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

James Breen
Monash University

Stuart McGraw
Quick Overview of Japanese Orthography

Japanese is written in a mixture of scripts: 
*kanji* (Chinese characters), used mainly for nouns and roots of verbs, adjectives, etc. Approx. 2,000 in common use.
most nouns in *kanji* use 2 or more characters, verbs typically use one *kanji*
the *hiragana* syllabary (46 symbols plus diacritics: あいうえおかきくけこ, etc.), used mainly for particles, inflections, conjunctives, etc.
Japanese is written in a mixture of scripts (cont...):
the *katakana* syllabary (アイウエオカキクケコ, etc.) used for loanwords, foreign names, scientific names, etc.
Latin alphabetics - in text mainly used for initials, acronyms, etc (*USB, bps, etc.*) or product names (*iPhone, Windows, etc.*)
e.g. デパートで旅行鞄と iPhone を買いました。
Open-Source Japanese-English Dictionary (other languages can be included)
used in many servers, apps, etc.
widely used in NLP research
Began in 1991 as EDICT (Electronic DICTIONary)
simple format, text file, initial DOS program, etc.
In 1999 migrated to XML format
richer structure (now about 80k entries)
legacy EDICT format maintained
In 2003 began using online forms for submitting new entries and amendments
still in text files, single editor
In 2008/2009 an online database and interface were developed
2010 cutover to the new system
runs on a cloud server (Linux)
entry-level linking from client systems (e.g. WWWJDIC)
daily generation of distributions. (now 170k entries)
The complexity of Japanese dictionary entries
multiple surface forms, e.g. 思い出す, 思いだす, 思出す, おもい出す
multiple pronunciations/readings
含嗽 (gargling) pronounced both うがい (ugai) and がんそう (gansou)
not all readings apply to all surface forms
うがい can also be written 嗽 and 漱, and がんそう can also be written 含漱
in polysemous terms, sometimes senses are limited to certain surface forms and/or readings
眼鏡 is read めがね (megane) or がんきょう (gankyou) and means “glasses/spectacles”
the めがね/megane reading also means “judgement/discernment”
Interface Design Requirements

We wanted a User Interface which would:
- enable anyone to propose a basic new entry or correction;
- enable a skilled user to handle all the entry structure complexity

Needed to handle workflow and record-keeping:
- partial edits (awaiting approval)
- approval of new entries/amendments
- change logging
- references, comments, discussion
Design Decisions

Opted for a simple text-based interface
Five text panels:
  - kanji part
  - reading(s) part
  - meaning(s) part
  - references
  - comments

Microstructure described by a simple language (JEL: JMdict Entry Language)
Design Decisions (2)

Two levels of user:

editor (account, login, can accept/reject changes, delete entries)

general (no login, can propose new entries and amendments)

Permanent record and complete visibility of all changes, comments, etc.
Where To From Here?

Generally well-accepted in the user community has spread the editorial/approval load still a bit daunting to unpractised users
Upgrading from Python 2 to Python 3 (in testing)
Planned enhancements:
  Incorporating non-English glosses
currently maintained separately and imported
selectable language views, e.g. Japanese/English/French
Extending to other dictionaries
  JMnedict (named entities, approx. 740,000 entries)
  Kanjidic (database of about 15,000 kanji)
Multi-language User Interface (Japanese, French, etc.)