Multiple Indexing in an Electronic Kanji Dictionary

James BREEN
Monash University
Clayton 3800, Australia
jwb@csse.monash.edu.au

Abstract

Kanji dictionaries, which need to present a large number of complex characters in an order that makes them accessible by users, traditionally use several indexing techniques that are particularly suited to the printed medium. Electronic dictionary technology provides the opportunity of both introducing new indexing techniques that are not feasible with printed dictionaries, and also allowing a wide range of index methods with each dictionary. It also allows dictionaries to be interfaced at the character level with documents and applications, thus removing much of the requirement for complex index methods. This paper surveys the traditional indexing methods, introduces some of the new indexing techniques that have become available with electronic kanji dictionaries, and reports on an analysis of the index usage patterns in a major WWW-based electronic kanji dictionary. This is believed to be the first such analysis conducted and reported.

1 Introduction

Unlike languages written in alphabetic, syllabic or similar scripts, languages such as Japanese and Chinese, which are written using a large number of characters: *hanzi* in Chinese, *kanji* in Japanese, require two distinct sets of dictionaries. These are:

a. the traditional "word" dictionaries, as used in most recorded languages. Such dictionaries are usually ordered in some recognized phonetic sequence, and typically include the pronunciation or reading of the word as well as the usual dictionary components: part-of-speech, explanation. etc.

b. character dictionaries, which typically have an entry for each character, and contain such information as the classification of the character according to shape, usage, components, etc., the pronunciation or reading of the character, variants of the character, the meaning or semantic application of the character, and often a selection of words demonstrating the use of the character in the language's orthography. These dictionaries are usually ordered on some visual characteristic of the characters.

A typical learner of Japanese needs to have both forms of dictionary, and the process of "looking up" an unknown word often involves initially using the character dictionary to determine the pronunciation of one or more of the characters, then using that pronunciation as an index to a word dictionary, in a process that can be time-consuming and error-prone.

The advent of electronic dictionaries has had a considerable impact on Japanese dictionary usage:

a. it has facilitated the integration or association of character and word dictionaries such that a user can index between them in a relatively straightforward manner. This integration was pioneered by the author in the early 1990s (Breen, 1995), and is now a common feature of almost all hand-held electronic Japanese dictionaries and PC-based dictionary packages;

b. it has allowed the direct transfer of words between text documents and dictionary software, thus removing the often-laborious character identification;

c. for kanji dictionaries, it has greatly increased the number of character indexing methods that can effectively be used, and has also provided the opportunity for new indexing methods that are not available to traditional paper dictionaries.

This paper will concentrate on the issues associated with Japanese kanji dictionaries. Many of these also apply to Chinese.

2 Indexing a Kanji Dictionary

The general problem confronting the publication of kanji dictionaries is the large number of kanji in use and the absence of an intrinsic and recognized lexical order for kanji. In the post-war educational reforms in Japan, the number of kanji taught in schools was restricted to a basic 1,850, which has now been increased to 1,945. This set of kanji, along with a small set designated for use in personal names, accounts for all but a small proportion of kanji usage in modern Japanese. Many dictionaries and similar reference books compiled for students are based on this set (Sakade, 1961; Henshall, 1988; Halpern, 1999; etc.). The main computer character-set standard used in Japan, JIS X 0208 (JIS, 1997), which extends to less-common kanji including those used in places-names, has 6,355 kanji. This set is the basis for several kanji dictionaries (Nelson, 1997; Spahn & Hadamitzky, 1996), while larger sets of kanji are covered in many dictionaries, e.g. the Kodansha *Daijiten* (Ueda, 1963) has 14,900 kanji and the 13-volume Morohashi *Daikanwajiten* (Morohashi, 1989) has over 45,000 kanji.

In this paper, the term "primary index" has been used for the method of ordering the kanji entries, and "secondary index" has been used for cross-reference lists of kanji based on alternative ordering systems.

The major traditional indexing technique for kanji and hanzi dictionaries has been the radical system (*bushu* in Japanese), based on 214 elements plus about 150 variants. These elements are graphic components of the character that occur frequently enough to be used for indexing purposes. For example, the kanji 村 (*mura*: village) is identified by the 木 radical, and in a dictionary would be grouped with other kanji identified by that radical (札, 杒, 李, 松, etc.), with the grouped kanji ordered by the number of strokes in the remainder of the kanji. Radical systems have been used in Chinese character dictionaries for nearly 2,000 years, and the dominant 214-radical system was first used in the 康熙字典 *(kangxi zidian)* published in 1716.

Virtually all major kanji dictionaries published in Japan use the radical indexing method as the primary index, as do a number of dictionaries published elsewhere. Some dictionaries use modified or reduced sets of radicals. The technique is not simple to use, and some skill and practice is required in correctly identifying the radical and counting the residual strokes. The difficulty has been compounded by recent simplifications of the glyphs of the kanji, which in some cases have modified or eliminated the radical.

There are a number of other techniques used for indexing kanji in a dictionary:

a. **reading.** The reading or pronunciation of a kanji is a common and useful method of identification, and virtually all kanji dictionaries have a separate reading/kanji index. The reading cannot be used effectively as the primary index, as in Japanese each kanji usually has two sets of readings, and some kanji have as many as fifteen distinct readings.

b. **shape/stroke.** A number of techniques have been used to decompose the shape of a kanji according to coded patterns. One, which was popular in China, is the Four-Corner code, which allocates a number (0-9) to the pattern of strokes at each corner of the character, leading to a four-digit index. Another method, which is quite popular, is the SKIP (System of Kanji Indexing by Patterns) used by Jack Halpern in his kanji dictionaries (Halpern, 1990, 1999). In this, a kanji is typically divided into two portions, and a code constructed from the division type and the stroke-counts in the portions. Thus 村 has a SKIP code of

1-4-3, indicating a vertical division into four and three-stroke portions.

c. **school grade.** In Japan the kanji to be taught in each grade of elementary school are prescribed, and some references either organize kanji in those groupings or provide a secondary index of grades.

d. **stroke count.** The number of pen or brush strokes making up a kanji, ranging from one to over forty, can be an effective indexing technique, particularly for the simpler kanji. Some dictionaries employ a secondary index using the total number of strokes in a kanji.

e. **frequency.** The ranking of kanji according to frequency-of-use can be a useful secondary index, especially for the commonly used kanji.

f. **code-point.** The standardization of character set code-points for kanji has led to the emergence of dictionaries with these as the primary index. The Sanseido Unicode Kanji Information Dictionary (Tanaka, 2000) uses the Unicode code-point as the primary index, and the first edition of the JIS Kanji Dictionary (Shibano, 1997) used the JIS X 0208 code-point. It is interesting to note that the second edition (Shibano, 2002) changed to the traditional radical system, with the codepoints being relegated to a secondary index.

A summary of the indices available in a selection of dictionaries and references is in Table 1. The "**P**" indicates the primary index and an "**S**" indicates a secondary index. (The original Nelson uses a slightly modified version of the traditional radical index, and the Spahn & Hadamitzky Kanji Dictionary uses a simplified 79-radical system.

	Index Type						
Dictionary	**Radical**	**Shape**	**Code-point**	**Grade**	**Reading**	**Frequency**	**Stroke Order**
Morohashi (1989)	P				S		
Ueda (1963)	P				S		
Nelson (1974)	P*			S	S		S
Nelson (1997)	P			S	S		
S&H (1996)	P*				S		
S&H (1997)	S				S	P	S
Halpern (1990)	S	P		S	S	S	
Halpern (1999)	S	P			S	S	
Sakade (1961)				P	S		
Henshall (1988)				P	S		S
Shibano (1997)			P		S		
Shibano (2002)	P		S		S		
Tanaka (2000)	S		P				

Table 1: Index Types in Printed Kanji Dictionaries.

3 Electronic Kanji Dictionaries

As mentioned above, electronic kanji dictionaries have an increased number of indexing methods available, and in particular have navigational advantages over traditional paper dictionaries:

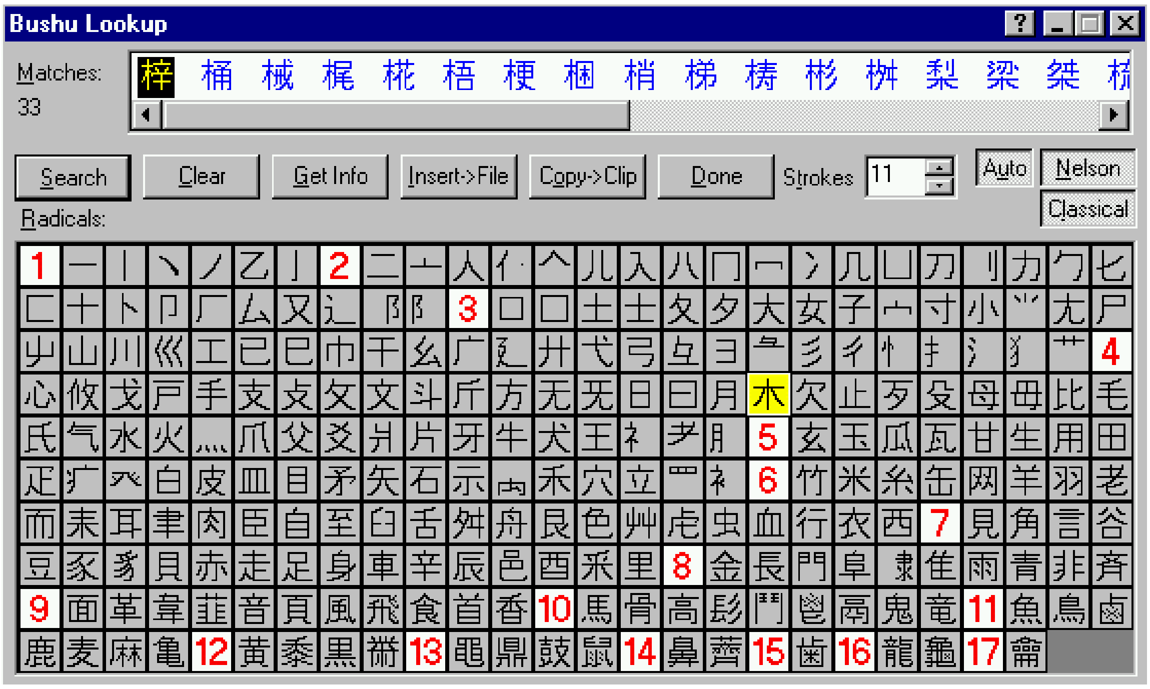
a. the concept of a "secondary" index no longer applies, as every index is capable of linking directly to the kanji entries;

b. dictionary users can choose flexibly between index methods according to preference, and can select a method appropriate to the characteristics of an individual kanji;

c. the above-mentioned capability to index directly to a kanji entry from a kanji selected from a text or application;

d. suitable GUIs can enhance the kanji lookup process by providing visual cues and a degree of interactivity.

Figures 1 and 2 show the GUIs for the *bushu* and SKIP methods in the kanji dictionary module of the JWPce word-processor (Rosenthal, 2002).

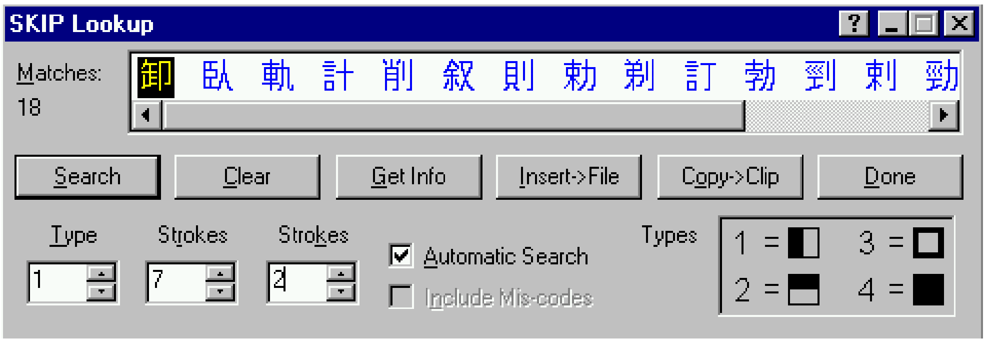


Matches: 33

梓 桶 械 梶 椛 梧 梗 梱 梢 梯 梼 彬 桝 梨 梁 桀 椛

Strokes 11

Fig. 1: Bushu Lookup GUI



Matches: 18

卸 臥 軌 計 削 叙 則 勅 剃 訂 勃 剄 剌 勁

Type 1 Strokes 7 Strokes 2

☑ Automatic Search

☐ Include Mis-codes

Types

1 = ◧ 3 = □

2 = ⬒ 4 = ■

Fig. 2: SKIP Lookup GUI

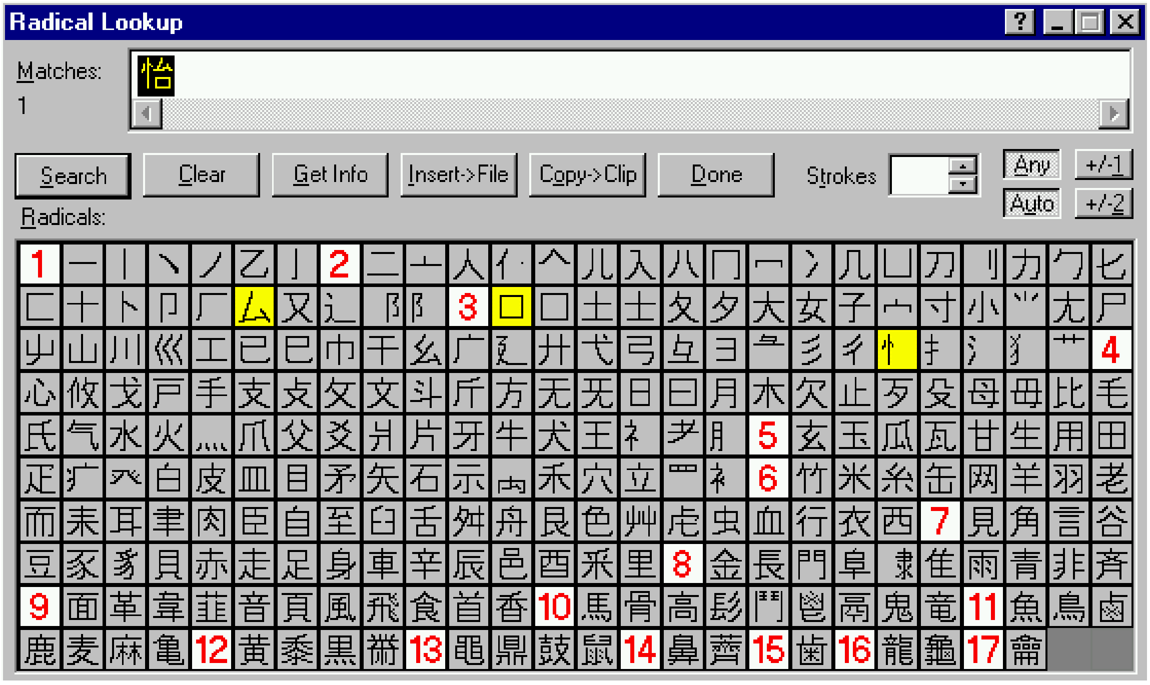
Among the new indexing methods introduced with electronic kanji dictionaries are:

a. indexing using the meaning of a kanji. The compilation of the KANJIDIC database (Breen 2004), which contains the English meanings for over 12,000 kanji, has enabled this technique to be employed. Searching for kanji meaning "castle" immediately gives 城, "fox" gives 狐, etc.

b. multi-radical searching. Most kanji are made up of several basic shapes drawn from a set of about 300 patterns. For example, the kanji 新 (*atarashii:* new) consists of the 立, 木 and 斤 patterns. Clearly, a traditional dictionary can only use one of these as an index (the index radical of 新 is 斤) but an electronic dictionary can use all the patterns to identify a kanji. A file of the visual components of the 6,355 kanji in the JIS X 0208 standard was prepared by a team of volunteers and is currently maintained by the author. The patterns used are similar to the traditional 214 radicals, but include common shapes such as ユ and マ that are not among the 214, and distinguishes between shapes such as 月 and 肉 that are regarded as variants of the same radical. The following is an extract from the file of kanji with the radical components identified.

旭 : 日 九
宛 : 夕 卩 宀
謂 : 月 言 田
韻 : 音 貝 口 日 立

This file is inverted, enabling dictionary software to identify the kanji containing a particular selection of radicals. Figure 3 shows the multi-radical lookup GUI in JWPce, having identified the 怡 kanji from its components.



1	一	丨	丶	丿	乙	亅	2	二	亠	人	亻	𠆢	儿	入	八	冂	冖	冫	几	凵	刀	刂	力	勹	匕
匚	十	卜	卩	厂	厶	又	⻌	阝	阝	3	口	囗	土	士	夂	夕	大	女	子	宀	寸	小	⺌	尢	尸
屮	山	川	巛	工	已	巳	巾	干	幺	广	廴	廾	弋	弓	彑	ヨ	⺕	彡	彳	忄	扌	氵	犭	艹	4
心	攸	戈	戸	手	支	攴	攵	文	斗	斤	方	无	旡	日	曰	月	木	欠	止	歹	殳	母	毋	比	毛
氏	气	水	火	灬	爪	父	爻	爿	片	牙	牛	犬	王	礻	耂	⺝	5	玄	玉	瓜	瓦	甘	生	用	田
疋	疒	癶	白	皮	皿	目	矛	矢	石	示	禸	禾	穴	立	罒	衤	6	竹	米	糸	缶	网	羊	羽	老
而	耒	耳	聿	肉	臣	自	至	臼	舌	舛	舟	艮	色	艸	虍	虫	血	行	衣	西	7	見	角	言	谷
豆	豕	豸	貝	赤	走	足	身	車	辛	辰	邑	酉	釆	里	8	金	長	門	阜	隶	隹	雨	青	非	斉
9	面	革	韋	韭	音	頁	風	飛	食	首	香	10	馬	骨	高	髟	鬥	鬯	鬲	鬼	竜	11	魚	鳥	鹵
鹿	麦	麻	亀	12	黄	黍	黒	黹	13	黽	鼎	鼓	鼠	14	鼻	齊	15	歯	16	龍	龜	17	龠		

Fig. 3: Multi-radical Lookup GUI

New indexing techniques, such as those described above, have to date been largely confined to non-commercial packages based on the author's KANJIDIC project files (Breen, 2004). The commercial electronic kanji dictionaries in Japan, which are typically based on published kanji dictionaries, usually only provide radical, reading and occasionally stroke-count indices.

4 Usage Patterns in an Electronic Kanji Dictionary

The availability of a large range of indexing techniques in an electronic kanji dictionary raises the question of how useful they actually are to users of such dictionaries, and which methods are preferred by users. With dictionaries provided as software packages, measurement of the usage of the differing indexing techniques would be limited to such things as surveys of users. To date no analysis appears to have been carried out on user preferences in indexing methods.

One form of electronic kanji dictionary which is amenable to the direct measurement of usage patterns is the kanji dictionary component of WWW-based Japanese dictionary, such as the WWWJDIC server (Breen, 2003) developed by the author. The WWWJDIC server provides over twenty indices to its database of over 13,000 kanji, including all the techniques mentioned earlier in this paper. The users are primarily students of Japanese and non-native speakers of Japanese.

The server code at the Monash University site was extended to provide detailed statistics of the accesses to the kanji dictionary module. Information was collected over a two-week period, during which time over 70,000 accesses to the kanji dictionary were made. Table 2 contains a breakdown of the accesses by index type. In the case of accesses using the multi-radical method, it is clear that users frequently have to make several selections of radicals to reach the correct kanji. From inspection of the raw statistics, it appeared that on average three accesses were made by each user of this method for each target kanji. Accordingly, the reported accesses for this method have been reduced to make a more meaningful comparison with the other methods. The "Direct" method involves access to the kanji in a word encountered in another dictionary function, whereas the "Cut-Paste" method refers to kanji transferred from another WWW page or application.

Access Method	**Access %**
Multi-radical	24.8
Reading (ja)	24.1
Direct	17.6
Cut-Paste	9.7
English Meaning	9.2
Code-Point	6.4
Stroke Count	2.8
Reading (cn,ko)	1.5
Radical/Bushu	1.4
SKIP, 4-Corner	1.3
Dictionary Index	0.6
Other	0.6

Table 2: Kanji Access Statistics

(In 20.3% of the accesses recorded in Table 2, the user opted to make a follow-on search of one of the "word" dictionaries on the server using a kanji as a search key.)

These results are interesting for a number of reasons:

a. the index methods which dominate are either those which have only become available with electronic dictionaries: multi-radical, direct access, English meaning, etc., or those which can only be used via a secondary index in traditional dictionaries.

b. the high levels of access based on the code-points of the kanji, which includes the Direct, Cut-Paste and Code-Point methods is an indication of the usefulness of operating an electronic dictionary in association with other software. The relatively high result for the Code-Point method, which involves supplying the server with the hexadecimal representation of the kanji's code, was investigated further. Over 60% of these accesses used the Unicode code-point, and on inspection of the server logs it transpired that most arrived as linkages from other WWW servers and database collections dealing with kanji and hanzi.

c. the relatively low usage of the traditional radical index and the SKIP method is an indication that while they may be suitable and accepted in paper dictionaries as the primary indices, they play only a minor role in electronic dictionaries, where users clearly find other methods more useful.

It is recognized that this survey of usage patterns reflects both the preferences of the particular set of users who have chosen to use it, and the biases introduced by the interface, which in the case of HTML forms is often not as easily used as, for example, a tailored GUI. It is, however, a strong indication of the sorts of indexing methods which are found to be useful by such a group. It is also worth noting that despite the clumsiness of the Multi-radical selection form, which has over 200 check-boxes, it is clearly among the most popular kanji index methods.

5 Conclusion

Kanji dictionaries have traditionally been published using indexing techniques developed for use in the printed medium. Electronic dictionary techniques provide the opportunity both to interface such dictionaries directly with text, and also to introduce new techniques more suited to the computer-human interface. Implementation of such techniques and the subsequent measurement of their usage in an environment where users can choose from a variety of indexing methods indicates a high level of acceptance of and preference for the new indexing techniques.

References

Breen, J.W. 1995. *Building an Electronic Japanese-English Dictionary,* JSAA Conference, Brisbane.

Breen, J.W. 2003. *A WWW Japanese Dictionary,* in "Language Teaching at the Crossroads", Monash Asia Institute, Monash University Press

Breen, J.W. 2004. *KANJIDIC - Kanji Database Project,* http://www.csse.monash.edu.au/~jwb/kanjidic.html

Rosenthal, G. 2002. *JWPce: Japanese Word Processor* http://www.physics.ucla.edu/~grosenth/jwpce.html

Halpern, J. 1990. *New Japanese-English Character Dictionary,* Kenkyusha/NTC.

Halpern, J. 1999. *Kanji Learner's Dictionary,* Kodansha

Henshall, K.G. 1988. *A Guide to Remembering Japanese Characters,* Tuttle.

Japanese Industrial Standards Committee. 1997. *JIS X 0208-1997 7-bit and 8-bit Coded Kanji Sets for Information Interchange,* Japanese Standards Association.

Morohashi, T. et al. (1989) *Daikanwa Jiten,* (Large Character Dictionary), Taishukan.

Nelson, A.N. 1974. *The Modern Reader's Japanese-English Character Dictionary,* (second revised edition), Tuttle.

Nelson, A.N. revised Haig, J.H. 1997. *The New Nelson Japanese-English Character Dictionary,* Tuttle.

Sakade, F. et al. 1961. *A Guide to Reading & Writing Japanese,* (second edition), Tuttle.

Shibano, K. et al. 1997, 2002. *JIS Kanji Dictionary,* (first and second editions), Japan Standards Association.

Spahn, M. & Hadamitzky, W. 1996. *The Kanji Dictionary,* Tuttle.

Spahn, M. & Hadamitzky, W. 1997. *Kanji & Kana: A Handbook of the Japanese Writing System* Tuttle.

Tanaka, Y. et al. 2000. *Unicode Kanji Information Dictionary,* Sanseido.

Ueda, K. et al. 1963. *Daijiten,* (Large Character Dictionary), Kodansha.