IJET-20

Compiling a Freeware Japanese-English Dictionary

Jim Breen

(Feb 2009)

Outline of Presentation

- Background – how it started
- The early days
- Structure and complexity
- Current system: status, size, etc.
- Issues: scope/inclusion, licensing, copyright
- Growth plans
- Future approach for maintenance
- Would I do it again?

Background

- Had wanted to do some “Japanese” software for years
- ~1991: MOKE and KD showed it was possible, and not too difficult
- It seemed like writing a dictionary program would be fun
- MOKE 2.0 had a simple dictionary
- First efforts: JDIC, JREADER [DOS]

The EDICT Dictionary

Very simple text file:

- 漢字 [かんじ] /kanji/
- カナだけ /only kana/
- EUC-JP coding
- Initially tiny – added a few thousand words
- Released along with JDIC via USENET and called for contributions

The EDICT Dictionary (2)

- Initially a small flood donated vocabulary lists, etc.
- Included proper names as well at first
- By ~1998 had ~50,000 entries plus ~120,000 names
- Names were split off to keep it fitting on a diskette (144kb)

Dictionary Software

JDIC: dictionary search

- Enter key in kanji/kana/English
- Integrated kanji dictionary

JREADER: text reading assistant

- Split screen: text / glosses
- Step through text, select words to gloss

All released as freeware

Dictionary Software (2)

Code recycled into other packages:

- *xjdic* – Unix X11 (GPL source)
- MacJDic (Mac OS)
- WWWJDIC (server code)

Many other EDICT-using packages have been written

Growing Pains

- Simple structure a major limitation
  - Orthographic variants: kanji, okurigana, etc.
  - Richer meta-information: PoS, senses, etc.
  - Linking to examples, other dictionaries, etc.
  - Able to be parsed correctly
- Began to search for a better structure
  - Retain back-compatibility (legacy software)

1999 – New Approach

- Decided to move to an “internal” database for maintenance and to generate multiple distributions formats
- Major distribution format would be XML
- Various bilingual models examined: TEI, Shoebox, etc. No suitable standard
- Cater for multilingual glosses

One Dictionary – Multiple Formats

Current distributions:

- JMdict (XML) – the whole file, along with German, French, etc. glosses
- edict_mac (XML) – for the Mac "D ic" app.
- EDICT – traditional format
- EDICTSUB – 20,000 "common"s ubset
- EDICT2 – extended EDICT (used in WWWJDIC)

Format Example

EDICT

華々しい [はなばなしい] /(adj-i) brilliant/magnificent/spectacular/(P)/

花々しい [はなばなしい] /(adj-i) brilliant/magnificent/spectacular/

華華しい [はなばなしい] /..../

花花しい [はなばなしい] /..../

Format Example (2)

EDICT2

華々しい (P); 花々しい ; 花花しい ; 華華しい

[はなばなしい] /(adj-i)
brilliant/magnificent/spectacular/(P)/

Format Example (3)

JMDict

<entry>

<ent_seq>1600960</ent_seq>

<k_ele>

<keb> 華々しい </keb>

<ke_pri>ichi1</ke_pri>

</k_ele> .....

Format Example (4)

JMDict (cont.)

<r_ele>

<reb> はなばなしい </reb>

<re_pri>ichi1</re_pri>

</r_ele>

Format Example (5)

JMDict (cont.)

<sense>

<pos>&adj-i;</pos>

<gloss>brilliant</gloss>....

<gloss xml:lang="fre">glorieux</gloss> ...

<gloss xml:lang="ger">prächtig</gloss>....

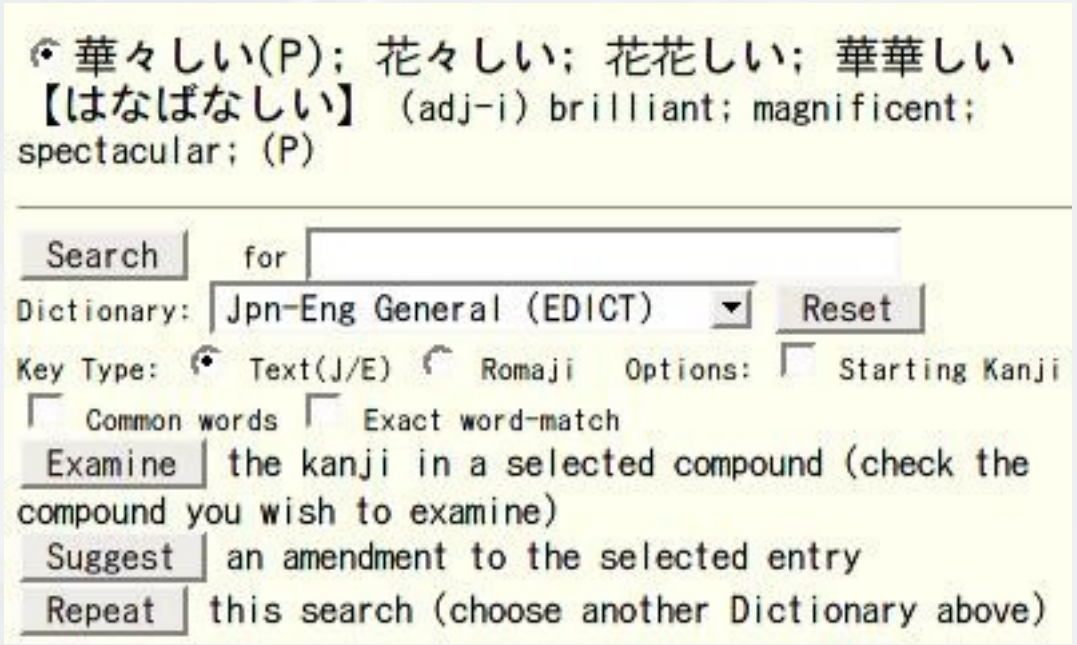
</sense>

</entry>

Current Dictionary

- 137,200 entries (161k headwords)
- 728,000 names (597k unique kanji forms)
- WWW-based amendment/new entry forms
  - Linked from WWWJDIC (others possible)
  - Discussion forum
  - Manual edits, semi-automatic insertion of new entries
- Daily generation and distribution

WWW JDIC Link



WWW JDIC Edit Page

Headword (kanji) 1	華々しい
Headword (kanji) 2	花々しい
Headword (kanji) 3	花花しい
Headword (kanji) 4	華華しい
Headword (kanji) 5	
Reading 1	はなばなしい
Reading 2	
Part-of-speech	adj-i
Translation (English, etc.) 1	brilliant
Translation (English, etc.) 2	magnificent
Translation	

Inclusion Issues

How big should it be? (reverse searches)

- Very uncommon words/expressions
- Long multi-word expressions
- Long expressions
- Inflected forms
- Classical forms, archaisms, etc.

(Google hits often the criterion, e.g. > 20)

Licensing Issues

- Early contributors wanted it free and unexploited
- Initial licence required “no commercial use”
- Later extended to commercial licences
- Changed to unrestricted use: acknowledgement; donations
- Now a “Creative Commons: Share-Alike” licence

Copyright Issues

- Very difficult issue; little relevant case law
- Varies by country: US is relaxed; EU strict
- Abstract information/concepts cannot be copyrighted, only form, wording, selection.
- Started cautiously with EDICT (Nelson problem, etc.)
- Now more relaxed (large, unique format, diluted)

Growth Plans

Polishing:

- Many early entries are too simple and context-specific

Expansion:

- Mostly compound nouns, multiword expressions, neologisms
- Huge potential to use existing sources
- Editors needed: verifying, rewording, etc.

Future Approach

- Online database (mid-2009?)
  - entries linked from WWWJDIC, etc.
  - Wiki-like environment
    - Anyone can enter/edit entries
    - Panel of editors: approve/amend/reject
  - entry complexity is a problem (using a simple markup language)
- Daily file distribution:
  - DB -> XML -> EDICT, etc. (via XSLT)

New Online Edit

Would I Do It Again?

- Probably not
  - At the time it started there was nothing
  - Now there are many good online resources
    - Commercial dictionaries (EXCEED, Daijirin, etc.)
    - Eijiro
    - Search engines
- Will it continue?
  - I certainly hope so
  - It needs a support community to thrive