators can express all the Chinese characters successfully[7].

These five methods proposed above are all related to components, but the structure of Chinese character is complex and easily cause ambiguous. For example, "卡" can (which menns stuck) is up-down structure, it can be described by "上" and "卜", also can be depicted "⊦"and "下". The same Character could be displayed differently by different interpreters. Also these description methods can not cover all exsiting Hanzi that are encoded and can not display particular variants of characters that are "unified"(treated as equivalent).

Then CDL can help resolve difficulties just mentioned. Compared with even the largest standard character set, CDL provides more precision: the ability to distinguish between unified variants. It also provides wider scope: a potentially infinite number of Han characters. CDL can also be used for describing and displaying characters that are not in any standard character set. The CDL instructions can be composed whenever the need arises (preferably using a graphical user interface), and included directly in a document using XML syntax.

III. CHARACTER DESCRIPTION LANGUAGE

Character Description Language (CDL) is for accurately describing and displaying the forms of all Han (CJKV) characters. This part we will presents the key features and syntax of the language.

A. The set of Basic Stroke Types

Due to the limitation of length, we only lists the set of 10 basic stroke types in Table 3.1, currently implemented in the CDL descriptions of more than40,000 ISO/IEC 10646 "CJK Unified Ideographs" are 39 types. The eleven headers A..K in Table Ⅰ are as follows[8]:

- A Sequential numbering [1..10] of all current types;
- B Numeric index for the 5 札 zhá types [1..5], with alphabetic sub-types [a..z];
- C Total number of 折 zhé 'transitional bends' (+1
- = number of segments) in the type;
- D Total number of control points currently implemented for the type;
- E Frequency of this type incurrent descriptions, as
- a percentage of total;
- F Glyph exemplifying the type in isolation (outside of compounds);
- G Provisional assignment of an ISO/IEC 10646& Unicode UCS (Unified Character Set) Scalar Value for each exemplar in F, or PUA (Private Use Area) for unencoded forms;
- H Name of the type in Han characters;
- I Romanization in pnyn of H;
- J Abbreviation for the pnyn name of the type in I (acronymic, except for 39);
- K Notes on the type, including structural analysis (not necessarily tied to the actual implementation), unified variants of the type, examples of usage in compounds, and cross- references to similar types.

TABLE I. THE SET OF 10 BASIC STROKE TYPES

A	B	C	D	E	F	G	H	I	J	K
1	1a	0	2	26.87	一	U+4e00	横	héng	h	horizontal;as in 大，木，三；
3	2a	0	2	15.77	丨	U+4e28	竖	shù	s	vertical;stroke 2 of 下, first stroke of 卜，stroke 3 of 丫
5	3a	0	2	12.54	丿	U+4e3f	撇	piě	p	falling to left, not very curved； as 1st stroke in 八, stroke 1 of 九
8	4a	0	2	09.59	丶	U+4e36	点	diǎn	d	taper + clockwise curve；as in 为; something to left, as 1st in 火
12	4e	1	3	00.11	乀	U+4e40	提捺	tí-nà	tn	提 tí+捺 nà; last stroke in 入,之,辶;
16	5c	1	3	03.28	𠂉	U+4e5b	**横钩**	héng-gōu	hg	一 h+left hook; 2 in 写，冗，军，农，冠，冗
17	5d	0	2	02.54	乚	U+200ca	竖折	shù-zhé	sz	一 h+丨 s；1st stroke in 山，or as 𠂆(丨 s+ 一 h) in 乐，东，互
21	5h	1	3	00.11	𠃌	U+21fe8	撇点	piě-diǎn	pd	丿 p+丶 d; stroke 1 in 女，妯，娌
28	5o	2	4	02.22	𠄎	U+200cc	**横折钩**	héng-zhé-gōu	hzg	一 h+丿 sg; 1st stroke in 力，2 in 月，stroke 3 of 舟
32	5s	2	5	01.84	乛	U+4e5a	**竖弯钩**	shù-wān-gōu	swg	乚 sw+up hook; as in 屯，儿，心

B. *Description of A Han Character*

CDL is an XML application, which means that it conforms to a widely-used standard syntax (usage of angle brackets < >, et cetera).

Here is a description for "行":

<cdl char= "行" >
<comp char= "彳 " points="0,0 40,128" />
<comp char= "亍" points="60,12 128,128" />
</cdl>

Positions are given as points with two-dimensional coordinates. The square enclosing the entire character has (x, y) coordinates ranging from (0, 0) for the top left corner, to (128, 128) for the bottom right corner.The numbers after "彳 " describe its bounding rectangle on the left side of "行": (0, 0) is its top left corner, and (40, 128) is its bottom right corner.

In order for the above CDL description to be carried out as a set of instructions (e.g., for displaying the character or counting its strokes), it is necessary for the interpreter to refer to the separate descriptions of the components, "彳 " and " 亍 ", as sequences of particular stroke types with specific coordinates.

Here is a description for "彳 ":

<cdl char="彳 ">
<stroke type="p" points="107,0 10,46" />
<stroke type="p" points="128,38 0,83" />
<stroke type="s" points="86,70 86,128" />
</cdl>

There are three strokes in 彳. The first two (from top to bottom) are both type 'p', which stands for 撇 piě, a curved stroke falling to the left. The third stroke is type 's', which stands for 竖 shù, a vertical falling stroke. For each of these simple stroke types, only two points are needed. For example, the first stroke starts at (107, 0) and ends at (10, 46).

C. *From Components to Strokes*

CDL uses flexible strokes to describe Chinese characters glyph. There is a form of recursion implied by CDL. For example, a description of 龍 has 16 strokes, its simplified character is 龙 [lóng] dragon. 龍 may refer (with a component) to a description of 立, which in turn may refer(with another component tag) to a description of 亠, which describes two individual strokes. A CDL interpreter will therefore typically process components within components within components, using recursive algorithms. Recursion stops when stroke elements are reached.

Any CDL description that uses comp elements can be transformed automatically into a description that uses only stroke elements. For example, 明 is described as a sequence of two components 日 and 月, each of which is in turn described as a sequence of 4 strokes. Alternatively, 明 could be described directly as a sequence of eight strokes like 3.2 we have mentioned. A straightforward recursive algorithm can transform the component description into the "strokes-only" descrition. Component descriptions are more generally useful as well as more concise.

D. *CDL and Unicode*

CDL is based on Unicode, the difference between them is the former has its own font database, the latter has Universal Character Set.

Unicode consists of a repertoire of more than 100,000 characters, a set of code charts for visual reference, an encoding methodology and set of standard character encodings, an enumeration of character properties such as upper and lower case, a set of reference data computer files, and a number of related items, such as character properties, rules for normalization, decomposition, collation, rendering and bidirectional display order.

The Unicode Consortium want to replace existing character encoding schemes with Unicode and its standard Unicode Transformation Format (UTF). Unicode can be implemented by different character encodings, the most commonly used encodings are UTF-8 (which uses 1 byte for all ASCII characters, which have the same code values as in the standard ASCII encoding, and up to 4 bytes for other characters). XML treated UTF-8 as its default character encoding form.

CDL is a character description language bulid on XML and Unicode, it provides wider scope than Unicode. By decomposing characters into strokes, CDL can cover all possible characters that existed.

IV. EXPERIMENTAL RESULT

In this section we use wenlindemo341[9] to display a hand-writing Chinese character called "土"(which means dust). First we use brush tools to write the character, just like Figure 1:

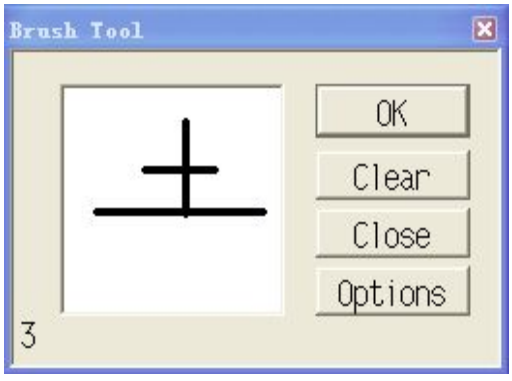

Figure 1. Using brush tools to write the character

From the bottom left corner we know this character has 3 stroke counts. Then we press ok to get lists in connection with "土" as follows:

- list characters containing 土 as a component
- list words containing 土

- list words starting with 土 (in alphabetical order)
- search files for 土
- radical 32 土
- Unicode 571f (GB cdc1) (Big5 a467)

The learners will find out "土" has 3 strokes: 一 十 土 . Considering "土" as a component, we can get characters like 在 [zài] at, 地 [dì] earth, 去 [qù] go and so on, characters containg 土 as a component have a lot. When getting wise to 土 as a radical, we have 坏 [huài] bad, 块 [huài] block, 址 [zhǐ] address and so on.

With wenlindemo341, we know character "土" can be written stroke by stroke like as Figure 2. The overseas students who believe chracter writing is painful will have fun according to use Character Description Language.

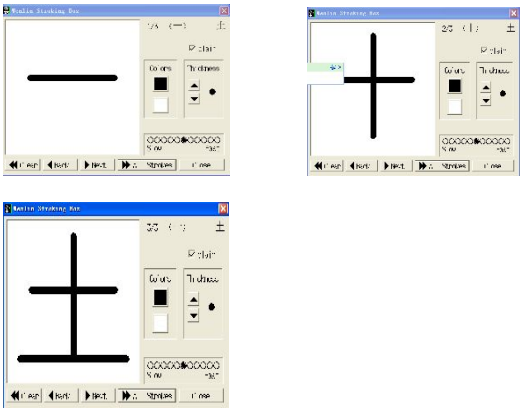

Figure 2. 1st, 2nd and 3rd stroke of 土

The wenlindemo341 has only a miniature dictionary, with a few dozen vocabulary items. The complete Wenlin dictionary has over 10,000 characters and about 200,000 compound words and phrases, including the entire ABC Chinese-English Dictionary edited by John DeFrancis.

V. CONCLUSIONS

CDL is based on Unicode, XML(Extensible Mark-up Language), and a few well-known characteristics of Han characters. It makes use of flexible strokes to display Chinese characters, maybe a helping aid to overseas students to learn Chinese characters writing.

There are many approaches to representing what might be seen as a Chinese character: bitmap, vector drawing, HanGlyph, SGML. The first two methods describe pure glyphs, whereas CDL sits somewhere between a character and a glyph, HanGlyph appears to describe almost pure haracters, but maintains some glyph-like operations that prescribe relative dimensions. SCML instead describes pure characters.

We discussed methods about character description, gived the key features and compared their differences. Decomposing a Chinese character into strokes and components is the trend of this research field, and further, application of XML will be added into these fields. All we can do is to make the character set cover all possible Chinese characters, some particular variants of characters that are "unified" can be described, traditional Chinese character and simplified chinese can both be recognized in the near future.

ACKNOWLEDGMENT

I would like to express my sincere appreciation to all the members of my team, for their encouragement and advice. Thanks to my family, they give me endless love. "The Research on Chinese Visualized Teaching Method's Model for Foreigner in Long Distance", The National Society Science Foundation of P.R. China Grant (No:2006BAK11B03) partly supported this paper."The Research on Cyberculture Security and its Warning System", Project in the National Science & Technology Pillar Program in the Eleventh Five-year Plan Period (No:07BYY033) partly supported this paper.

REFERENCES

[1] Li Xiangnong, Zhang Yi, "Design of Long-distance Visual TCFL Platform", JOURNAL OF YUNAN NORMAL UNIVERSITY(TEACHING AND RESEARCH ON CHINESE AS A FOREIGN LANGUAGE), Vol. 6, No.1, Jan., 2008.

[2] http://en.wikipedia.org/wiki/Chinese_character.

[3] Tom Bishop, Richard Cook, "A Specification for CDL Character Description Language.", 2003, pp. 2-5.

[4] 上海通大学汉字编码组. 汉字信息字典[M]. 北京：北京大学出版社，1988.

[5] 国家语言文字工作委员会.GF3001-1997 信息处理用 GB13000.1 字符集汉字部件规范[S]. 北京. 语文出版社，1997.12.1 发布，1998. 5. 1 实施。

[6] http://www.sinica.edu.tw/～cdp.

[7] SUN Xing-Ming, YIN Jian-Ping, CHEN Huo-Wang, WU Quan-Yuan, JING Xin-Hai, "ON MATHEMATICAL EXPRESSION OF A CHINESE CHARACTER", Journal of Computer Research and Development, 2002, pp.1-4.

[8] Tom Bishop, Richard Cook,. "Character Description Language (CDL): The Set of Basic CJK Unified Stroke Types",May 23, 2004, pp. 1-5.

[9] Lin Min, Song Rou., "Pattern Computing-Oriented Formal Description of Chinese Character Glyph",JOURAL of CHINESE INFORMATION PROCESSING,VOL.22, NO.3, MAY 2008, pp. 1-2.

[10] http://www.wenlin.com/cdl.