Chinese Character Synthesis using METAPOST

Candy L.K. Yiu

Department of Computer Science, Hong Kong Baptist University
email: candyyiu@comp.hkbu.edu.hk

Wai Wong

Department of Computer Science, Hong Kong Baptist University
email: wwong@comp.hkbu.edu.hk

Abstract

A serious problem in Chinese information exchange in this rapidly advancing Internet time is the numerousness of characters. Commonly used character encoding systems cannot include all characters, and often fonts do not contain all characters, either. In professional and scholarly documents, these unencoded characters are quite common. This situation hinders the development of information exchange because special care has to be taken to handle these characters, such as embedding the character as an image. This paper describes our attempt towards solving the problem. Our approach utilizes the intrinsic characteristic of Chinese characters, that is each character is formed by combining strokes and radicals. We defined a Chinese character description language named *HanGlyph*, to capture the topological relation of the strokes in a character. We are developing a Chinese Character Synthesis System CCSS, which transforms *HanGlyph* descriptions into graphical representations. A large part of the CCSS is implemented in METAPOST.

Introduction

The rapid advancement of the Internet and the Web provides an effective means of information exchange. However, there is a very serious problem in exchanging Chinese documents. This is because the number of Chinese characters that ever exist is unknown. Furthermore, new characters are being created continually. Therefore, no character set can encode *all* Chinese characters.

Even if a character set could encode all Chinese characters, it is very expensive to create Chinese fonts using typical methods and a fairly large number of Chinese characters would be so rarely used that expense would be very difficult to justify.

One possible solution to this problem is to create an unencoded character according to its composition of strokes and radicals. Several experiments along this line were attempted in past, but none were very successful. The key reason is that the composition of the strokes and radicals is very complex, and the previous attempts did not effectively divide and resolve the complexity. Section 1 gives a brief survey of some previous attempts.

Our approach to Chinese character synthesis resolve the complexity in two ways. First, we defined a high-level Chinese character description language, *HanGlyph*. It captures the abstract and topological relation of the strokes. Thus, the character description is compact and can be targeted to a variety of rendering styles. Section 1 describes the *HanGlyph* Chinese character description language in more detail. Secondly, we use METAPOST as our rendering engine to take its advantages of meta-ness and the ability of specifying paths and solving linear equations.

The *HanGlyph* language is defined based on many studies of Chinese characters. Section explains the basic structure of Chinese characters for the benefit of readers who are not familiar with them. *HanGlyph* defines 41 basic strokes, 5 operators and a set of relations. A character is built by combining strokes using the operators recursively. *HanGlyph* allows the user to define macros to represent a stroke cluster which can then be re-used in building more complex characters.

The CCSS(stands for Chinese Character Synthesis System) takes *HanGlyph* expressions and renders the characters. It can be divided into three parts: a front-end to translate *HanGlyph* expressions into METAPOST programs, a set of primitive strokes and a library of METAPOST macros to implement the operators, relations. By varying the parameters to these macros, or redefining the basic stroke macros, Chinese characters in different styles can be formed. Thus, it can create a variety of different fonts from the same *HanGlyph* description. The implementation of CCSS is described in Section 1.

The structure of Chinese Characters

Chinese characters, or *hanzi*, has its root in a very long history. A large body of literature on the study of the written form of Chinese language dated from as early as more than 2000 years ago in *Han* dynasty up to now are available. The written form of Indo-European languages consists of around 30 characters. Words are formed using these characters in a linear fashion. In contrast, written Chinese language is denoted by tens of thousands of *hanzi*. The exact number of *hanzi* that have ever in existence is never known.

Many studies have pointed out that each Chinese character is composed from strokes. The number of strokes in a character varies from one for the simplest, and up to around 50 for the most complex. Unlike the linear composition of words from characters in Indo-European languages, the arrangement of the strokes in *hanzi* is two-dimensional.

According to the convention of writing Chinese characters, a stroke is a continuous movement of the brush over the writing surface without being lifting up. It is commonly agreed that there are five most basic strokes: ㇐ (橫héng[1]), ㇑ (豎shù), ㇒ (撇piě), ㇏ (捺nà) and ㇔ (點diǎn).

In practice, each of these basic strokes has some variations depending on the position in a character. For example, the stroke ㇒ 撇can have two more variances: ㇒ (平撇pìngpiě) (as the top stroke in 千) and ㇓ 豎撇(as the left most stroke in 月). In addition, a number of combinations of these basic movements are considered as strokes because they are connected in a natural way in writing. For example, a ㇐ 橫followed by a ㇒ 撇is a single stroke ㇇ called (橫折撇hèngzhépiě). Modern studies on Chinese characters [1] [9] identified a small set of around 40 strokes as the basic elements of *hanzi*.

Although the arrangements of strokes to form a *hanzi* is very complex, there are some rules that guide the formation of characters. Further, some stroke arrangements are relatively stable and appear in many characters. Some of these arrangements are themselves *hanzi*, for example, 日月; some of them are known as *radicals* which are used in Chinese dictionery to index characters, for example, 彳匚卩. There are still some relative stable arrangements that are not *hanzi* themselves, neither radicals, but appear in many characters. We will use the term *components* to refer to all these kinds of stroke arrangements, while we use a more general term *stroke clusters* to refer to any arrangements of several strokes.

Except a small number of very simple characters, like 人,二,十, which cannot be divided into component parts, all *hanzi* can be considered as compositions of certain components. The ways of composing *hanzi* from components are known as the *structure* of the character. Many studies, such as [2] and [8], have identified around 10 different types of structures if one considers to compose character from only two componenets. This does not place serious restrictions because the composition process can be performed recursively. Figure 1 illustrates the commonly used structures.

Related works

Based on the studies of Chinese characters, several attempts have been carried out to create *hanzi* from a structural composition approach.

Toshiyuki, *et al* [10] proposed a way of describing Chinese character using sub-patterns. In principle, their method is similar to the approach of *HanGlyph* because the underlying theory of character structure is intrinsic to all Chinese characters.

Dong [11] and Fan [5] reported their work on the development of a Chinese character design system which took a parametric approach to create characters in different styles. Lim and Kim [7] developed a system for designing oriental character fonts by composing stroke elements.

[1] The word héng following the *hanzi* name of the stroke is in *pinyin*, a phonetic transcription of Chinese characters.

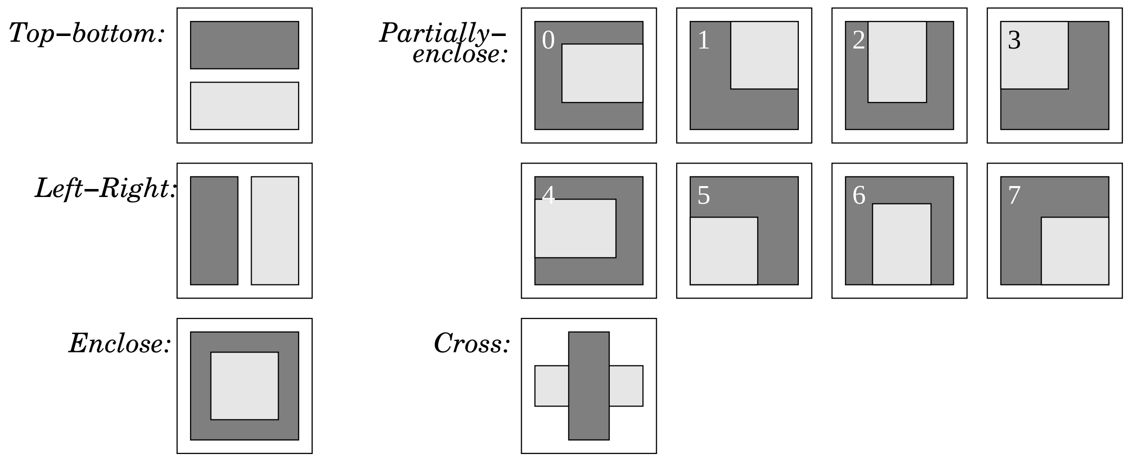

Figure 1: The basic structure of *hanzi*

Inspired by the success of METAFONT[6] in creating latin character fonts, Hobby and Gu[3] attempted to generate Chinese characters of different styles using METAFONT. A small set of strokes were defined in METAFONT. A small set of radicals were then defined as METAFONT macros by using the strokes. Characters can then be specified as METAFONT programs using these macros as building blocks. By varying some parameters governing the shapes of the strokes, fonts of different styles can be generated. However, the research was not conclusive because they only generated fonts of a very small character set (128 characters).

Another attempt similar to Hobby and Gu was done by Hosek [4] who aimed at generating *hanzi* from a small sets of components.

A common theme of the works mentioned is the difficulty of handling the complexity of the structures and the numerousness of characters. Our approach handles the complexity using an abstract description and a layered CCSS to decompose the complexity into several sub-problems. On the *HanGlyph* level, we consider strokes as abstract objects. We need to specify the relative positions between these abstract objects only. On a lower level, we can work out the outline of the strokes and fine tune the positions.

Another theme of the mentioned works is that they are mainly aimed at the disign and generation of character fonts. Our approach can certainly be applied in font generation. However, a very important application area, namely for the exchange of Chinese character information, is made possible with our character description language *HanGlyph*.

The *HanGlyph* Language

Based on the analysis described in previous sections, we defined a Chinese character description language, named *HanGlyph*. The most crucial characteristic of this language is that it is abstract and it captures only the topological relation of the strokes that form a character.

The essential information needed to distinguish a Chinese character is the arrangement of strokes. The precise location of each stroke can vary in a large extent upto certain threshold, and the character can still be recognized correctly. For example, the following two characters, 士and 土, comprise exactly the same strokes and in exactly the same arrangement. The only difference between them is the relative length of the two horizontal strokes. How much longer is a horizontal stroke in these characters is unimportant for distinguishing between them. To recognized the character 土, the threshold is that the upper horizontal stroke must be shorter than the lower one. Therefore, the *HanGlyph* language does not describe the precise geometric information of the characters.

The strokes After studying a number of Chinese linguistic and graphological works, we selected a small set of 41 strokes as the primitives of *HanGlyph*. Each primitive stroke is assigned a Latin letter as its code so that users can write *HanGlyph* expressions using a standard qwerty keyboard easily. Table 1 lists these primitive strokes.

Table 1: *HanGlyph* primitive strokes

Stroke	Name	Code	Examples	Stroke	Name	Code	Examples
㇔	點	`d`	衣主沙	㇒	撇	`p`	大人少
㇒	左點	`D`	心快熱	㇒	平撇	`P`	看千毛
㇔	長點	`f`	不	㇓	豎撇	`q`	用月兒
㇛	撇點	`g`	女好巡	㇜	撇折	`r`	么絲去
㇐	橫	`h`	二三	㇏	捺	`n`	人大丈
㇕	橫折	`i`	口四國	㇝	平捺	`v`	走趕這
㇆	橫折鉤	`j`	勾狗月	㇀	提	`t`	刁打地
㇇	橫折撇	`k`	又水冬	㇀	點提	`U`	冰清
㇍	橫折彎	`l`	沿船般	㇅	橫折折	`N`	凹
㇈	橫折彎鉤	`m`	乙吃	㇎	橫折折折	`L`	凸
㇖	橫鉤	`a`	冠皮軍	㇡	橫折折鉤	`J`	乃仍
㇑	豎	`s`	十中用	㇊	橫折提	`E`	语计
㇗	豎折	`b`	山區忙	㇌	橫撇彎鉤	`K`	隊部
㇄	豎彎	`c`	四酒	㇞	豎折折	`B`	鼎亞
㇟	豎彎鉤	`w`	見兒	㇉	豎折折鉤	`C`	馬弓
㇙	豎提	`e`	衣根	㇉	豎折撇	`Q`	专
㇚	豎鉤	`S`	丁寸利	⺄	橫斜鉤	`M`	飛風颱
㇁	彎鉤	`X`	狗狼家	㇋	橫折折撇	`R`	廷建及
㇂	斜鉤	`Y`	我氣代	㇒	撇點	`z`	学应
㇃	臥鉤	`W`	心感	㇇	橫折斜	`F`	子魚矛
㇠	橫豎彎鉤	`o`	九几				

The operators and relations To form a Chinese character, one combines primitive strokes using operators. Five operators are defined as listed in Table 2. Figure 1 illustrates the composition performed by these operators. Each operator takes two operands and combines them to form a stroke cluster. This operation continues recursively until the desire character is formed. For example, to describe the character 土, one may first combine two horizontal strokes using the top-bottom operator, then use the cross operator to add a vertical stroke. The *HanGlyph* expression for this character (written in ASCII characters) is `h h=s+`. (*Note*: the expression is in postfix notation.)

However, with only these operators, some characters, like 土and 士mentioned above, cannot be distinguished. To resolve situations like this, we can augment the operator with a number of *relation specifiers* to describe the operation in more specific terms. For our sample character 土, the proper *HanGlyph* expression should be `h h=< s+_` where the symbol `<` denotes the relation that the length (i.e., the horizontial

Table 2: *HanGlyph* operators

Name	Symbol	Example
top-bottom 上下	`=`	昌早李
left-right 左右	`|`	明他你
fully enclose 全包	`@`	回國困
half enclose 半包[†]	`^`	凶問區
cross 穿插	`+`	十半木

[†]A digit ranging from `0` to `7` should suffix the half enclose operator to indicate the direction of the opening.

dimension) of the upper horizontal strokes must be shorter than the lower one, and the symbol _ denotes the relation that the two operands of the cross opeartor, namely 二and 丨, are aligned at the bottom.

The following five kinds of relations are defined:

1. Dimension — The relations in this group specifies the relative dimension of the operands, i.e., comparing the width and height of their bounding boxes. There are four boolean relations: less than(`<`), greater than(`>`), not less than(`!<`), not greater than(`!>`).
2. Alignment — This specifies how the operands are aligned. The possible alignments are at top(`‘`), at bottom(`_`), at left(`[`), at right(`]`) and centred(`#`). More than one alignment can be added to an operation, for example, to aligh at bottom right (`_]`).
3. Touching — This specifies whether the operands can touch each other. The possible relations are touching (`~`) or not touching (`!~`). When combining two elements to form a new character, the strokes next to the interface of the two elements may or may not touch each other. In general, if the strokes on either sides of the interface have the same direction, they will not touch each other, for example, 昌晶. Otherwise, the strokes may touch each other, for example, 香相.
4. Shifting direction (`*`) — This is used to specify the direction along which the second operand is moved against the first operand. The direction is indicated by a number same as the half-enclosing operator. Another number follows the direction indicated when the second operand stops.
5. Scale (`/`)— This is used to adjust the width and height of the resulting character after the operation.

The *HanGlyph* macros It will be very tedious if every character is described down to all its primitive strokes. It can be seen that certain arrangements of strokes are very common, such as 日月, and so on. They are used to build up characters. We call them *components*. *HanGlyph* allows *macros* to be defined to stand for a component. For example, the component 日is a macro with the name `ri_4`. It is defined in terms of another macro `sih` representing the component 口. The actual definitions are as below:

```
let(sih){s i|/h=/}
let(ri_4){sih h@}
```

With a small number of operators and relations, and using a postfix notation, the syntax of the *HanGlyph* language is very simple. This facilitates the development of simple language processors. Appendix A shows the concrete syntax of *HanGlyph*.

After defining the *HanGlyph* language, we have written descriptions of more than 3755 Chinese characters (the first level characters in the GB2312-80 character set). We found that the *HanGlyph* language is adequate for its purpose, that is to capture the topological relation of the strokes.

The CCSS

The purpose of the Chinese Character Synthesis System (CCSS) is to render the *HanGlyph* expressions into a visual representation. For example, the *HanGlyph* expression `h h = <` only tells there are two horizontal strokes, one above another, and the upper stroke should be shorter than the lower one. However, it does not tell us about the exact distance between two strokes. In addition, it does not tell us exactly how much shorter is the upper stroke than the lower one.

The task of the CCSS is to determine and calculate the precise geometric information for each stroke so that a good rendering of characters can be generated.

The CCSS consists of three modules: basic strokes, composition operations and a *HanGlyph*-to-METAPOST translator. The first two modules are implemented as METAPOST macros. The translator is a C program.

Basic strokes Each of the 41 basic strokes listed in Table 1 is implemented as a set of three METAPOST macros:

- *Control-point* macro defines the control points, handle points and their properties. For example, the stroke ㇅ (横折折鉤héngzhézhégōu) has six control points and a handle point as shown in Figure 2(a). The locations of the points are specified relative to a referece point. The properties of a control point indicates whether it is a beginning, an end point, or a turning point. These poperties will be used in creating the outline since its shape at different types of points will be different, for example, the second control point is a turning point, the outline at this point will have a serif shape. The properties will also

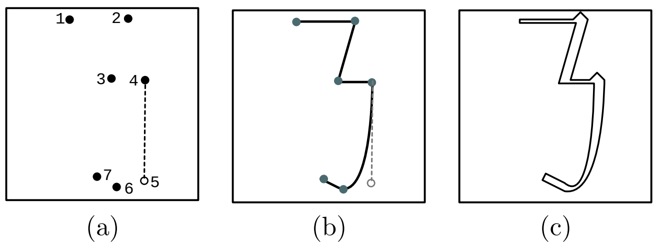

Figure 2: Basic stroke macros for the stroke ㇡

be used in the composition operations to determine whether certain transformation and positioning are required.

- *Skeleton* macro specifies a path passing through the control points. This path is very important. Given two points, the path can be straight or curvy, therefore, this macro traces out the exact stroke skeleton. Firgure 2(b) shows the skeletal path of the stroke ㇡.
- *Outline* macro creates the outline for the stroke. It is defined relative to the control points and the skeletal path. Firgure 2(c) shows the skeletal path of the stroke ㇡.

The first reason for organizing the strokes into three macros is to avoid distortion. In composition operations, each stroke will be transformed sereval times before the whole character is formed. If the stroke including the outline is represented in one macro, the transformation will distort the stroke thickness and even the direction in certain slant strokes.

Another reason is to provide meta-ness and flexibility. This organization provides several levels of style changes. The first level is to vary the parameters of the outline macros. For instance, changing the stroke thickness parameter will result in characters of varying stroke thickness. If we change the set of outline macros, we can create a completely different font styles, but they may still recognised as a family because the locations of control points are unchanged. More variations can be achieved if the control point macros and the skeleton macros are also changed. The result will be a completely different font family. Figure 1 shows three variations of the same skeleton. Figure (a) is the simple skeletal path of the strokes. Figure (b) is the outline with serif. Figure (c) is generated by stroking the skeleton with a pen angled at 25 degree.

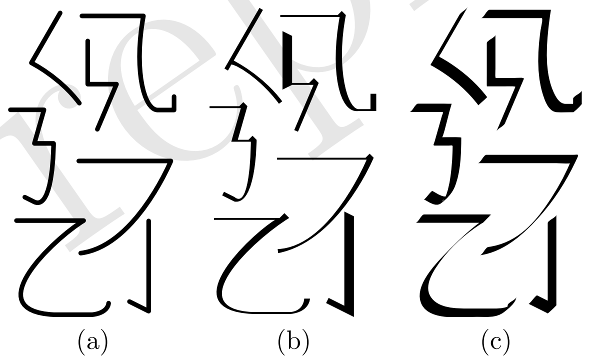

Figure 3: Variations of strokes having the sme skeleton.

Composition operations CCSS implements five operations corresponding to the five operators defined in *HanGlyph*. These operations are implemented as METAPOST macros. Again, the operators in *HanGlyph* respresents abstract operations, like the *Top-bottom* operator (=) only means 'put an operand on top of another'. It carries no precise geometric information. Given this abstract instruction, the macro implementing this operation has to calculate the exact location and dimension of each operand. The resulting rendering should be a well-balanced and well-positioned arrangements of strokes.

One important task the composition operation has to handle is to estimate the relative size and position for its operands so that the result is visually well-balanced. For example, Figure 4 illustrates two characters having the same radical 木(mù) on their left side. The width of this radical in the first character 林(lín) is larger than that in the second character 樹(shù) because the right-hand side of 樹has many more strokes. We have found that the ratio of the widths of the two components is proportional to the ratio of the sums of the lengths of the strokes and the number of strokes of the components.

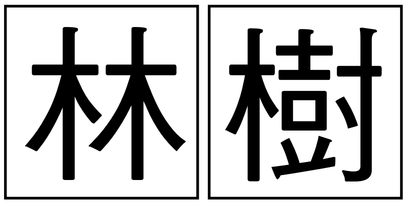

Figure 4: The same radical having different widths.

HanGlyph expressions may include a number of relations to augment the operators. The composition operations need to calculate the exact dimension and transformation to apply to each operand. For example, the character 人(rén) composed of two strokes where the right-hand one is shorter and the two are aligned at the bottom. The composition operation will first scale the right-hand stroke down to a default size, and then translate it so that the bottom lines of the two strokes are aligned to the bottom of the character box.

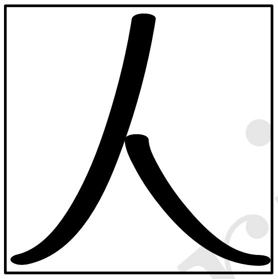

Figure 5: A character composed of two strokes aligned to bottom.

While we are talking about transforming the strokes, in fact, only the control points and the skeletal path are transformed. After all strokes forming a character have been put at the right position, the outline is drawn. This avoids the outline being distorted by the transformations.

***HanGlyph* to METAPOST traslation** The front-end of the CCSS is the translator that converts*HanGlyph* expressions into METAPOST programs. The current implementation of the translator puts each *HanGlyph* expression into a METAPOST figure. Within each figure, the appropriate sequence of composition operation macros are called to render the character. The output of the system is a set of PostScript files.

This implementation provides a simple way to render the *HanGlyph* expressions and obtaining previews of the characters. It facilitates the fine-tuning of the composition operations. Future implementation can streamline the process according to the requirements of the target application. For instance, a back-end processor can be added to convert the PostScript output into a particular format, such as a PostScript Type 3 font.

Conclusion

This paper describes an attempt to synthesize Chinese characters from an abstract description. A Chinese character description language known as *HanGlyph* has been defined. A Chinese Character synthesis system is being developed. It is implemented in METAPOST and C, and the output is rendered in PostScript. The preliminary results show that the approach is very promising. Some of the characters generated by the CCSS are shown in Figure 6.

Currently, we in the process of fine-tuning the composition parameters. We hope the system is able to produce visually pleasing characters. There are many factors that may affect the quality of the output, for example, the thickness of the strokes, the allocation of the space occupied by each component, and so on.

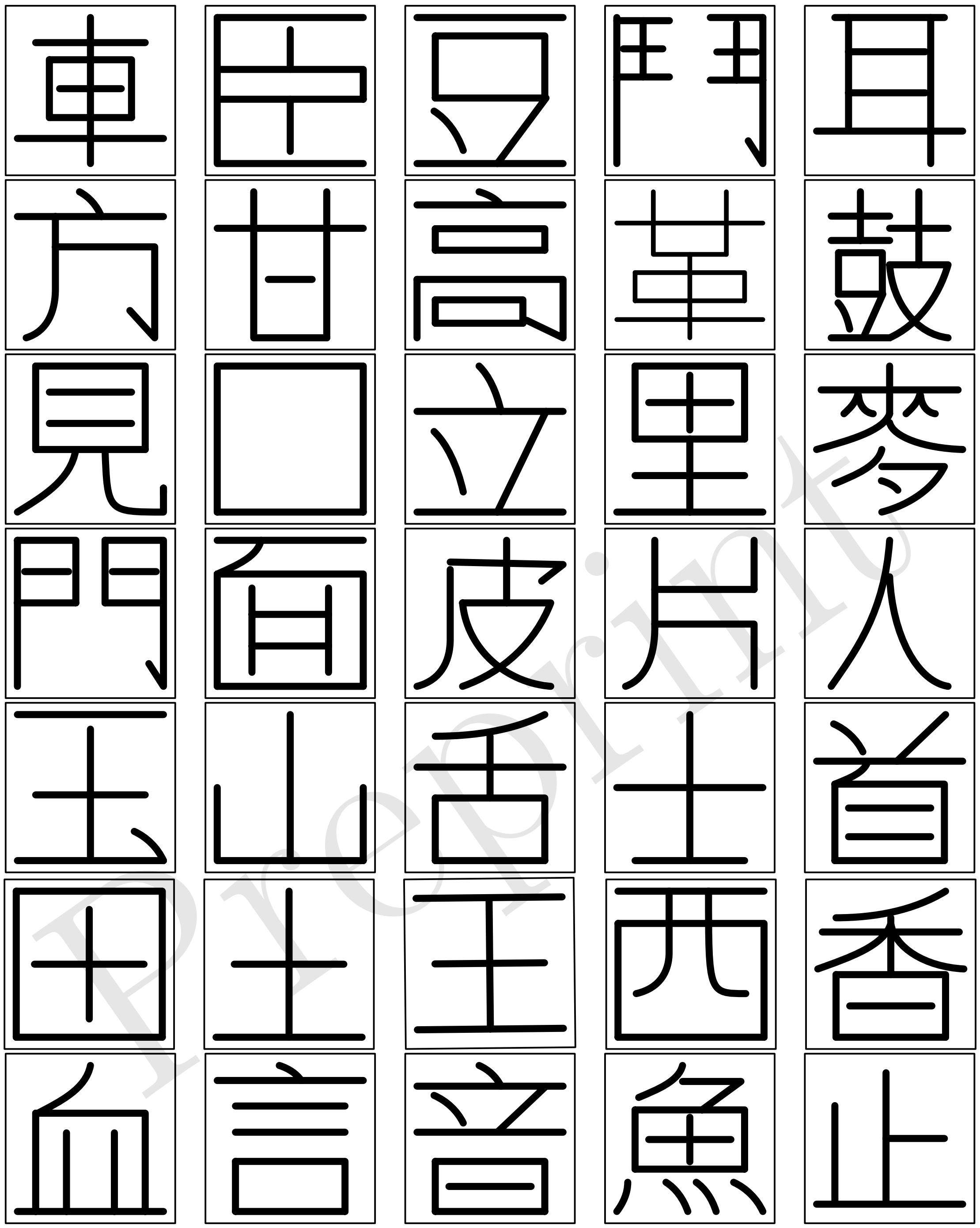

Figure 6: Some characters generated by CCSS

There are many applications of such system. The most important ones are in exchanging Chinese textual information in an open, heterogenous environment, and in Chinese font generation.

A The syntax of *HanGlyph*

This appendix shows the concrete syntax of *HanGlyph* in a augmented BNF notation.

```
       ⟨hanglyph⟩  ::=  ⟨expr⟩ +                               (1)
           ⟨expr⟩  ::=  ⟨glyph_expr⟩ | ⟨macro⟩ | ⟨char⟩        (2)
     ⟨glyph_expr⟩  ::=  ⟨glyph⟩;                               (3)
          ⟨macro⟩  ::=  let(⟨id⟩){⟨glyph⟩}                     (4)
           ⟨char⟩  ::=  char(⟨code⟩){⟨glyph⟩}                  (5)
          ⟨glyph⟩  ::=  ⟨glyph⟩⟨glyph⟩⟨operation⟩               (6)
                    |   ⟨stroke⟩
                    |   ⟨id⟩
      ⟨operation⟩  ::=  ⟨operator⟩⟨rels⟩?                      (7)
                    |   ⟨cross_operator⟩⟨cross_rels⟩?
       ⟨operator⟩  ::=  = | | | @ | ⟨half_op⟩                  (8)
 ⟨cross_operator⟩  ::=  +                                      (9)
        ⟨half_op⟩  ::=  ^⟨dir⟩                                 (10)

            ⟨dir⟩  ::=  0 | 1 | 2 | 3 | 4 | 5 | 6 | 7          (11)
           ⟨rels⟩  ::=  ⟨dimens⟩?⟨aligns⟩?⟨touch⟩?⟨scale⟩?    (12)
     ⟨cross_rels⟩  ::=  ⟨dimens⟩?⟨align⟩?                      (13)
                        (⟨align⟩ | (*⟨dir⟩⟨+int⟩))?⟨scale⟩?
         ⟨dimens⟩  ::=  ⟨comp⟩(⟨comp⟩ | ⟨num⟩)?                (14)
                    |   ⟨num⟩⟨comp⟩?
                    |   ⟨num⟩,⟨num⟩
           ⟨comp⟩  ::=  < | > | !< | !> | -                    (15)
         ⟨aligns⟩  ::=  ⟨align⟩⟨align⟩?                        (16)
          ⟨align⟩  ::=  ‘ | _ | [ | ] | #                      (17)
          ⟨touch⟩  ::=  ~ | !~                                 (18)
          ⟨scale⟩  ::=  /⟨num⟩?                                (19)
```

References

[1] 蘇培成. 二十世紀的現代漢字研究. 書海出版社, 2001.

[2] 劉連元. 漢字拓扑結構分析. In 漢字. 上海教育出版社, 1993.

[3] John Hobby and Gu Guoan. Chinese meta-font? *TUGBoat*, 5(2):119–136, 1984.

[4] Don Hosek. Design of oriental characters with MF. *TUGboat*, 10(4):499–501, 1989.

[5] Fan Jiangping. Towards intelligent chinese character design. In *Raster Imaging and Digital Typography II (RIDT91)*, pages 166–176. Cambridge University Press, 1992.

[6] Donald E. Knuth. *The METAFONT Book*. Addison-Wesley, 1986.

[7] Soon-Bum Lim and Myung-Soo Kim. Oriental character font design by a structured composition of stroke elements. *Computer-aided design*, 27(3):193–207, 1995.

[8] 傅永和. 漢字結構和構造成分的基楚研究. 上海教育出版社, 1993.

[9] 傅永和. 中文信息處理. 廣東教育出版社, 1999.

[10] Sakai Toshiyuki, Nagao Makoto, and Terai Hidekazu. A description of chinese characters using subpatterns. *Information Processing Society of Japan Magazine*, 10:10–14, 1970.

[11] Dong YunMei and LI Kaide. A parametric graphics approach to chines font design. In *Raster Imaging and digital typography II(RIDT91)*, pages 156–165. Cambridge University Press, 1992.