Japanese Dictionaries and Multiple Surface Forms: Issues and Solutions

Jim Breen
jimbreen@gmail.com

September 8, 2024

Japanese Dictionaries and Multiple Surface Forms

Why are we discussing this issue?

- Japanese has a very flexible orthography, leading to (potentially) many "surface forms" for terms
  - choice of *kanji* (often two or more are available)
  - choice of using *kana* for all or part of a term
    - めぐり合う or 巡り会う or 巡り合う (めぐりあう)
  - variations in *okurigana* usage
    - 引き換え or 引換え or 引換 (ひきかえ)
  - variations in the transcription of loanwords into *katakana*
    - ダイアモンド or ダイヤモンド
- A particular challenge for JSL learners
  - don't have a background in written Japanese
  - are currently exposed to significant amounts of "in the wild" Japanese text
    - WWW pages
    - social media

Japanese Dictionaries and Multiple Surface Forms (2)

- Challenges for lexicographers compiling dictionaries for JSL use (decoding or passive dictionaries)
  - which forms to include?
  - how multiple forms should be presented?
    - ordering
    - indication of status
    - visible or search-only
- Challenges when providing lexicons for NLP/text-processing, e.g. glossing systems; translation support, etc.

Japanese Dictionaries and Multiple Surface Forms (3)

A messy example of variation between dictionaries - the verb てこずる *(tekozuru)* "to have much trouble; to have a hard time"

- てこずる - several Japanese-English dictionaries have only this (48% of usage)
- 手こずる - not in paper dictionaries (48% of usage)
- 梃子摺る - 実用日本語表現辞典 (3%)
- 手古摺る - 新和英大辞典第5版 (<1%)
- 梃摺る - 広辞苑, 大辞泉, 日本国語大辞典 (<1%)
- 手子摺る - 大辞林, 大辞泉, 日本国語大辞典 (<1%)

Context Of This Discussion

The main context of this discussion/analysis is the JMdict (Japanese-Multilingual Dictionary) project.

- open source dictionary compilation, operating since the early 2000s;
- anyone can contribute, with an editorial team setting policies and vetting additions and changes;
- primary database (currently over 200,000 entries) handles Japanese-English - other language glosses are appended from related projects.

Surface Form Variation

Of the current 200,000 JMdict entries:

- approx. 160,000 entries have *kanji* forms
  - of these over 30,000 have multiple forms;
  - over 250 have 4 or more forms
  - over 20 have 8 or more forms, e.g. 磨りガラス, 擦りガラス. 磨り硝子, etc. (すりガラス)
- of the 40,000 *kana-only* entries:
  - about half have multiple surface forms
  - over 20 have 5 or more forms
  - several have 10 or more, e.g. ロイヤリティ, ロイヤルティ, ロイヤリティー, ロイヤルティー. etc.

Multiple surface form issues:

- which forms should be collected/recorded?
- which should be displayed, and in what order?
- what status details should be included?

Determination Of Surface Form Frequencies

Important to know which surface forms are in common use and their relative frequency

- - published dictionaries vary a lot and concentrate on "official" *kanji* versions
- search-engine counts are quite unreliable (Google insider advice)
- smaller corpora quite limited
  - the BCCWJ only has one of the kanji forms for てこずる (手子摺る)

A very major resource is the 2007 Google N-gram Corpus

- based the entire Japanese WWW content in mid-2007
- contains 1.3 billion counted text strings (terms, phrases, etc.)
- flexible on-line/batch lookup tool

Examples from the Google N-gram Corpus

Ranked with affixes:

- 学校 48633879
- 学校の 7319568
- 学校に 4398391
- 学校で 3872167
- 学校を 1881651
- 学校が 1507934

Verb inflections

- 食べる 19179416
- 食べます 1168041
- 食べない 1882110
- 食べぬ 1561
- 食べず 354272
- 食べません 281265
- 食べた 10032006

Issues of Choice of Kanji

Traditional vs Simplified

- since the 1940s reforms simplified *kanji* have dominated (学 vs 學, 気 vs 氣, etc.)
- computer *kanji* coding standards mainly used simplified forms
- Unicode has made more traditional forms available
- standards and government bodies trending back to traditional forms (鹼 instead of 鹸, 彎 instead of 弯)

Similar-looking *kanji*

- 柿 (persimmon) *(かき)* vs 柿 (wood shaving) *(こけら)*
- 気慨 vs 気概 *(きがい)* - often from keyboard errors

Mixed Kanji/Kana Forms (交ぜ書き)

Frequently seen in compound verbs:

- 絞り込む (to squeeze; to wring out) is sometimes written 絞りこむ or しぼり込む
- 嵌まり込む (to fit in) is more commonly written はまり込む

Compound nouns often have *kana* in place of rare *kanji:*

- 痤瘡 (acne) is more often written ざ瘡 (or just as ざそう)
- 握り寿司 is more often written にぎり寿司
- 発癌 (carcinogenesis) is often written 発ガン or 発がん - *katakana* is often used in medical terms

Loanword Transcription Issues

- A large number of loanwords used Japanese - most from English, but many from French, German, Dutch, etc.
- Usually transcribed in the *katakana* syllabary
- Considerable variation in transcription approaches:
  - handling vowels, e.g. ダイアグラム/ダイヤグラム for diagram, ダイアモンド/ダイヤモンド for diamond
  - handling consonants, e.g. バイオリン/ヴァイオリン for violin
  - lengthening vowels - the standard is " ー" but often イ or ウ is used, as in デイ/デー (day) and メイト/メート (mate) (the " ー" is often omitted)
  - many have combinations of variations - vibraphone is usually ビブラフォン/ヴィブラフォン but versions with ホン/ホーン/フォーン and バイブラ/ヴァイブラ are used
- published dictionaries tend to use "standard" forms.

Recording and Display Approaches

- In JMdict all forms in "reasonably common" use are recorded
  - based on corpora counts and proportions, and on references
  - case-by-case decision for less common forms
- Relatively uncommon forms are tagged as "search-only"
  - signal to dictionary sites and apps
  - should be used as a search key, need not be displayed
- Forms are usually recorded in descending frequency order
- Terms which usually written only in *kana* are tagged to enable sites/apps to indicate this.

Search-Only Form Example

Basic entry display for ベビーシッター (WWWJDIC)

ベビーシッター*[P]*; ベビー・シッター (n) babysitter; *[P]* [Links] [Sch]

うちの子供たちは、ベビーシッターに世話をしてもらった。 Our little children were taken care of by the babysitter.[Amend]

Optional display of search-only forms

ベビーシッター*[P]*; ベビー・シッター (n) babysitter; *[P]* [Links] [Sch]

Search-only Keys X
ベイビーシッター
ベビーシッタ
ベビシッター
ベービーシッター

Example of Entry Usually in Kana

Search Key: かご Current Dictionary: Jpn-Eng General (

かご*[P]*; カゴ*[P]* 《籠*[P]*; 篭*[P]*》 (n) (uk) basket (shopping, etc.); hamper; cage; *[P]* [Links] (Name entry: 篭)

そのかごはリンゴでいっぱいだった。 The basket was full of apples.

Indication of Status of Multiple Forms

It is important to indicate to users the relative status of forms. For visible forms, JMdict uses meta-information tags for this purpose:

- rarely-used forms. These are not hidden as they are in published dictionares, e.g. 併記 *(へいき)* (writing side by side) has the rare 並記 variant. [rK, rk]
- incorrect *kanji* forms, for example 年棒 instead of 年俸 *(ねんぽう)* (annual salary) - common enough to keep visible and warn users [iK].
- incorrect *kana* forms, for example お待ちどおさま is more commonly written お待ちどうさま, and アサインメント (assignment) is often written アサイメント. [ik]
- outdated forms, e.g. 雄辯 for 雄弁 *(ゆうべん)* (eloquence; fluency). [oK,ok]

Multiple Surface Forms - Summary

- wide range of surface forms in use in Japanese texts
- focus on Japanese-English decoding dictionaries (JSL, etc.)
- need to direct users to the "correct" entries
- emphasis on commonly-used forms rather than "standard" forms
- extensive use of corpora for determining term/form frequencies
- major use of "search-only" forms for less common versions
- tagging of status of multiple surface forms