1

A WWW JAPANESE DICTIONARY

J.W. Breen
School of Computer Science & Software Engineering
Monash University.

February 16, 2000

Abstract

This paper presents a non-technical overview of the provision of an extended Japanese-English dictionary service on the World-Wide Web (WWW). The service described it the author's WWWJDIC server, which is part of the EDICT project. As well as providing linked waei and kanwa dictionary services, it also has the facility to provide English glosses of words in Japanese text, including other WWW pages.

Introduction

Since 1991, the author has been engaged in the EDICT (Electronic DICTionary) project to develop a series of computer-based Japanese-English dictionaries, capable of being used both as traditional dictionaries and as semi-automated aids for reading Japanese text. The main EDICT glossary file now has over 60,000 entries, and has been joined by subject-specific files covering bio-medical terminology, legal terms, computing, telecommunications, business, etc., as well as a proper names file with 160,000 entries and a kanji database covering over 12,000 kanji. A variety of software packages have been released for use on a number of computer systems, and the files are used within several free or shareware Japanese word-processor systems. The files, which have also been used in a number of natural-language processing (NLP) and machine translation (MT) projects, are all available free of charge for non-commercial use.

The development of the World-Wide Web as an information retrieval system on the Internet in 1993 opened the possibility of providing a comprehensive dictionary facility from a small number of servers. The facilities within the WWW to combine server-based software with text input from almost any browser has meant that an identical service can be provided regardless of the user's type of computer. Also complex software distribution and installation is avoided, and the central lexicographical databases can be continually expanded and the services enhanced without causing problems for the users.

The first WWW-based dictionary using the EDICT files began operating in 1993, and since then approximately 10 different server systems have been developed to use these files. This article describes the dictionary and related services provided by the author's

*Paper delivered at a Japanese Studies Centre Symposium, July 1999, Melbourne, Victoria, Australia.

WWWJDIC server[1], which operates at Monash University, and from mirror servers in the USA, Canada, Chile and Japan.

WWWJDIC Facilities

The WWWJDIC server provides the following facilities:

a. a keyword search in one of the eight lexicographic files currently available. Each entry in the file typically consists of a jukugo, its reading in kana, and a short English gloss. The keywords entered in the search can either be in Japanese or in English. In the case of a Japanese keyword, it can be in kanji and kana, entered using an IME (Input Method Editor) or cut and paste from another screen or program, or entered in rōmaji. Figure i shows an example of a typical word search in WWWJDIC.

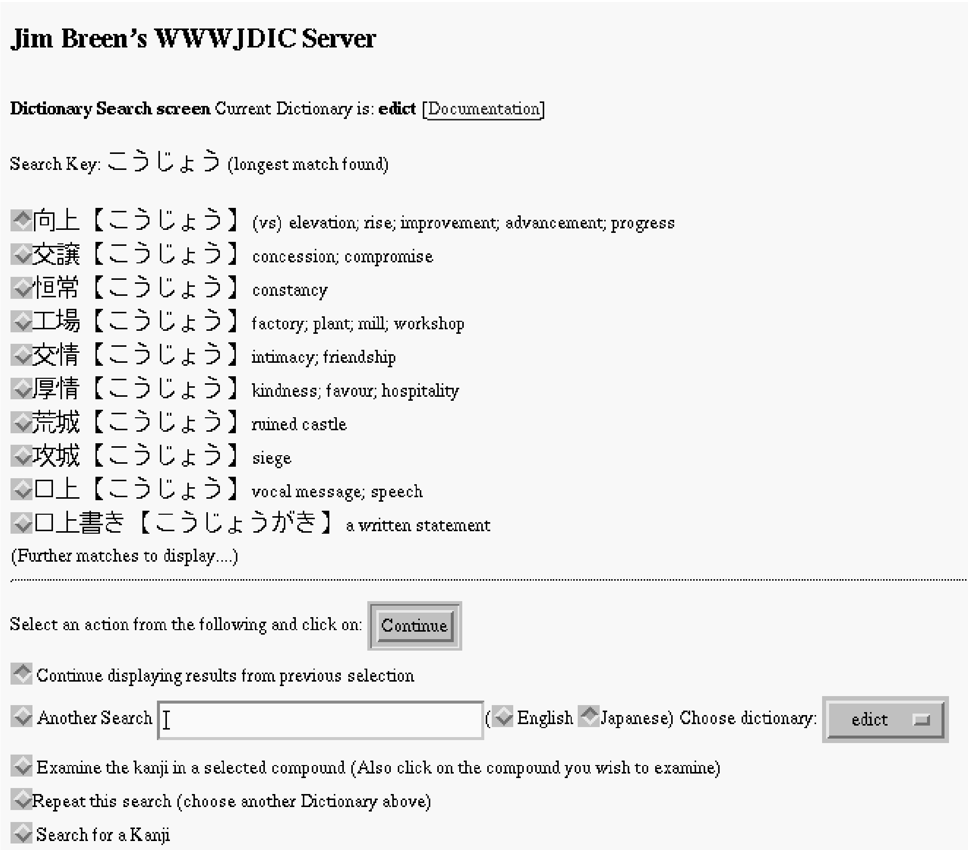


Figure i: WWWJDIC result when searching for こうじょう.

b. a kanji selection facility, in which kanji can be identified by a wide variety of methods ranging from traditional bushu/stroke-count to coding systems such as Halpern's SKIP, De Roo codes, Four Corner[2], etc. Kanji readings and English senses can also

[1] `http://www.csse.monash.edu.au/~jwb/wwwjdic.html`

[2] these are all numeric codes based on the stroke-counts of identifiable portions of kanji. Halpern's SKIP (System of Kanji Indexing by Patterns) is used to order and index kanji in his New Japanese-English Character Dictionary (Kenkyusha, Tokyo 1990) and Kanji Learner's Dictionary (Kodansha, Tokyo 1998). De Roo's code is used in his "2001 Kanji" (Bonjinsha). The Four Corner code was developed by Wang Chen in 1928 and is widely used in Chinese and Japanese dictionaries. As an example, the kanji 村 has a SKIP of 2-4-3 indicating a vertical division into 4 and 3 stroke portions, a De Roo code of 1848

be used. One novel feature is the classification of kanji according to their basic shape components, with kanji being identifiable by several components instead of a single bushu. Figure ii shows an example of the result of a kanji selection. The coded information after the kanji includes indexes into several dictionaries: Nelson, Halpern, Spahn & Hadamitzky, Morohashi, etc., as well as readings in Korean and Chinese.

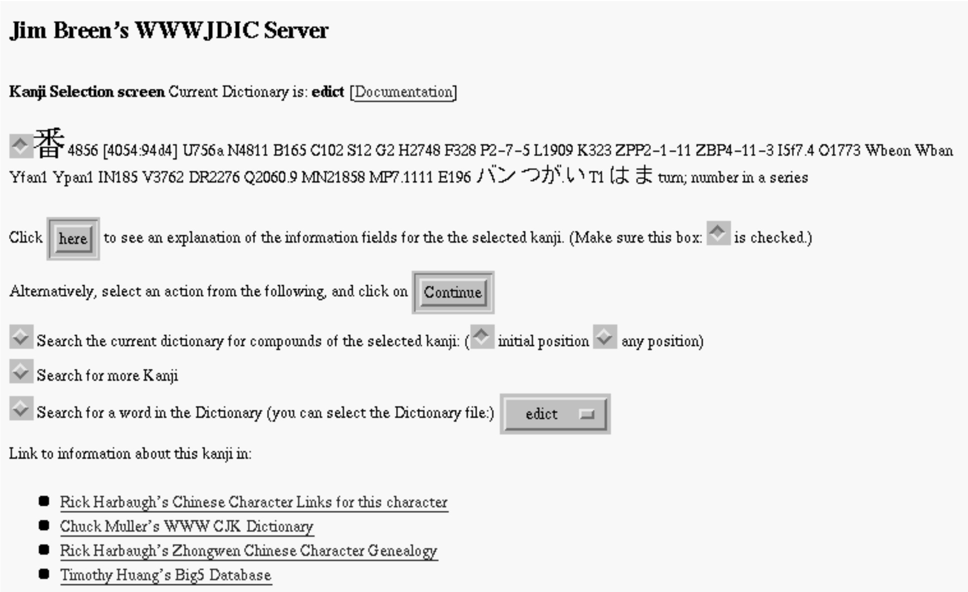


Figure ii: Kanji dictionary display for 番。

c. the capability for the user to move flexibly between the kanji-oriented and text-oriented dictionary files. For example, having identified a kanji, it is possible to retrieve entries in the dictionary files which contain that kanji, either in the first character position or in any position in a word. Similarly, it is possible to examine the details of any kanji from a retrieved dictionary entry. In this sense the WWW dictionary is able to combine the features of both a Japanese-English/English-Japanese dictionary and a kanwa dictionary.

d. the capability to annotate Japanese text with the English glosses of the words within it. The text can either be cut and pasted from another page or program, or can come from a selected WWW page. Figure iii shows an example of this facility. This is a major feature of the WWWJDIC server and is described in the following section.

Text Glossing

The ability to use dictionary files to gloss text is a powerful adjunct to computerized dictionaries. The files of the EDICT project have often been used for this purpose, with earlier examples including the author's JREADER program[3], Hatasa & Henstock's AutoGloss/J Package[4], Yamamoto's Mailgloss system[5], Kitamura & Tera's DLink system[6],

---

representing 木 (18) and 寸 (48), and a Four Corner code of 4490 because there is a 十 (4) at the top two corners and a 小 (9) at the bottom left.

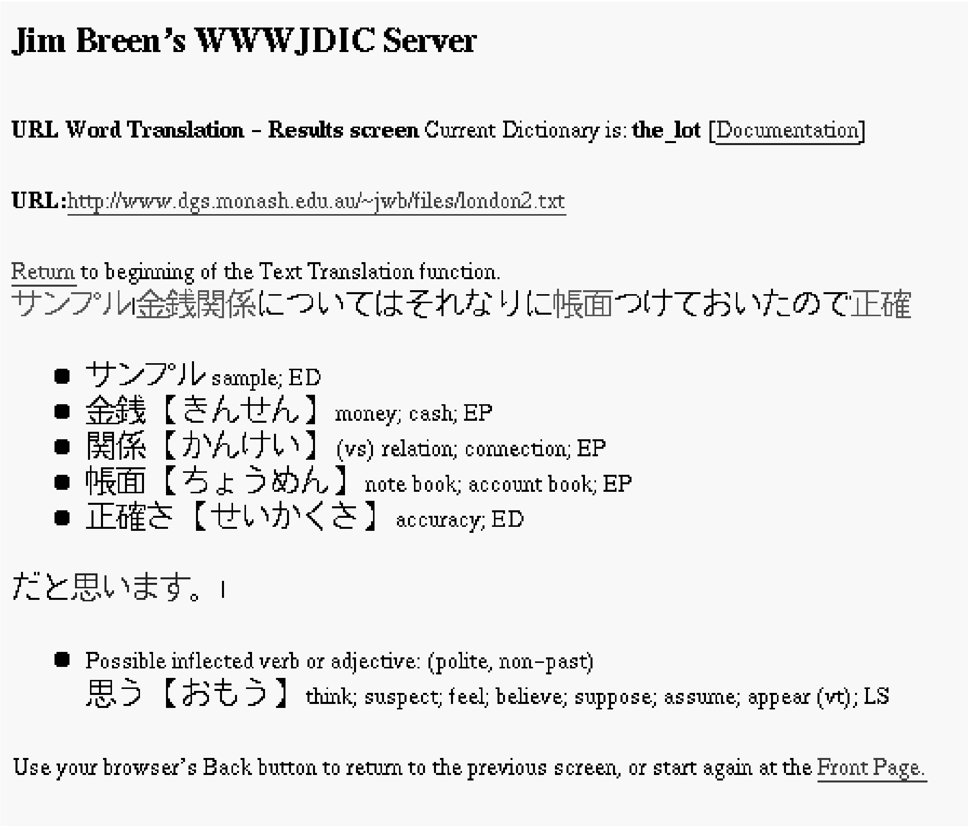
[3] `http://www.csse.monash.edu.au/~jwb/japanese.html#edict_proj`

[4] `http://www.sla.purdue.edu/academic/fll/JapanProj/AutoGloss/AutoGloss.html`

[5] `http://www.intersc.tsukuba.ac.jp/~yamagen/mg/`

[6] `http://basil.cs.inf.shizuoka.ac.jp/~kitamura/DLS/dls-e.html`

etc.



Jim Breen's WWWJDIC Server

URL Word Translation - Results screen Current Dictionary is: **the_lot** [Documentation]

URL:http://www.dgs.monash.edu.au/~jwb/files/london2.txt

Return to beginning of the Text Translation function.
サンプル|金銭関係についてはそれなりに帳面つけておいたので正確

- サンプル sample; ED
- 金銭【きんせん】money; cash; EP
- 関係【かんけい】(vs) relation; connection; EP
- 帳面【ちょうめん】note book; account book; EP
- 正確さ【せいかくさ】accuracy; ED

だと思います。|

- Possible inflected verb or adjective: (polite, non-past)
  思う【おもう】think; suspect; feel; believe; suppose; assume; appear (vt); LS

Use your browser's Back button to return to the previous screen, or start again at the Front Page.

Figure iii: Example of the glossing of words in Japanese text.

In carrying out a glossing of Japanese text, a degree of processing of the text must be carried out beforehand, in particular to segment the text into its lexemes and to convert the inflected forms of words into their dictionary forms. These tasks are non-trivial for Japanese text, and have led to the development of powerful morphological analysis software tools such as ChaSen[7] and JUMAN[8]. These tools are generally too large and slow to use with the WWW, where a rapid response is essential.

With WWWJDIC a simpler approach to segmentation has been employed in which the text is scanned to identify in turn each sequence of characters beginning with either a katakana or a kanji. The dictionary is searched using each sequence as key, and if a match is made, the sequence is skipped and the scan continues. Thus the dictionary file itself plays a major role in the segmentation of the text in parallel with the accumulation of the glosses. The technique cannot identify grammatical elements and other words written only in hiragana, however is it quite successful with gairaigo and words written using kanji.

A further element of preprocessing of text is required for inflected forms of words, as the EDICT files only carry the normal plain forms of verbs and adjectives. A technique previously employed in the author's JREADER program is used here, wherein each sequence comprising a kanji followed by two hiragana is treated as a potential case of an inflected word. Using a table of inflections, a list of potential dictionary form words is created and tested against the dictionary file. If a match is found, it is accepted as the

[7]http://cactus.aist-nara.ac.jp/lab/nlt/chasen/
[8]http://pine.kuee.kyoto-u.ac.jp/nl-resource/juman.html

appropriate gloss. The table of inflections has over 300 entries and is encoded with the type of inflection which is reported with the gloss. Although quite simple, this technique has been extensively tested with Japanese text and correctly identifies inflected forms in over 95% of cases. (In Figure iii this can be seen where 思います has been identified as an inflection of 思う.)

When preparing glosses of words in text, it is appropriate to draw on as large as a lexicon as possible. For this reason, a combination of all the major files of the EDICT project is used, unlike the single word search function where users can select which glossary to use. This can introduce other problems as the inappropriate entry may be selected. For example, for the word 人々 the ひとびと entry must be selected, not the much less common にんにん. To facilitate this, a priority system is employed in which preference is given in turn to entries from:

a. a 12,000 entry file of more commonly used words;

b. the rest of the EDICT file;

c. the other subject-specific files;

d. the file of names.

Use of WWWJDIC by other systems

It is possible to interface other WWW systems to the WWWJDIC software. An interesting example of this is the Japanese Text Initiative at the University of Virginia library[9]. As part of this project, a "portal" system has been developed which allows individual words to be selected from texts and examined via WWWJDIC.

Conclusion

The WWW, with its ability to associate central data files and server software, and be accessed flexibly by innumerable users, has opened the possibility of extensive sophisticated dictionary facilities being provided to many people at little cost. These facilities can extend beyond those of traditional paper dictionaries by providing additional serves such as integrated kanji and text dictionaries, access using several different keys, and automated glossing of text.

At present many of the systems are experimental, however as more extended lexicons become available online, and as server and browser software become more advanced, the WWW is likely to play an increasingly important role in language study and multi-lingual communications.

[9] http://etext.virginia.edu/japanese/