JMdict: a Japanese-Multilingual Dictionary

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

Abstract

The JMdict project has at its aim the compilation of a multilingual lexical database with Japanese as the pivot language. Using an XML structure designed to cater for a mix of languages and a rich set of lexicographic information, it has reached a size of approximately 100,000 entries, with most entries having translations in English, French and German. The compilation involves information re-use, with the French and German translations being drawn from separately maintained lexicons. Material from other languages is also being included. The file is freely available for research purposes and for incorporation in dictionary application software, and is available in several WWW server systems.

1 Introduction

The JMdict project has as its primary goal the compilation of a Japanese-multilingual dictionary, i.e. a dictionary in which the headwords are from the Japanese lexicon, and the translations are in several other languages. It may be viewed as a synthesis of a series of Japanese-Other Language bilingual dictionaries, although, as discussed below, there is merit in having this information collocated.

The project grew out of, and has now subsumed, an earlier Japanese-English dictionary project (EDICT: Electronic Dictionary) (Breen, 1995, 2004a). With Japanese being an important language in world trade, and with it being the second most common language used on the WWW, it is not surprising that there is considerable interest in electronic lexical resources for Japanese in combination with other languages.

2 Project Goals and Development

As mentioned above, the JMdict project grew out of the bilingual EDICT dictionary project. The EDICT project began in the early 1990s with a relatively simple goal of producing a Japanese-English dictionary file that could be used in basic software packages to provide traditional dictionary services, as well as facilities to assist reading Japanese text. The format was (and is) quite simple, comprising lines of text consisting of a Japanese word written using kanji and/or kana, the reading (pronunciation) of that word in kana, and one or more English translations.

By the late 1990s, the file had outgrown its humble origins, having reached over 50,000 entries, and having spun off a parallel project for recording Japanese proper nouns (see below). The material has partly been drawn from word lists, vocabulary lists, etc. in the public domain, and supplemented by material prepared by large numbers of users and other volunteers wishing to contribute. While it had been used in a variety of software systems, and as a source of lexical material in a number of projects, it was clear that its structure was quite inadequate for the lexical demands being made by users. In particular, it was not able to incorporate a suitable variety of information, nor represent the orthographical complexities of the source language. Accordingly, in 1999 it was decided to launch a new dictionary project incorporating the information from the EDICT file, but expanded to include translations from other languages with the Japanese entries remaining as the pivots. The project goals were:

a. a file format, preferably using a recognized standard, which would enable ready access and parsing by a variety of software applications;

b. the handling of orthographical and pronunciation variation within the single entry. This addressed a major problem with the EDICT format, as many Japanese words can be written with alternative kanji and with varying portions in kana *(okurigana),* and may have alternative pronunciations. The EDICT format required each variant to be treated as a separate entry, which added to the complexity of maintaining and extending the dictionary;

c. additional and more appropriately associated tagging of grammatical and other information. Certain information such as the part of speech or the source language of loan words had been added to the EDICT file in parentheses within the translation fields, but the scope was limited and the information could not easily be parsed;

d. provision for differentiation between different senses in the translations. While basic indication of polysemy had been provided in the EDICT file by prepending (1), (2), etc. to groups of translations, the result was difficult to parse. Also it did not support the case where a sense or nuance was tied to a particular pronunciation, as occurs occasionally in Japanese;

e. provision for the inclusion of translational equivalents from several languages. The EDICT dictionary file was being used in a number of countries, and several informal projects had begun to develop equivalent files for Japanese and other target languages. A small Japanese-German file (JDDICT) had been released in the EDICT format. There was considerable interest expressed in having translations in various languages collocated to enable such things as having a single reference file for several languages, cross-referencing of entries, cross-language retrieval, etc. as well as acting as a focus for the possible development of translations for as yet unrepresented languages;

f. provision for inclusion of examples of the usage of words. As the file expanded, many users of the file requested some form of usage examples to be associated with the words in the file. The EDICT format was not capable of supporting this;

g. provision for cross-references to related entries;

h. continued generation of EDICT-format files. As a large number of packages and servers had been built around the EDICT format, continued provision of content in this format was considered important, even if the information only contained a sub-set of what was available.

An early decision was to use XML (Extensible Markup Language) as a format for the JMdict file, as this was expected to provide the appropriate flexibility in format, and was also expected to be supported by applications, parsing libraries, etc.

An examination was made of other available dictionary formats to ascertain if a suitable formatting model was available. It was known that commercial dictionary publishers has well-structured databases of lexical information, and some were moving to XML, but none of the details were available. A large number of bilingual dictionary files and word lists were in the public domain; however in general they only used very simple structures, and none could be found which covered all the content requirements of the project. The dictionary section of the TEI (Text Encoding Initiative), which at the time of writing has a well-developed document structure for bilingual dictionaries, was at that stage quite limited (Sperberg-McQueen et al, 1999). Accordingly, an XML DTD (Document Type Definition) was developed which was tailored to the requirements of the project.

The EDICT file was parsed and reformatted into the JMdict structure, and at the same time, many of the orthographical variants were identified and merged. The initial release of the DTD and XML-format file took place in May 1999. At that stage, it contained the English translations from the EDICT file and the German translations from the JDDICT file. As described below, it has been expanded considerably since then, both in terms of number of entries and also in multi-lingual coverage.

3 Project Status

The JMdict file was first released in 1999, and updated versions are released 3-4 times each year along with versions of the EDICT file, which is generated at the same time from the same data files. The file now has over 99,300 entries, i.e. the size of a medium-large printed dictionary, and the growth in numbers of entries is now relatively slow, with most updates dealing with corrections and expansion of existing entries.

The file is available under a liberal licence that allows its use for almost any purpose without fee. The only requirement is that its use be fully acknowledged and that any files developed from it continue under the same licence conditions.

4 Structure

The JMdict XML structure contains one element type: **<entry>,** which in turn contains sequence number, kanji word, kana word, information and translation elements. The sequence number is used for maintenance and identification.

The kanji word and kana word elements contain the two forms of the Japanese headwords; the former is used for representations containing at least one kanji character, while the latter is for representations in kana alone. The kana word is effectively the pronunciation, but is also an important key for indexing the dictionary file, as Japanese dictionaries are usually ordered by kana words. The minimum content of these fields is a single word in the kana word element. In addition, each entry may contain information about the words (unusual orthographical variant, archaic kanji, etc.) and frequency of use information. The latter needs to be associated with the actual words rather than the entry as a whole because some combinations of kanji and kana words are used more frequently than others. (For example, 合気道 and 合氣道 are orthographical variants of the one word *(aikidô)*, but the former is more common.)

The kana used in the elements follows modern Japanese orthography, i.e. *hiragana* is used for native Japanese words, and *katakana* for loan words, onomatopoeic words, etc.

In most cases an entry has just one kanji and one kana word (approx. 75%), or one kana word alone (15%). In about 10% of entries there are multiple words in one of the elements. In some cases a kana reading can only be associated with a subset of the kanji words in the entry. For example, *soyokaze* (そよかぜ: breeze) can be written either 微風 or そよ風 (the latter is more common as そよ is a non-standard reading of the 微 kanji). However 微風 can also be pronounced *bifuu* (びふう) with the same meaning, but clearly this pronunciation cannot be associated with the そよ風 form, as the kana portion is read *"soyo"*. XML does not provide an elegant method for indicating a restricted mapping between portions of two elements, so when such a restriction is required, additional tags are used with each kana word supplying the kanji word with which it may be validly associated.

The information element contains general information about the Japanese word or the entry as a whole. The contents allow for ISO-639 source language codes (for loan words), dialect codes, etymology, bibliographic information and update details.

The translation area consists of one or more sense elements that contain at a minimum a single gloss. Associated with each sense is a set of elements containing part of speech, cross-reference, synonym/antonym, usage, etc. information. Also associated with the sense may be restriction codes tying the sense to a subset of the Japanese words. For example, 水気 can be pronounced *suiki* (すいき) and *mizuge* (みずけ); both meaning "moisture", but the former alone can also mean "dropsy".

The gloss element has an attribute stating the target language of the translation. In its absence it is assumed the gloss is in English. There is also an attribute stating the gender, if for example, the part-of-speech is a noun and the gloss is in a language with gendered nouns. Figure 1 shows a slightly simplified example of an entry. The **<ke_pri>** and **<re_pri>** elements indicate the word is a member of a particular set of common words.

```
<entry>
<ent_seq>1206730</ent_seq>
<k_ele>
<keb>学校</keb>
<ke_pri>ichi1</ke_pri>
</k_ele>
<r_ele>
<reb>がっこう</reb>
<re_pri>ichi1</re_pri>
</r_ele>
<sense>
<pos>&n;</pos>
<gloss>school</gloss>
<gloss g_lang="nl" g_gend="fg">school</gloss>
<gloss g_lang="fr" g_gend="fg">école</gloss>
<gloss g_lang="ru" g_gend="fg">школа</gloss>
<gloss g_lang="de" g_gend="fg">Schule</gloss>
<gloss g_lang="de"
g_gend="fg">Lehranstalt</gloss>
</sense>
</entry>
```

Fig. 1: Example JMdict entry

The potential to have multiple kanji and kana words within an entry brings attention to the issues of homonymy, homography and polysemy, and the policies for handling these, in particular the criteria for combining kanji and kana words into a single entry. As Japanese has a comparatively limited set of phonemes there are a large number of homophonous words. For example, over twenty different words have the kana representation こうじょう *(kôjô).* If we regard homography as only applying to words written wholly or partly with kanji, there are relatively few cases of it, however they do exist, e.g. 川柳 when read せんりゅう *(senryû)* means a comic poem, but when read かわやなぎ *(kawayanagi)* means a variety of willow tree.

The combining rule that has been applied in the compilation of the JMdict file is as follows:

a. treat each basic entry as a triplet consisting of: kanji representation, matching kana representation, senses;

b. if for any basic entries two or more members of the triplet are the same, combine them into the one entry;

i. if the entries differ in kanji or kana representation, include these as alternative forms;

ii. if the entries differ in sense, treat as a case of polysemy;

c. in other cases leave the entries separate.

This rule has been applied successfully in a majority of cases. The main problems arise where the meanings are similar or related, as in the case of the entries: (放す, はなす, to separate; to set free; to turn loose) and (離す, はなす, to part; to divide; to separate), where the kana words are the same and the meanings overlap. Japanese dictionaries are divided on 放す and 離す; some keeping them as separate entries, and others having them as the one entry with two main senses. (The two words derive from a common source.)

5 Parts of Speech and Related Issues

As languages differ in their parts of speech (POS), the recording of those details in bilingual dictionaries can be a problem (Al-Kasimi, 1977). Traditionally bilingual dictionaries involving Japanese avoid recording any POS information, leaving it to the user to deduce that information from the translation and examples (if any). In the early stages of the EDICT project, POS information was deliberately kept to a minimum, e.g. indicating where a verb was transitive or intransitive when this was not apparent from the translation, mainly to conserve storage space. As there are a number of advantages in having POS information marked in an electronic dictionary file, a POS element was included in the JMdict structure, and publicly available POS classifications were used to populate much of the file. About 30% of entries remain to be classified; mostly nouns or short noun phrases.

In the interests of saving space an early decision had been made to avoid listing derived forms of words. For example, the Japanese adjective 高い *(takai)* meaning "high, tall, expensive" has derived forms of 高さ *(takasa)* "height" and 高く *(takaku)* "highly". As this process is very regular, many Japanese dictionaries do not carry entries for the derived forms, and some bilingual dictionaries follow suit. Another such example is the common verb form, sometimes called a "verbal noun", which is created by adding the verb する *(suru)* "to do" to appropriate nouns. The verb "to study" is 勉強する *(benkyôsuru)* where 勉強 is a noun meaning "study" in this context. Again, Japanese dictionaries often do not include these forms as headwords, preferring to indicate in the body of an entry that the formation is possible.

The omission of such derived forms means that care needs to be taken when constructing the translations so that the user is readily able to identify the appropriate translation of one of the derived forms.

In a multilingual context, the omission of derived forms can have other problems. The recording of する verbs only in their noun base form has been reported to lead to some discomfort among German users, as German language orthographical convention capitalizes the first letters of nouns but not verbs (the WaDokuJT file has する verbs as separate entries for this reason).

6 Inclusion and Maintenance of Multiple Languages

As mentioned above, part of the interest in having entries with translations in a range of languages came from the compilation of a number of dictionary files based on or similar to the EDICT file. There are a number of issues associated with the inclusion of material from other dictionary files, in particular those relating to the compilation policies: coverage, handling of inflected forms, etc. (Breen, 2002) There is also the major issue of the editing and maintenance of the material, which has the potential to become more complex as each language is incorporated.

The approach taken with JMdict has been to:

a. maintain a core Japanese-English file with a well-documented structure and set of inclusion and editing policies;

b. encourage the development and maintenance of equivalent files in other languages paired with Japanese, which can draw on the JMdict/EDICT material as required;

c. periodically build the complete multi-lingual JMdict from the different components.

This approach has proved successful in that it has separated the compilation of the file from the ongoing editing of the components, and has left the latter in the hands of those with the skills and motivation to perform the task.

At the time of writing, the JMdict file has over 99,300 entries (Japanese and English), of which 83,500 have German translations, 58,000 have French translations, 4,800 have Russian translations and 530 have Dutch translations. A set of approximately 4,500 Spanish translations is being prepared, with the prospects that some 20,000 will be available shortly.

The major sources of these additional translations are:

a. French translations from two projects:

i. approximately 17,500 entries have come from the Dictionnaire français-japonais Project (Desperrier, 2002), a project to translate the most common Japanese words from the EDICT File into French;

ii. a further 40,500 entries drawn from the 仏語補完計画 (French-Japanese Complementation Project) at http://francais.sourceforge.jp/ (This project is also based on the EDICT file.)

b. German translations from the WaDokuJT Project (Apel, 2002). This is a large file of over 300,000 entries; however, unlike JMdict it includes many phrases, proper nouns and inflected forms of verbs, etc. The overlap of coverage with JMdict is quite high, leading to the large number of entries that have been included in the JMdict file.

One of the issues that can lead to problems when incorporating translations from other project files is that of aligning the translations when an entry has several senses. In the case of the French translations, the project coordinator has marked the translations of polysemous entries with a sense code, thus enabling the translations to be inserted correctly when compiling the final file. For other languages, the translations are being appended to the set English translations. The appropriate handling of multiple senses is an item of future work.

7 Examples of Word Usage

When the project was begun and the DTD designed, it was intended that sets of bilingual examples of usage of the entry words would be included. For this reason an <example> element was associated with each sense to allow for such example phrases, sentences, etc, to be included.

In practice, a quite different approach has been taken. With the availability since 2001 of a large corpus of parallel Japanese/English sentences (Tanaka, 2001), it was decided to keep the corpus intact, and instead provide for the association of selected sentences from the corpus with dictionary entries via dictionary application software (Breen, 2003b). This strategy, which required the corpus to be parsed to extract a set of index words for each sentence, has proved effective at the application level. It also has the advantage of decoupling the maintenance of the dictionary file from that of the example corpus.

8 Related Projects

Apart from a few small word lists involving several European languages, the only other major current project attempting to compile a comprehensive multilingual database is the Papillon project (e.g. Boitet et al, 2002). See http://www.papillon-dictionary.org/ for a full list of publications. The Papillon design involves linkages based on word-senses as proposed in (Sérasset, 1994) with the finer lexical structure based on Meaning-Text

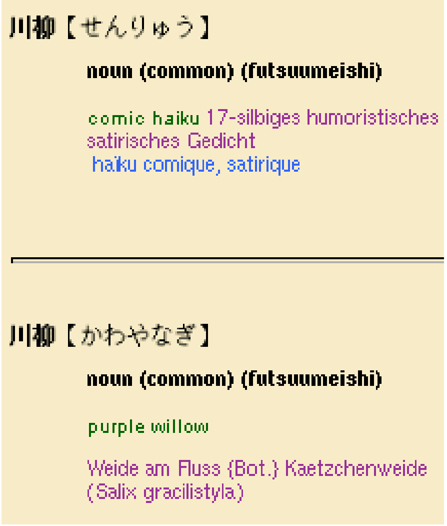
Theory (MTT) (Mel'cuk, 1984-1996). At the time of writing the Papillon database is still in the process of being populated with lexical information.

Closely related to the JMdict project is the Japanese-Multilingual Named Entity Dictionary (JMnedict) project. This is a database of some 400,000 Japanese place and person names, and non-Japanese names in their Japanese orthographical form, along with a romanized transcription of the Japanese (Breen, 2004b). Some geographical names have English descriptions: cape, island, etc. which are in the process of being extended to other languages. The JMnedict file is in an XML format with a similar structure to JMdict.

Another multilingual lexical database is KANJIDIC2 (Breen, 2004c), which contains a wide range of information about the 13,039 kanji in the JIS X 0208, JIS X 0212 and JIS X 0213 character standards. Among the information for each kanji are the set of readings in Japanese, Chinese and Korean, and the broad meanings of each kanji in English, German and Spanish. A set of Portuguese meanings is being prepared. The database is in an XML format.

9 Applications

While there are a number of experimental systems using the JMdict file, the only application system using the full multilingual file at present is the Papillon project server. Figure 2 shows the display from that server when looking up the word 川柳. The author's WWWJDIC server (Breen, 2003a) uses the Japanese-English components of the file. Figure 3 is an extract from the WWWJDIC display for the word 小人, which is an example of an entry with multiple kana words, and senses restricted by reading. (The (P) markers indicate the more common readings.)



川柳【せんりゅう】

noun (common) (futsuumeishi)

comic haiku 17-silbiges humoristisches satirisches Gedicht
haïku comique, satirique

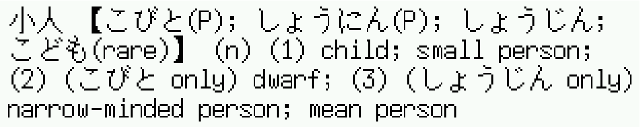
川柳【かわやなぎ】

noun (common) (futsuumeishi)

purple willow

Weide am Fluss {Bot.} Kaetzchenweide (Salix gracilistyla)

Fig. 2: Papillon example for 川柳



小人 【こびと(P); しょうにん(P); しょうじん; こども(rare)】 (n) (1) child; small person; (2) (こびと only) dwarf; (3) (しょうじん only) narrow-minded person; mean person

Fig. 3: WWWJDIC example for 小人

The EDICT Japanese-English dictionary file, which is generated from the same database as the JMdict file, continues to be a major non-commercial Japanese-English lexical resource, and is used in a large number of applications and servers, as well as in a number of research projects.

10 Conclusion

The JMdict project has successfully developed a multilingual lexical database using Japanese as the pivot language. In doing so, it has reached a lexical coverage comparable to medium-large printed dictionaries, and its components are used in a wide range of applications and research projects. It has also demonstrated the potential for re-use of material from related and cooperating lexicon projects. The files of the JMdict project are readily available for use by researchers and developers, and have the potential to be a significant lexical resource in a multilingual context.

References

Al-Kasami, A.M. 1977 *Linguistics and Bilingual Dictionaries,* E.J. Brill, Leiden

Apel, U. 2002. *WaDokuJT - A Japanese-German Dictionary Database,* Papillon 2002 Seminar, NII, Tokyo

Boitet, C, Mangeot-Lerebours, M, Sérasset, G. 2002 *The PAPILLON project: cooperatively building a multilingual lexical data-base to derive open source dictionaries & lexicons,* Proc. of the 2nd Workshop NLPXML 2002, Post COLING 2002 Workshop, Ed. Wilcock, Ide & Romary, Taipei, Taiwan.

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

Breen, J.W. 2002. *Practical Issues and Problems in Building a Multilingual Lexicon,* Papillon 2002 Seminar, NII, Tokyo.

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

Breen, J.W. 2003b. *Word Usage Examples in an Electronic Dictionary,* Papillon 2003 Seminar, Sapporo.

Breen, J.W. 2004a. *The EDICT Project,* http://www.csse.monash.edu.au/~jwb/edict.html

Breen, J.W. 2004b. *The ENAMDICT/JMnedict Project,* http://www.csse.monash.edu.au/~jwb/enamdict_doc.html

Breen, J.W. 2004c. *The KANJIDIC2 Project,* http://www.csse.monash.edu.au/~jwb/kanjidic2/

Desperrier, J-M. 2002. *Analysis of the results of a collaborative project for the creation of a Japanese-French dictionary,* Papillon 2002 Seminar, NII, Tokyo.

Mel'cuk, I, et al. 1984-1996. *DEC: dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques,* Vols I-IV, Montreal Univ. Press.

Sérasset, G. 1994. *SUBLIM: un Système Universel de Bases Lexicales Multilingues et NADIA: sa spécialisation aux bases lexicales interlingues par acceptions,* (Doctoral Thesis) Joseph Fourier University, Grenoble

Sperberg-McQueen, C.M.. and Burnard, L. (eds.) 1999. *Guidelines for Electronic Text Encoding and Interchange.* Oxford Univ. Press.

Tanaka, Y. 2001. *Compilation of a Multilingual Parallel Corpus* PACLING 2001, Japan.