JMdictDB - Online Development and Maintenance System for the Japanese-Multilingual Dictionary

James Breen
Monash University, Australia

Stuart McGraw

Abstract

This paper describes an online database system for maintaining the dictionary files in the JMdict (Japanese-Multilingual Dictionary) project (Breen: 2004). The system and database have been designed with the goals of (a) fully supporting the complex microstructure of Japanese dictionary entries; (b) providing a relatively simple and flexible user interface; and (c) enabling open access to dictionary users and contributors.

Keywords: database; dictionary; Japanese; maintenance; on-line

1. Introduction

The JMdict/EDICT Japanese-English dictionary is a popular and widely-used resource which has been compiled by a team of volunteers over the last two decades. The relatively simple EDICT plain-text version was first released in 1991, and the XML-format JMdict with support for a much more complex and complete microstructure was first released in 1999. At present the dictionary contains nearly 170,000 entries.

During the initial period of the development of the dictionary, it was maintained as a set of text files using a markup system to support the internal structure and to enable generation of the full XML releases. A simple WWW-based submission system for new entries and amendments was used, however all edits had to be made centrally and mostly by hand (by the first author), which meant there was a bottleneck in processing changes, there was little scope for using multiple editors, and there was limited opportunity for discussing aspects of the entries among contributors.

The issues were widely discussed in email forums associated with the project, and it was recognized that although there were several systems available for editing dictionary entries which were being used successfully by other projects, they were not capable of supporting the relative complexity of some of the entries in the JMdict dictionary (this is discussed below.) A key requirement was to find a balance between an interface which could be used with minimal training and experience to submit straightforward edit and new entries, and one which could also handle the complex edge cases.

After some prototyping of alternatives, a system was designed and developed (largely by the second author) which has successfully met both goals. The system has been developed in the Python language, uses PostgreSQL databases, and operates under an Apache WWW server.

2. Issues in Japanese dictionary entry structures

While many issues associated with the internal structures of entries in electronic dictionaries are common to all languages (part-of-speech tagging, sense tags, cross-referencing, etc.), the nature of Japanese orthography and morphology adds complexities that are not usually found in many languages. Japanese is written in a mixture of four scripts:

a. *kanji* (Chinese characters), of which approximately 2,000 are in regular use and considerably more are used occasionally;
b. *hiragana*, a Japanese syllabary of 46 basic characters plus diacritics. In modern Japanese it is used primarily for particles, conjunctions, pronouns, inflections, etc. although it is possible to use it for virtually everything, and is often used in the place of less-used kanji;
c. *katakana,* another syllabary which in modern Japanese is used primarily for loanwords, transcription of foreign names, scientific names of flora and fauna, etc.

d. alphanumerics. These are mainly used for acronyms, etc. and for words such as "iPhone" which are not transliterated. (It is possible to transcribe Japanese into alphabetics, resulting in ローマ字 (*romaji*), however this is used little in Japan apart from transcribing place-names, station-names, etc. for benefit of foreigners. It is also used in the early stages of teaching Japanese as a second language.)

Japanese words and phrases which typically make up the body of lexical items/headwords can and do often have a variety of surface forms. A lexical item in a Japanese dictionary typically has two parts:

a. a *kanji* part which contains the form of the item which contains one or more kanji;

b. a *yomikata* or reading part which has the pronunciation of the term(s) in the *kanji* part. This part is required because in Japanese *kanji* can have several alternative pronunciations or readings. Some lexical items will only have this part when there is no *kanji* form, such as with loanwords or particles.

The complexity arises because in many cases there is no canonical form either of the kanji form(s) or of the reading. For example:

a. alternative *kanji* can be used for the same meaning and pronunciation. The common verb かける *(kakeru)* can be written 掛ける or 懸ける.

b. the use of *okurigana* (writing part of a word in *hiragana*) can vary. The word 生け花 (*ikebana* - flower arrangement) is sometime written 生花. (The surface form 生花 has another pronunciation - *seika* - which also mean "flower arrangement", but has an additional meaning: "fresh flowers".)

c. *hiragana* can be substituted for *kanji* in some cases, e.g. the word 陥穽 (かんせい, *kansei,* trap) is often written 陥せい, as the 穽 *kanji* is not common;

d. a *kanji* form can sometimes have several pronunciations, e.g. the term 遺言書 (will; testament) can be pronounced both ゆいごんしょ (*yuigonsho*) or いごんしょ (*igonsho*).

e. on occasions where there are several *kanji* forms and several pronunciations, some pronunciations do not apply to all the kanji forms;

f. sometimes where there are multiple surface forms in a polysemous entry, different surface forms may be more associated with one sense than another, or may even be restricted to a subset of the senses. Similarly, differing pronunciations may be associated more with one sense than another or restricted to a subset of senses.

The JMdict XML design has structures to enable complex entries to be recorded in a consistent manner. For example the 生け花 entry has, in part:

```
<k_ele><keb>生け花</keb></k_ele>
<k_ele><keb>生花</keb></k_ele>
<r_ele><reb>いけばな</reb></r_ele>
<r_ele><reb>せいか</reb><re_restr>生花</re_restr></r_ele>
```

indicating that there are two *kanji* forms and two readings but the せいか reading is restricted to the 生花 *kanji* form.

It must be emphasized that the sorts of complexities described above only apply to a minority of dictionary entries; most have a single *kanji* part and reading.

3. The user interface

As mentioned above, the design of the JMdictDB user interface had to meet the twin requirements of being suitable for untrained users entering or editing straightforward entries, while having the capability of allowing more complex structures to be described and edited efficiently. Various alternatives were considered including having drop-down menus of the options for the various elements, however it was concluded that these led to an interface that was over-complex.

The approach that has been adopted is to provide users with free-form text areas for the three key parts of an entry: Kanji, Reading and Meaning, and to use a relatively simple embedded markup language which we have called JEL (JMdict Entry Language) to record the contents of each part. To illustrate this, the relatively simple entry:

関頭 (かんとう) n, critical moment; crucial point; crossroads

is recorded:

Kanji: 関頭
Reading: かんとう
Meaning: [1][n] critical moment; crucial point; crossroads

The JEL components are always marked by square brackets, and in the entry above they simply indicate the sense number (there is only one) and the part-of-speech.

The general form of the JEL components is "[type=value]", however where the value code is unique, the "type=" may be omitted. Thus "[pos=n]" can be simplified to just "[n]", as in the above example. There are two sets of tags available for the Kanji field (frequency, information), three for the Reading field ((frequency, information, restriction) and eight for the Meaning field (POS, domain, miscellaneous, restriction, source-language, cross-reference, etc.)

The flexibility of the JEL approach can be seen in the following example of a relatively complex entry:

Kanji: 鬼[ichi1,news1,nf08]
Reading: おに[ichi1,news1,nf08]；き
Meaning: [1][n] ogre; demon
[2][n] spirit of a deceased person [see=1518610・亡魂[1]]
[3][n] [restr=おに] ogre-like person (i.e. fierce, relentless, merciless, etc.)
[4][n] [restr=おに] it (i.e. in a game of tag)
[5][n][fld=astron] [restr=き] Chinese "ghost" constellation (one of the 28 mansions)
[see=2176790・二十八宿[1]] [see=2176850・朱雀・すざく[2]]
[6][pref][sl] [restr=おに] very; extremely; super- [see=1429340・超[1]]

This entry appears in the display from the WWWJDIC server as:

鬼 【おに(P); き】 (n) (1) ogre; demon; (2) (See 亡魂) spirit of a deceased person; (3) (おに only) ogre-like person (i.e. fierce, relentless, merciless, etc.); (4) (おに only) it (i.e. in a game of tag); (5) (き only) {astron} (See 二十八宿, 朱雀・すざく・2) Chinese "ghost" constellation (one of the 28 mansions); (pref) (6) (おに only) (sl) (See 超・1) very; extremely; super-

The text in each of the fields is parsed at the time an edit is submitted and the values loaded into the appropriate database tables. When an entry is edited, the database tables are converted back into the JEL and text format.

In addition to the Kanji, Reading and Meaning fields, the interface also has Reference and Comment fields. Users are requested to provide appropriate references for submissions and edits, such as supporting information from other dictionaries and examples of usage in context. Submitters and editors may make comments about the edit as it proceeds. Both these fields are retained in the database along with the details of all edits made, and provide both an audit trail and a record of any discussion between submitters and editors.

4. Submission and editorial process.

In order to encourage submissions to the dictionary project, both new entries and amendments, it was decided to make the process as "open" as possible. To enable this a two-level submission/approval process has been implemented. Anyone accessing the dictionary maintenance server may submit new entries or amendments. There is no requirement for prior registration or approval.

Submissions are placed in the database tables associated with each entry as provisional updates, where they are visible along with any previous edits to the entry. Further edits may be made, resulting in a chain of interim updates, before a final decision is made.

Approval of a new entry can only be made by a registered editor. Upon approval the updates are committed to the database along with the record of changes, comments, etc., and the edited entry will be included the next data distribution.

5. JMdictDB in operation

The JMdictDB system has been fully operational since mid-2010, and currently handling an average of 20 new entries and 50 entry edits per day. It has demonstrated that it is meeting its goals of allowing a wide range of users to propose entries and amendments, while at the same time enabling the creation of more complex entries.

Access to the database can either be via its own powerful search facility which enables entries to be located by a variety methods, including entry contents, date; author, etc., or by links from other systems, such as the WWWJDIC dictionary server (Breen: 2009). Developers of online systems using the dictionary are encouraged to include such links.

The contents of the database can be downloaded in several formats, and the main one current used is the JMdict XML format, in which it is downloaded daily and from which other distribution formats are generated.

At present there are six registered editors and about fifty regular submitters of entries and amendments. One aspect of its operation that was unexpected is the use by submitters and editors of the accumulated set of comments as a form of "blog" associated with an entry. On occasions there are quite lengthy and lively dialogues about a proposed edits.

6. Future developments

At present the database and interface only handles the English translations for JMdict. Translations of the entries in other language, e.g. French, German, Dutch, etc. are maintained elsewhere and added at the sense level prior to distribution. Extensions are being planned to enable multiple language glosses to be supported, and to enable users to view either all languages or just a selection. Internationalization of the interface is also being examined.

The system has been designed to support other dictionaries, in particular the companion named-entity dictionary, JMnedict, which has approximately 740,000 entries and is currently maintained as a text file. It is hoped that maintenance of JMnedict can be brought into the JMdictDB system during 2013.

7. Conclusion

The JMdictDB system has been designed and developed to support an open Japanese dictionary maintenance operation in an online environment. It has demonstrated the effectiveness of using a simple embedded markup language to support a range of entry types from the simple up to the very complex. The system is currently supporting the 170,000-entry JMdict dictionary, and is planned to expand to other dictionaries and languages.

References

Breen, J. 2004. JMdict: a Japanese-Multilingual Dictionary, *COLING Multilingual Linguistic Resources Workshop,* Geneva, 71-78.

Breen, J. 2009. WWWJDIC - a feature-rich WWW-based Japanese dictionary, *eLexicography in the 21st century: New Challenges, New Applications,* Louvain-la-Neuve, Belgium, 2010, 381-386.