Building a Graphetic Dictionary for Japanese Kanji— Character Look Up Based on Brush Strokes or Stroke Groups, and the Display of Kanji as Path Data

Ulrich Apel
National Institute of Informatics
Hitotsubashi 2-1-2 Chiyoda-ku
Tokyo 101-8430
Japan
ulrich_apel@t-online.de

Julien Quint
National Institute of Informatics
Hitotsubashi 2-1-2 Chiyoda-ku
Tokyo 101-8430
Japan
quint@nii.ac.jp

Abstract

Reading and writing Japanese isn't easy for Japanese and foreigners alike. While Japanese learn these skills at school, foreigners should be helped by good teaching material and dictionaries.

Kanji lexica have to be very different from other dictionaries. Unfortunately existing lexica normally expect that the users already have a lot of information on a character to look it up—the character's stroke count, its radical or its pronunciation. Beginners normally don't have such information.

This project creates data to allow for easier and more flexible look up of Japanese characters and to build better teaching material. It develops different approaches to make use of this data.

1 Introduction: Kanji as Main Obstacle for Learning Japanese

The modern Japanese writing system is considered as one of the most complex ones in the world. In Japanese it is described as *kanji kanamajiri bun* (漢字仮名交じり文), meaning that Japanese uses the originally Chinese characters—kanji—together with kana, i.e. hiragana and katakana, which are cursive or shortened versions of former kanji and which represent only the sounds of syllables.

For foreigners kanji are very difficult to learn, to memorize, to read and to write. To make things worse, most teaching material on kanji contains too little information on how kanji are actually written. This concerns for example correct stroke order, stroke directions, possible glyph variations or kanji components.

2 History of Kanji and Variations in Kanji Forms

Chinese characters were developed in the second millennium before Christ. Kanji were introduced in Japan in the fourth century AD.

There are kanji dictionaries with tens of thousands of kanji, but in Japan there were never more than about 6,000 in actual use.

Developments after WW II lead to a separation of kanji forms in Japan, China, Taiwan and Korea. Japan and even more mainland China introduced shortened forms of kanji, while e.g. Taiwan uses the traditional kanji.

The Japanese Ministry for Education selected a number of around 2000 kanji for official and general public use. It also published instructions on the stroke order and the kanji forms, that should be taught at school for about 900 kanji (Monbushô 1958). Calligraphers stress, that these proposals often don't show the mainly used character forms or stroke orders (Emori 2003: 8–16).

3 Problem: Learning to Read and to Write Kanji

Although nowadays kanji are very often written with a word processor, it is still very important to be able to write them by hand too.

For learners of Japanese writing kanji by hand is still one of the best ways to memorize them. To recognize Japanese handwriting one has to identify the original strokes which are often joined, seeming hardly to be individual strokes. On computer screens strokes are also often almost unrecognizable.

Most lexica don't give information concerning the stroke order, and if they do, they only do so for a small number of characters. Normal kanji lexica contain little material on stroke forms and their variations in kanji components. Ordinary paper lexica are of little help if one is able to recognize only parts of a kanji. One has to recognize the whole kanji in order to be able to look it up.

4 Kanji Strokes, Stroke Groups and Path Data

In this project, kanji are considered as graphic information, hence, they are analysed in different ways. This new data is combined with existing data.

A very abstract and basic way to analyse kanji consists in a graphetic approach which leads to the recognition of graphemes. Graphemes are the smallest meaning distinguishing units. In the case of kanji this can be stroke length (as in 末 and 未), angle of the stroke (as in 三, 川 and 彡), stroke direction (as in 干 and 千), or ending of a stroke (as in 干 and 于).

A more concrete analysis which also takes the act of writing into account would use strokes as basic units of kanji. A stroke is a graphical element that can be drawn e.g. with a brush or a pencil without interruptions. Most kanji consist of more than one stroke.

Our analysis of strokes uses 25 basic forms of brush strokes for kanji. It considers stroke direction, bending of the strokes, stroke endings (blunt or with a short bend) and so on. The stroke forms are numbered and every stroke of a kanji is assigned with the corresponding number of its stroke form.

Strokes can be grouped together not only to build full kanji but also to combine smaller units which frequently occur in kanji. We call these smaller units grapheme elements. Many kanji dictionaries use a subset of such grapheme elements to classify characters (部首 *bushu*, engl. “radicals”). For the time being our analysis of the grapheme elements uses mostly existing kanji or given radicals.

The data concerning stroke forms, grapheme elements and relative position can, of course, be used for kanji look up.

To display the collected information concerning a kanji and its components, graphical data is needed. This is achieved in our case using a vector graphics software (Adobe Illustrator). Here kanji strokes are represented by paths. The stroke order is identical with the order of the path input.

To allow for later review and to have more flexible data, numbers for the stroke order are put beside the strokes.

4.1 Possible Applications of the Data

The data presented here allows new ways to look up kanji:

- Search for kanji by the form of their different strokes;
- Input of stroke forms on the numbers block of the keyboard in a matrix like style;
- Search for strokes in the correct stroke order;
- Search for grapheme elements:
- Search for stroke forms, radicals and grapheme elements according to there position.

The data allows new ways to display kanji:

- Kanji can be build up according to their stroke order. This could be used in dic-

tionaries or new teaching material (see figure 1);
- Practising sheets to write kanji can be generated automatically;
- Animation can be achieved automatically;
- Grapheme elements and stroke groups can be highlighted (see figure 2) etc.

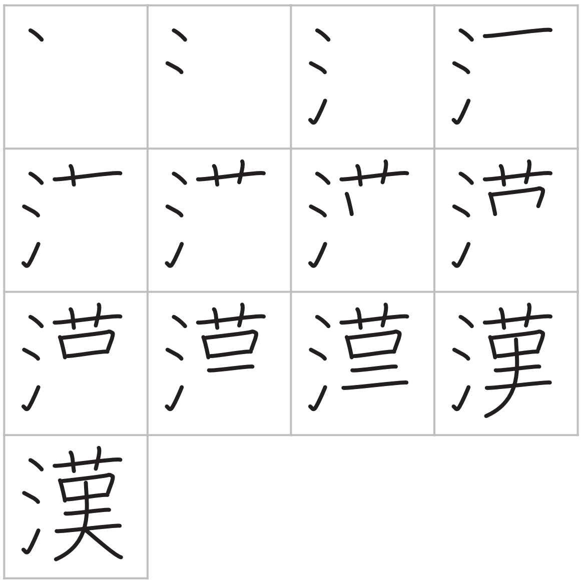

Figure 1: Building up a kanji by its strokes according to their stroke order

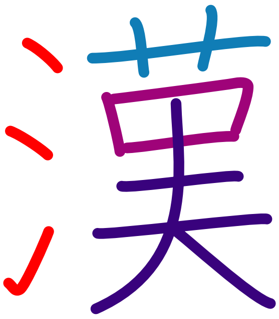

Figure 2: Highlighting grapheme elements with colours

4.2 An Example: Automatic Animation of Kanji Strokes

The path data created with Illustrator was exported into Scalable Vector Graphics format. SVG is an application of XML proposed by the World Wide Web Consortium. It provides a clear description of the graphical data well suited for the task at hand.

The graphical description of a kanji consists mostly of an ordered list of strokes. In SVG, we represent a stroke by a path element. For instance, the first stroke of the kanji 漢 (*kan*, “Han-China”) is:

```
<path d="M21.38, 19.75, c3.31,
   1.47, 8.54, 6.05, 9.37, 8.34"/>
```

The d attribute of the path element contains the path data in a compact form. This data is a list of drawing commands that an SVG renderer will execute to draw the path. The path data for every stroke will consist of a sequence of Bézier curves, which are parametric curves defined by four control points.

Several paths can be grouped together under a group element, which allows the association of groups of paths (i.e., lists of strokes) with every grapheme element of a kanji. It is then possible to deal directly with grapheme elements in the graphic representation of the kanji, in order to highlight such elements (as in figure 2) or to link them to other SVG files—e.g. clicking on the left component of 漢 would link it to the kanji 水 (*mizu*, “water”), which is this component’s standard form.

The SVG data available so far is static. Our goal is to present it in a dynamic fashion, showing strokes one by one, in the order and the direction in which they should be drawn. We will add an animated child element to every path in the static SVG file to create its animated counterpart. The animate element controls the moment at which the path is drawn, and the shape it should take.

Unfortunately there is no special command in SVG to draw a path progressively. A solution is to divide every path in several smaller ones, and to draw each segment one after another, giving the impression of an invisible pen drawing the kanji. Our division strategy is to segment every curve in a path into a fixed number of elements. That number of element is set to a power of two, because dividing Bézier curves into two is very easy to do. Longer strokes will consist of more curves than shorter ones, and it will take more to time to draw them; the distribution of the control points along the curves makes the animation look quite natural.

At the end, an animation is controlled by two parameters: the number of segments into which a curve is split and the time between the drawing of two strokes. Modifying these values will make the drawing slower or faster, and more or less smooth.

The first stroke of our example kanji will now look like shown below. The animation will start at time 0; it lasts for 0.45 seconds and it will iterate over the values given by the value's attribute. The d attribute in the path parent element will take these successive values over time.

```
<path d="">
<animate attributeName="d"
  begin="0" dur="0.45s"
  values="M21.38 19.75 C21.79
  19.93 22.23 20.16 22.69 20.43;
  M21.38 19.75 C21.79 19.93 22.23
  20.16 22.69 20.43 C23.16 20.7
  23.63 21.01 24.12 21.35;…"/>
</path>
```

4.3 Future Work

Based on the existing data, it is easy to develop further data concerning variations in stroke order or kanji form.

Especially the stroke descriptions could be used for better graphical character recognition. It may even lead to software that is able to recognize incorrect input, and is capable of explaining the user how to correct it.

So far we deal only with Japanese kanji, but, of course, the same approach could be used for other characters like hiragana and katakana or the traditional and the shortened Chinese characters.

5 Conclusion

Neither the analysis of kanji into strokes or grapheme elements, nor animation of kanji stroke order or the highlighting of grapheme elements are totally new.

The problem until now has been that for each of these tasks one had to build new data practically from scratch. This is the reason why existing grapheme analysis, movie files with animated strokes, or kanji graphics with numbered strokes deal only with a few hundred kanji (see some examples in the references). In contrast our data cover already several thousands of characters and glyph variations. Corresponding animations etc. can be generated automatically.

In sum, our data seem to be very adaptable. They have the potential of being used for a wide range of applications—many of which we haven't even thought of ourselves.

References

Hadumod Bußmann. 1990. *Lexikon der Sprachwissenschaft*. Kröner, Stuttgart.

Emori Kenji. 2003. *Kai gyô sô—hitsujun jitai jiten*. Sanseido Tokyo.

Eduardo Fazzioli. 1987. *Gemalte Wörter. 214 chinesische Schriftzeichen – Vom Bild zum Begriff*. Lübbe, Bergisch Gladbach.

John Ferraiolo, Fujisawa Jun and Dean Jackson, editors. 2003. *Scalable Vector Graphics (SVG) 1.1 Specification*. http://www.w3.org/TR/SVG11/.

Wolfgang Hadamitzky. 1995. *Langenscheidts Handbuch der japanischen Schrift – Kanji und Kana 1. Handbuch*. Langenscheidt, Berlin *et al*.

Kôdansha Encyclopedia of Japan. 1998. Kôdansha, Tokyo.

Monbushô. 1958. *Hitsujun shidô no tebiki*. Hakubundô, Tokyo.

Berthold Schmidt. 1995. *Einführung in die Schrift und Aussprache des Japanischen*. Buske, Hamburg.