Development of a Kanji Dictionary Layout Tool

Paul R. Dixon

1 Introduction

This presentation describes the development of a tool for the rapid creation of high-quality printed Japanese kanji dictionaries. The tool allows a precise control of the kanji entries an their example compounds, making it specifically suitable for creating dictionaries whose contents are matched to specific closed lexicon for example language proficiency examinations or course teaching material.

The basic operation of the dictionary creation tool is to take a list of kanji characters and vocabulary list and output a kanji dictionary in PDF format. The layout and the insertion of all the additional information is handled automatically by the software. To perform the automatic data lookup the tool internally has several dictionaries. A stroke order dictionary which currently covers most of the *Jouyou* kanji created from scratch. The databases containing additional kanji data such as radical or grade are derived from KANJIDICT[1]. The Automatic word translations are currently taken EDICT[2].

The initial version of the system used a native PDF library to directly create the final documents. However the amount of code required to deal with the manual layout became unmangable. Therefore a switch was made instead to generate LaTeX source files and let the LaTeX tool deal with the tedious manual placement. The disadvantage of this approach however is the dependency upon a LaTeX distribution. Two large components are the development of the kanji stroke dictionary and a grapheme to phoneme aligner for manipulating the compounds. These two modules have uses in other computational Japanese applications.

2 Dictionary Organisation

The system takes the input kanji and vocabulary data and generates the dictionary in the following order.

- **Front matter** is just composed of a table of contents and a title page. Any additional detailed front matter such as a user manual or a copyright notice can simply be included via a separate text file.
- **Main Matter** The main processing loop works through the kanji list sequentially assigning each entry a unique identifier and then writing the detailed entry information (described in Section 3). During this phase two other tasks are also carried out. For the purpose of creating indexes lists are maintained that can map each character to an identifier and a set of readings, radicals and stroke count. Page markers are constantly updated that indicate the reading and identifier range contained on each page.
- **Indexes** In the final stage reading radical and stroke indexes are written.

After the source files are written the dictionary is compiled using LaTeX. For a dictionary containing 1000 kanji and approximately 5000 vocabulary entries the entire process takes roughly 1.5 minutes. For enhanced on screen viewing hyper-links are used within the PDF document to allow for direct jumping from the indexes and special attention is paid to the PDF TOC to allow faster navigation. It is also possible to embed a Flash animation as opposed to the stroke order diagrams. However, this change will mean the stroke order information will be lost if the dictionary is also printed.

3 Entry Organisation

Figure 1 shows the layout of each kanji character entry.

1. **Kanji Character**
2. **Stroke Order Diagram:** This shows the correct stroke order, a separate frame is used for each stroke.
3. **Entry Identifier:** Each entry has a unique numerical identifier which is used for cross-referencing.
4. **On Pronunciation:** The sino-Japanese reading (*on-yomi*) written in katakana.

[1] http://www.csse.monash.edu.au/~jwb/kanjidic.html
[2] http://www.csse.monash.edu.au/~jwb/j_edict.html

5. **Kun Pronunciation:** The native Japanese reading (*kun-yomi*) written in hiragana.
6. **Meaning:** The translated meaning.
7. **Information Fields:** The following information is given for each character entry: radical, grade, usage frequency and stroke count.
8. **Usage Examples:** The set of words from the vocabulary list which contain the entry kanji.
9. **Furigana Annotation:** This is added to kanji characters which occur in compounds but do not have their own dedicated entry within the dictionary.

The two most complicated elements of the system are the stroke order diagrams and the sorting of the examples which will be described in more detail.

3.1 Usage Examples

For each character the set of usage examples is generated by finding every word in the vocabulary list in which the character occurs. For each example a grapheme-to-phoneme alignment is performed, the aligner also outputs the reading type and if a morpho-phonemic change has occurred a base reading for the purpose of comparison. The examples are placed in classes depending of the reading type the character generates. These are sino-Japanese readings (on-yomi), native Japanese readings (kun-yomi) or irregular readings such as *gikun*. Alignment failures are also placed into the irregular readings class. Each reading class is further divided into subsets based on the reading. These subsets are then sorted according to the characters position and to match the ordering of the readings in the pronunciation fields. When a word contain a kanji character that is not an entry in the dictionary a furigana annotation is shown above that character.

3.2 Kanji Stroke Order Diagrams

The kanji stroke order diagrams are derived from custom built stroke order dictionary. Each character is represented as a set of strokes. Every stroke is path composed of sequence of polylines and or cubic bezier curves. The vector information is written as postscript commands for typesetting. The vector nature of the stroke data allows for efficient storage and smooth scaling of the characters. The data is also suitable for use in dynamic applications such as animated stroke diagrams.

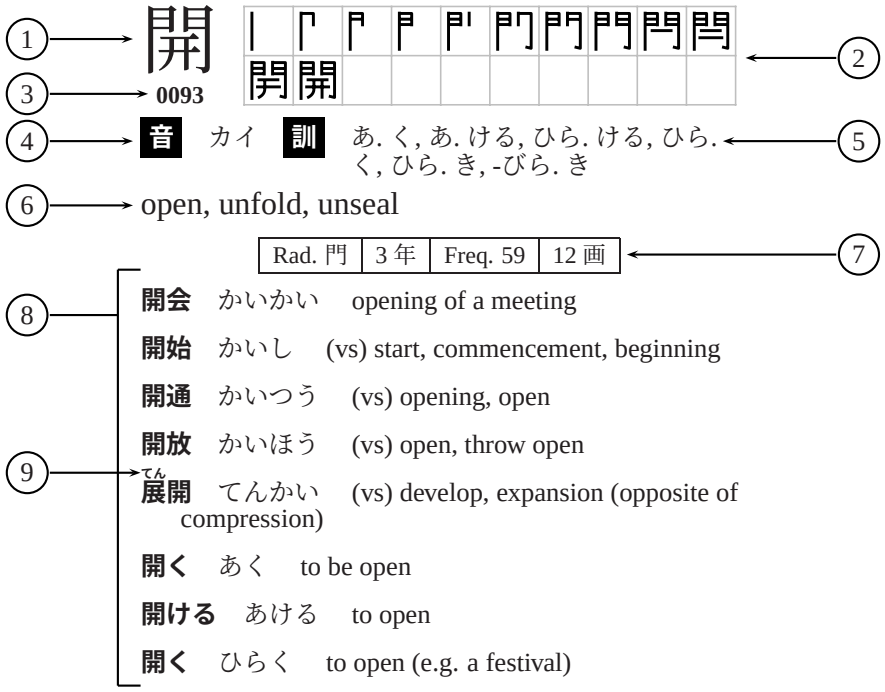

Figure 1: Example dictionary entry.