BUILDING AN ELECTRONIC JAPANESE-ENGLISH DICTIONARY

J.W. Breen
Monash University.

June 22, 1995

Abstract

This paper describes an on-going project to develop and maintain a comprehensive electronic Japanese-English dictionary capable of use within a variety of search-and-display, electronic-text reading support, and machine translation environments. The project consists of two parts: (a) the compilation of two major datafiles; a Japanese-English lexicon (EDICT), and a kanji information database (KANJIDIC). At the time of writing, the former has over 100,000 entries, and (b) the development of software to index, search and display entries in the data files. This software, which has now been released on a variety of computing platforms, including Unix, PC (DOS and Windows) and MacIntosh, can operate as either a stand-alone dictionary, providing the functions of both normal word/phrase dictionaries and character dictionaries, or as a support package for reading electronic text, by automatically glossing selected words and phrases.

1 INTRODUCTION

This paper describes a project to build an electronic Japanese-English Dictionary system. The project has been under way since early 1991, and combines both a compilation of dictionary material in computer files, and the development of computer software to search the files and display the contents.

The project began as experimentation by the author with techniques for the manipulation and display of Japanese text on a PC. Until the late 1980s, there had been virtually no use of Japanese text in computers in the West, and there was a common perception that it was an inordinately difficult task, involving specialized hardware to hold and display kana and kanji. As can be expected, there had been considerable development in these fields in Japan, including the establishment of comprehensive corporate and national standards for the representation of kana and kanji, and the development of local variants of operating and file systems which incorporated these techniques, as well as a mass of application software. The developments within Japan all relied on either extensions to the operating systems, or special hardware incorporating ROMs of character fonts, or both, which greatly inhibited its use outside Japan.

During 1989 things began to change. Ken Lunde produced and disseminated over the Internet his seminal `japan.inf` file [1], which described the coding standards and text manipulation techniques. (In 1993 he expanded the contents of this file greatly to produce his book "Understanding Japanese Information Processing" [2].) Izumi Ohzawa at Berkeley produced and released the "KD" package which enabled the display of Japanese text on non-Japanese PCs, and Mark Edwards began work on a kana/kanji text editor (MOKE) capable of being used on the simplest of PCs. It was the purchase of a copy of MOKE V2.0 in September 1990, which incorporated a simple 2,000-entry Japanese-English glossary file and a crude technique for searching and adding to that file, that inspired (or provoked) the author into experimenting with text handling software, using the development of more sophisticated dictionary techniques as a pretext.

From this relatively humble beginning has developed both a major set of dictionary files: EDICT, a Japanese-English dictionary file which now has over 100,000 entries; and KANJIDIC, an information file with details on each of the 6,353 kanji included in the JIS X 0208 standard, as well as a growing number of software packages by the author and others which implement a variety of dictionary functions on most modern personal computers and workstations. These packages, which have been distributed without charge, have proved very popular, and have thousands of users worldwide.

*Paper delivered at the Japanese Studies Association of Australia Conference, July 1995, Brisbane, Queensland, Australia.

2 DICTIONARY FILE DESIGN

The compilation of the dictionary files, and the lexicographic principles employed are described fairly fully elsewhere [3]. In summary, the main EDICT file is a simple text file, with a single line for each entry, each of which has the format:

KANJI [kana] /English/English/..../

if the head-word includes kanji, or:

kana /English/English/..../

if the head-word is in kana alone.

This is the format employed in the first version of the file supplied with MOKE, and has been retained, despite a number of inherent limitations, in order to enable existing software to continue to use the files. The file need not be in any particular order, although it is usually kept sorted to facilitate updating.

The following are examples of EDICT entries.

借り宅 [かりたく] /rented house/
借り住い [かりずまい] /living in rented quarters/
借り間 [かりま] /rented room/
借り店 [かりだな] /rented shop/

The KANJIDIC file is also a text file with one line per kanji. The information includes the on and kun readings, the primary radical and stroke count, and typical meaning(s) of the kanji, the indices to the entries in many popular kanji dictionaries (Nelson, Halpern, Morohashi, Gakken, Spahn & Hadamitsky, etc.) and indexing codes such as Four-Corner and SKIP. Both Jack Halpern and Mark Spahn from their respective dictionaries for inclusion in the file.

The following are some examples of entries in the KANJIDIC file:

借 3c5a U501f N490 B9 S10 G4 H122 F888 P1-2-8 L1186 K996 Q2426.1 MN781 MP1.0837 E502 Yjie4 Yji1 シャク か.りる {borrow} {rent}

宅 4270 U5b85 N1279 B40 S6 G6 H2174 F421 P2-3-3 L1916 K371 Q3071.4 MN7064 MP3.0898 E928 Yzhai2 Yzhe4 タク T け たか たけ や やけ {home} {house} {residence} {our house} {my husband}

The development of the EDICT and KANJIDIC files has been a cooperative effort by scores of contributors around the world, under the coordination of the author, who has carried out the overall compilation and provided editorial control. The development has triggered related compilations, the most important being a Japanese-German file in EDICT format compiled by Helmut Goldenstein from the Langenscheidt edited by Wolfgang Hadamitzky (with permission), and the Life Sciences Dictionary file of bio-medical terms compiled by Shuji Kaneko of the Pharmacology Department at Kyoto University and his associates.

In EDICT file has also been used to assist in the building of lexicons in several Machine Translation projects.

3 DICTIONARY SEARCHING

One of the main goals of the project was to exploit the capabilities of computer files and software in order to provide and experiment with a wide range of techniques for finding, selecting and displaying dictionary entries. At a minimum, such techniques should provide the same functions as a traditional paper dictionary, however it was hoped that additional techniques could be developed which would go beyond the traditional facilities.

Also, with the compilation of both word/phrase and kanji files, it was intended that the system be capable of emulating the facilities of the two traditional types of dictionary, and be able to move easily between the two.

As well as being able to select dictionary entries according to a number of criteria, it was also considered desirable that the entries be displayed in appropriate lexical orders, e.g. alphabetical order for English keywords, or gojuuon for kana keywords. In order to support a rapid search of the files for matching keywords, and at the same time produce an ordered display, it was decided to use an auxiliary index file in conjunction with the text files. As well as providing all the search and display support, the separation of the index from the dictionary file meant that the latter could be kept in a simple format, and continue to be edited using Japanese-capable text editors. A relatively simple utility program regenerates the index whenever the text file is modified.

The indexing technique developed for use in this system consists of parsing the entire dictionary file and generating an index to the start of each lexical token, a token in this case being a sequence of alphabetic characters (i.e. an English word) or a sequence of Japanese characters. In addition an index was generated for every kanji occur-

```
XJDIC SEARCH KEY: supplant

Searching for: supplant

8: supplanting 【下克上】 (げこくじょう)
retainer supplanting his lord; juniors
dominating seniors
8: supplanting 【下剋上】 (げこくじょう)
juniors dominating seniors; retainer
supplanting his lord
End of 8 character matches. Continue for shorte
r matches? (y/n)

ROMAJI ENTRY: はんしん
Searching for: はんしん
4: はんしん 【反身】 bending backward;
strutting
4: はんしん 【半神】 demigod
4: はんしん 【半身】 half the body; half
length
4: はんしん 【阪神】 Osaka-Kobe
4: はんしん 【叛臣】 rebellious retainer
4: はんしん 【叛心】 rebellious spirit
4: はんしんぞう 【半身像】 half-length statue
or portrait; bust
4: はんしんはんぎ 【半信半疑】 half in doubt;
dubious; incredulous
4: はんしんふずい 【半身不随】 paralyzed on
one side
4: はんしんよく 【半身浴】 sitz bath
4: はんしんろん 【汎神論】 pantheism
End of 4 character matches. Continue for shorte
r matches? (y/n)
```

Figure 1: Example of `xjdic` Display

ring within a sequence in order to be able to find compounds which contain a specified kanji in a non-initial position. Common English words such as "the, "for" and "but" are excluded, as are all words of only one or two characters. The table of indices is then sorted according to the lexical value of the token associated with each index, producing an ordered index file, which is used in a "binary search" in order to identify tokens which match a desired search key.

Figure 1 shows the output from searches involving English and Japanese (kana) keywords.

A similar parse/index/binary-search approach is also followed for the KANJIDIC file, although there are some differences to cater for numeric keys (e.g. Nelson index number) or for searching for kanji via combinations of criteria such as bushu plus stroke-count.

4 APPLICATION SOFTWARE

4.1 MS-DOS

The first software to be made available from the project was version 1.0 of the DOS dictionary program JDIC, released in March 1991. At that stage the EDICT file contained only 6,000 entries, and the program only supported searching with English and kana keys. Entry of kana keys is made by setting the input to kana-mode, and then typing in either Hepburn or kunrei ro-maji, which is converted to kana immediately. This basic dictionary display has continued in later releases, and has been enhanced through the addition of additional features such as the ability to select and write entries to a file for later study or editing, and the ability to select search keys from the display.

With the compilation of the first part of the KANJIDIC file later in 1991, an enhanced version of JDIC was released which provided for the selection of kanji by a number of criteria, the display of summary information about the kanji, and an optional display of dictionary entries which either begin with or contain that kanji. This facility has been progressively enhanced to increase the number of selection criteria, and to enable the movement of selection fields in both directions between the kanji function and the ordinary dictionary function.

The process of using the kanji-display function is:

- the specification of a selection criterion, e.g. index number, bushu, stroke-count, reading, etc.
- where more than one kanji meets the criterion, the display of those kanji, followed by the selection of the desired kanji.
- the display of all the information about that kanji.
- an optional display of entries containing the kanji.

Figures 2, 3 and 4 show these steps.

```
        J D I C - Japanese-English Dictionary System  Version 2.5       (H)
 Kanji Selection  - Radical No. 85 - 9 stroke characters
▬洶    洳    洙    洸    洽    洫    洌    洵    洲    洩    洛    涑    津    洪
 0     0     0    387    0     0     0    383   391    0     0     0    390   386
 9     9     9     9     9     9     9     9     9     9     9     9     9     9
 洒    洞    派    浄    浅    洋    洗    活    海
 0    380   381   382   389   392   388   385   384
 9     9     9     9     9     9     9     9     9









Position Cursor and Enter to select, (M)ore Display or e(X)it function:
```

Figure 2: Example of JDIC Display of kanji meeting a bushu/stroke count combination

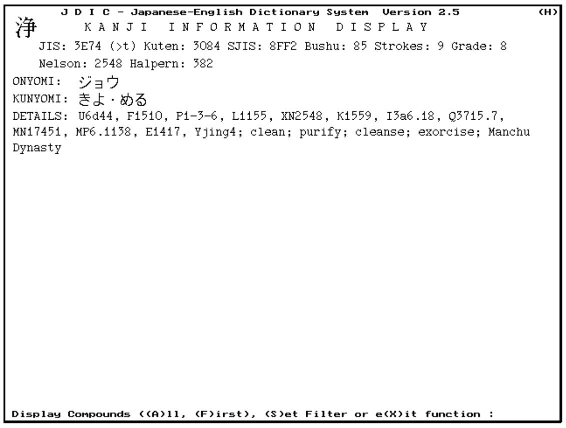

Figure 3: Example of JDIC Kanji Information Display

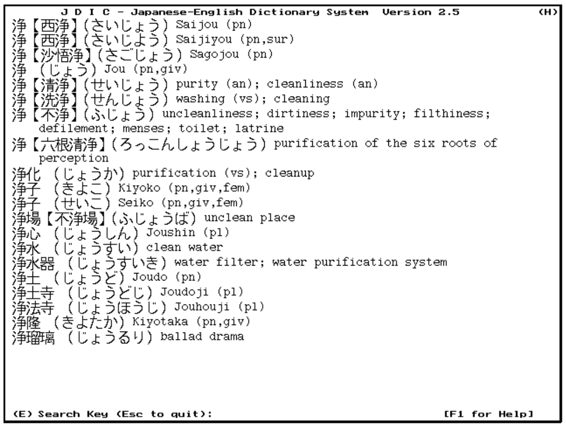

Figure 4: Example of JDIC Display of entries containing a selected kanji

Where a kanji has been selected from the dictionary display, the procedure begins at the third step.

As one of the major potential uses of an electronic dictionary is to support the reading of text, which is increasingly available in electronic form, a second (DOS) PC program: JREADER was developed and released in early 1992. This program combines two functions:

- the display of lines of a text file, which may include Japanese characters in any of the common coding methods (JIS, EUC, Shift-JIS). The usual text-reading functions such as page-up/down, skip to specified text, etc. are available. This display takes place in an upper window on the screen.
- a dictionary display in a lower window, in which the results of dictionary searches for keywords from the upper window are displayed. Keywords are selected from the upper window by placing the cursor on the word and pressing a key to indicate the type of search. Both the normal (EDICT) dictionary search, and the kanji display are available. Also available is a kanji compound search against a "reverse-henkan" file which contains the readings of over 250,000 compounds. (This latter file has been compiled from the kana-kanji henkan files of the SKK and WNN systems, which are public-domain kana-to-kanji conversion systems for Unix workstations.)

Figure 5 shows a typical JREADER display.

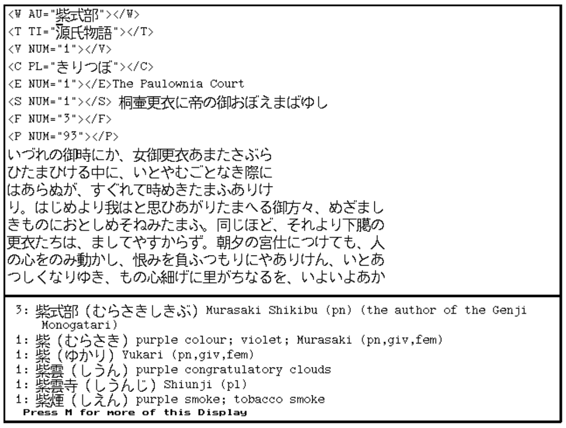

Figure 5: Example of JREADER Display

There are two ways in which the JREADER dictionary search differs from that use in JDIC:

- instead of seeking an exact match of keyword against dictionary entry, the longest possible match(es) are obtained and displayed, followed by the next longest, and so on.
- where the indicated search keyword is a kanji followed by one or two kana, the kana portion is checked against a table of common verb and adjective inflections. If a match is found, the possible dictionary-form words are sought and, if a match is found, displayed first.

The current releases of JDIC and JREADER are V2.4, which have been available since early 1994. V2.5 of these programs are expected to be released in mid-1995.

The JREADER and JDIC programs have been successfully operated on all types of PCs, from slow XTs with CGA graphics up to Pentiums. An interesting application has been their successful operation on the tiny "palmtop" PCs, such as the HP100LX. In this environment, they are often seen as a competitor to the tailored hand-held electronic dictionaries, such as the Canon Wordtank.

4.2 WINDOWS

A Windows dictionary application, WinJDic, which uses the same dictionary and index files as JDIC, was written in 1993/4 by Mark Edwards. It has some but not all of the search functions as JDIC, but has a more elegant and friendly user interface. WinJDic is freeware, and a commercial version, Kihon, has also been released.

4.3 UNIX

During 1992, the author developed a variant of the dictionary software to operate in the X-Windows environment on workstations using versions of the Unix operating system. This package, known as `xjdic`, must operate within a Japanese-capable "xterm" window, such as is provided by the kterm (kanji xterm) package. The `xjdic` program draws heavily on the services of the X-Windows and kterm environments, in particular the built-in Japanese character displays, and the ability to "cut and paste" text from one window to another using mouse functions. This means that that a copy of `xjdic` operating in one window can support the reading of a text file displayed in another window, or a document being keyed in another window.

The `xjdic` program, as is usually the case with freely available Unix software, is distributed as a source program which is compiled and installed on each installation. It has been successfully operated on workstations of all varieties in many different locations, and the source code has been used by other developers of Japanese text-handling software.

Figure 1 is a sample of a display window from `xjdic`.

The current version of `xjdic` is 1.2, and version 2.0 is planned for release in mid-1995. As well as a number of other enhancements, this version will be capable of operating in a client/server mode, with a single dictionary file and search program able to serve requests from many users on a network.

Another dictionary display system which uses the EDICT and KANJIDIC files is the `lookup` package developed by Jeffrey Friedl, an engineer working at Omron in Kyoto. This program uses a simpler indexing technique which does not produce lexically ordered displays, but is capable of very flexible searches, including the capability of over-riding the distinctions between おお and おう, or ず and づ. This program is also the basis of his popular WWW dictionary server (`http://www.wg.omron.co.jp/cgi-bin/j-e`).

4.4 MACINTOSH

In 1993 Dan Crevier, a graduate student in biophysics at Harvard, took the `xjdic` program and converted it to operate on MacIntosh computers equipped with the Kanjitalk or JLK (Japanese Language Kit) operating systems. He also used some of the code provided to him by the author from JDIC for this redevelopment. The resulting program, called MacJDic, has proved to be very popular, particularly in Japan where the MacIntosh is beginning to achieve considerable penetration. At the time of writing it has been included on four CD-ROMs, and as been listed among the 100 most popular MacIntosh programs in Japan.

The author was pleased to hear recently that the team at the East-West Center at the University of Hawaii who are working on the new edition of the Nelson character dictionary for Tuttle are using MacJDic to assist them in their work. (Dr John Haig from that Center has kindly agreed to make the index numbers from the New Nelson available for inclusion in the KANJIDIC file.)

The current version of MacJDic is 1.3.4. Dan Crevier has also written his own dictionary display system, for which he intends a commercial release.

5 WORD PROCESSORS

Three low-cost PC word-processor packages use the EDICT and KANJIDIC files to provide dictionary services for users: MOKE (uses EDICT alone), NJSTAR (uses EDICT with its own indexing system, and a kanji file derived from KANJIDIC), and JWP (uses the same data and index files as JDIC.)

6 PLANNED ENHANCEMENTS

For the past three years the basic structure of the EDICT and KANJIDIC files has remained unchanged, with the major work going into increasing the amount of information contained in them, and enhancing the functionality of the dictionary software. There are some major structural limitations with both files, in particular the problem with EDICT in the recording of compounds with alternative readings, and compounds with okurigana variants. With the large amount of software now dependent on those files, it is difficult to move to a new format without causing considerable disruption.

Two enhancements to the structure of the EDICT are planned for introduction in 1995. These changes, which will have minimal impact on existing software, and have the potential to enhance considerably the usefulness of the files and software, are:

- Priority English Keywords

  Although the EDICT file, in conjunction with appropriate software, has the potential to operate as an English-Japanese dictionary, this function has become less useful as the size of the file has grown. For example, the word "house" occurs in over eighty entries, so a search for that word is of limited use, particularly for a learner. It is planned to prepend the character "@" to the occurrence of English words within entries where it is appropriate that that word be treated as a head-word. Appropriately equipped dictionary software will be able to operate in two modes: a "find-all-occurrences" mode and a "find-priority-keywords" mode.

  The software modifications to the JDIC and `xjdic` programs to implement this enhancement have been completed, and will be released during 1995, although it may take some years to mark all the appropriate words in the EDICT file. Once this task is completed, the file and software will be able to function far more effectively in an English-Japanese orientation.

- Extension File

  A limitation of the present EDICT file is its rather terse format, and its inability to include examples of the use of words, as Japanese characters cannot occur in the English translational equivalent fields.

  It is planned to extend the information in selected EDICT entries by including short articles in an EDICT Extension File. This will be a text file in which each article will be marked by one or more (Japanese) keywords. The articles can contains such things as examples of the use of the Japanese words, further explanatory information, etc. As each article is added to the Extension File, the matching entries in the EDICT file will have an appropriate tag such as "qv" added to them to indicate to users that further information is available.

  Modifications to the JDIC, JREADER and `xjdic` programs have been prepared which will enable users to request the display of the article from the Extension File for any tagged entry. As this is a separate file, its development will have no impact on existing software.

7 CONCLUSION

The software and files developed in the Japanese-English electronic dictionary project described in this paper have amply demonstrated the power and amenity of such a system. The system has proved capable of providing a service at least equivalent to traditional dictionaries, and in several areas, such as the ability to identify kanji via a variety of methods, the integration of the functions of the kanji dictionary and the Japanese-English dictionary, and the automatic glossing of electronic text, it has proved to be of greater power and capacity than any combination of traditional dictionaries.

The packages from the project have filled an important niche, and will doubtless continue to be used for years to come. In many ways the scholar and student working in Japanese and English now has better electronic dictionary facilities than those available for virtually any other pair of languages.

Ironically, however, the technology which has supported their development is also leading to the development of commercial systems which, for the serious scholar and translator, will inevitably become the favoured electronic tool. The growth of the CDROM-based Electronic Book (EB) industry in Japan has been rapid, with many major Japanese dictionaries appearing in that form. Already in 1995 we have seen a version of the Halpern character dictionary appear in EB form, and the new editions of the Nelson dictionary and the Spahn and Hadamitzky dictionary planned for completion in 1995 are both intended to incorporate EB versions.

8 REFERENCES

[1] K. Lunde, `japan.inf`, March 1992, (ftp.cc.monash.edu.au:pub/nihongo/japan.inf and other ftp sites)

[2] K. Lunde, *Understanding Japanese Information Processing*, O'Reilly & Associates Inc., 1993

[3] J.W. Breen, *A Japanese Dictionary Project (Part 1: The Dictionary Files)*, Department of Robotics & Digital Technology Technical Report, Monash University, 1993.