

# Japanese Dictionaries and Multiple Surface Forms: Issues and Solutions

James Breen

Monash University, Australia (jimbreen@gmail.com)

## Abstract

This paper explores the issues associated with the wide variety of surface forms that can be used in Japanese as they apply to the compilation of Japanese dictionaries, especially decoding dictionaries aimed at language learners. These varieties arise from such things as: the choice of *kanji*, variations in *kana* usage, replacing *kanji* with its *kana* equivalent, and variations in transcription of loanwords.

The processes leading to the multiple surface forms are examined in detail, along with the identification of the levels of usage of the various forms in computer and web-based corpora.

The handling of multiple surface forms within the open-source JMdict (Japanese-Multilingual Dictionary) project is described, including the approaches to ordering and displaying the alternative forms, along with the inclusion of meta-information to indicate the status of the various forms.

**Keywords:** surface forms, decoding dictionaries, Japanese, *kanji*, *kana*

## 1. Introduction

A major challenge for compilers and maintainers of Japanese dictionaries, especially in the digital age, is the wide variety of surface forms that can be used in Japanese. The variety of surface forms presents a particular challenge to compilers of decoding dictionaries aimed at non-Japanese speakers, and also for lexicons supporting text analysis and glossing software.

Since the development and spread of Internet-based platforms, such as the WWW and social media, large quantities of text have been, in effect, published without the editorial controls that were employed to a degree in the past to keep language use and orthography reasonably regular. While native speakers of a language are usually able to associate less common and irregular forms with their "correct" versions, language learners and non-native speakers can find such forms found in the wild quite a challenge, especially if they cannot be readily located in published dictionaries.

The issue of variation in surface forms can be illustrated by the verb *tekozuru* (to have a hard time, etc.) which is recorded in major published dictionaries in four *kanji* forms: 手古摺る, 梃摺る, 手子摺る and 梃子摺る. In fact contemporary Japanese usage, as can be seen in the Google n-gram corpus (described below), is that the verb is almost invariably written in either a simpler mixed *kanji/kana* form (手こずる) or in *kana* alone (てこずる). (These two forms account for over 95% of the usage of the term.)

In this paper we explore approaches to identifying, selecting and presenting variant surface forms in entries in order to optimize the usefulness of the displayed entry, while at the same time enabling the entries to be correctly identified using all of the known forms.

## 2. Context

The particular context for the investigation and evaluation reported in this paper is the JMdict (Japanese-Multilingual Dictionary) project (Breen, 2004). (See the online project documentation.) This project, which developed from an earlier simple text-based dictionary project, aims to compile and maintain a comprehensive decoding dictionary, primarily for use by non-native Japanese speakers.

The project operates an open online database of dictionary entries (Breen, McGraw, 2013). While anyone can submit new entries or amendments, an editorial team evaluates and edits all submissions prior to them being accepted. The dictionary is distributed daily in an XML format which has a rich micro-structure. The distribution is under a Creative Commons "CC-SA" licence and can be used without charge. The dictionary database covers Japanese-English bilingual entries, however the daily XML distribution includes an expanded form in which translations for Japanese to a range of other languages (German, French, Spanish, Dutch, etc.) from other open-source projects are included. A version with a collection of example Japanese and English sentence pairs linked to specific entries is also provided.

The particular focus of the dictionary is on modern Japanese, including colloquialisms and slang which can be demonstrated to be in common use. It also sets out to record variant surface forms which are common enough to warrant collection.

The dictionary is used by a wide variety of websites (jisho.org, WWWJDIC, etc.) and apps (ImiWa, AEdict, etc.) and has a large international following (Breen 2009). While many of the systems that use the dictionary are of the traditional "look up a term" variety, many others use it to support text-glossing and other education-oriented functions.

The JMdict database currently has over 200,000 entries. Of these over 160,000 have surface forms which include *kanji*, of which over 30,000 (19%) have more than one form. Of the over 40,000 *kana*-only entries, about half have multiple surface forms.

More detailed information about the project can be found in the online documentation; see the links listed at the end of the paper.

### 3. Issues of Japanese Orthography and Surface Forms

Lexical items in Japanese are written using a combination of one or more of four scripts:

- a. *kanji* (Chinese characters). e.g. 辞書, 言葉, etc. About 2,000 *kanji* are studied at school and several thousand more are in reasonably common use. They are used primarily to write nouns and the roots of verbs, adjectives, adverbs, etc.
- b. the *hiragana* syllabary (あいうえおかきくけこ...) This consists of 46 characters plus diacritics. It is used primarily for particles, conjunctions, short common words, and the inflections of verbs and adjectives.
- c. the *katakana* syllabary (アイウエオカキクケコ...) which also consists of 46 characters plus diacritics. In modern Japanese orthography it is used primarily to transcribe loanwords and non-Japanese names. It is also commonly used for the taxonomical names of plants, animals, etc., for slang and colloquial terms, and for emphasis.
- d. the Roman alphabet. This is mainly used for acronyms and initialisms, both of foreign terms and names (e.g. CD, DJ) and Japanese terms (e.g. NHK, JR).

As mentioned above, a wide variety of surface forms can be used in Japanese. This variety in surface forms comes about for a number of reasons, including:

- a. the choice of *kanji*; often two or more *kanji* are available with the same meaning and pronunciation.
- b. replacing one or more *kanji* in a term with their reading(s) in *kana*.
- c. variations in the writing of parts of morphemes in *kana*.
- d. variations in approaches to the transcription of loanwords.

In evaluating the surface forms, a number of major Japanese reference dictionaries are regularly consulted. The dictionaries include the following Japanese-language dictionaries (国語辞書: *kokugojiten*): 広辞苑 (*Kōjien*, Iwanami), 大辞泉 (*Daijisen*, Shōgakusan), 大辞林 (*Daijirin*, Sanseidō), 明鏡 (*Meikyō*, Ōbunsha), and the Kenkyūsha 新和英大辞典 (*Shin Waei Daijiten*) Japanese-English dictionary.

It is apparent from examination of published Japanese dictionaries that surface form variations are not widely included. Often alternative *kanji* forms are included, but these usually relate to rarely used historical forms. Relatively little attention is paid, for example, to mixed *kanji/kana* forms or variant loanword transcriptions. It is noted that in the case of online editions of some of these reference dictionaries mixed forms get some degree of lookup support (see the mention of the Kenkyūsha system below.)

### 4. Establishment of Term and Form Frequency

An important part of this study has been the detection and confirmation of variant forms of Japanese terms. Recent work in corpus linguistics has led to the compilation of major collections of terms which have the potential to be used for this purpose. One such collection is the Balanced Corpus of Contemporary Written Japanese (BCCWJ) which has been developed by the Center for Language Resource Development within the National Institute for Japanese Language and Linguistics (NINJAL) It is a publicly-available word-list with over 2 million items (NINJAL 2014). Unfortunately this and other word-lists usually do not include many variant surface forms. For example, of the six plain forms for the verb *tekozuru* mentioned above, only one, the comparatively rare 手子摺る, occurs in the BCCWJ list.

An important resource which has proved invaluable in detecting forms "in the wild" has been the Google Japanese N-gram Corpus (Kudo, Kazawa, 2009). The corpus, which was constructed from a crawl of web pages in 2007, contains 1.3 billion short text segments along with their frequency counts. It has been shown that frequencies drawn from web-based corpora are generally of high quality, and comparable with those from other corpora. (Keller, Lapata, 2003)

A set of interactive and batch tools has been developed for exploring this corpus and enabling determination of the frequency of use of the various alternative forms. (Breen, 2024)

For example, in this corpus the comparative counts and frequencies for the various forms of *tekozuru* can easily be extracted and are shown to be:

Table 1: N-gram Frequencies for *tekozuru*

Surface Form	Count	Percentage
手こずる	23,647	48.3%
梃子摺る	1,665	3.4%
手古摺る	135	0.3%
手子摺る	39	0.1%
梃摺る	29	0.1%
てこずる	23,487	47.9%

## 5. Specific Issues in Surface Form Variation

We will now examine in detail the issues that lead to surface form variation, as they arise from several issues related to Japanese script usage.

### 5.1 Choice of *Kanji*

A significant number of *kanji* have two or more forms which are coded independently in the character codes used in computer text, and people are able to choose freely which they use. The availability of alternative forms has increased in recent years with the wide acceptance and implementation of the Unicode coding system. Before that the coding system commonly used in Japan had fewer alternative forms available (Breen, 2021). The variety of alternative *kanji* forms and the issues associated with their use are described below.

- Traditional and simplified *kanji* forms. A number of characters in use in both China and Japan had, over an extended period, seen the development of simplified forms which were used along with the traditional and generally more complex forms. A typical example is the pair 学 and 學 which have the same meaning and pronunciation. After some limited reforms in the early 1900s a major reform to *kanji* usage was conducted in the late 1940s which led to the establishment of the 当用漢字 (*tōyōkanji*: general use *kanji*) set of 1,850 characters for use in schools, textbooks, etc. (Seeley 1991, Twine 1991, Unger 1996) This set placed emphasis on the simplified forms, commonly known as 新字体 (*shinjitai*: new character form) in contrast to the 旧字体 (*kyūjitai*: traditional character form). Dictionaries published from then tended to include only the forms using simplified characters. The standards for *kanji* computer codes, established initially in the 1970s, also emphasized simplified versions. (Breen, 2021)
- Although the 当用漢字 set and its successor the 常用漢字 (*jōyōkanji*: daily use *kanji*) emphasized the simplified forms of characters, the expanded character set provided with Unicode has seen some drift back to more complex forms. For example a recent expansion of the 常用漢字 added the 頰 *kanji* instead of the simpler 頰 form which was (and still is) more commonly used.
- In 2000 the 文化庁 (*bunkachō*: Agency for Cultural Affairs) published a list of 1,022 non-当用漢字 commonly used in print (表外漢字字体表). In this list it gave preference to traditional *kanji* forms in a number of cases. For example it recommends 鹵 be used instead of 鹵. This form only became available in computer text with Unicode, and words such as *sekken* (soap) are still far more commonly written as 石鹵 rather than 石

鹼. It has been noticed that recent editions of dictionaries have begun to include these "preferred" *kanji* forms. Other cases where traditional forms are promoted in this list include 弯/彎, 鷗/鷗 and 啞/啞.

- d. Apart from official and semi-official variants, there are many other cases where alternative *kanji* can be used. For example, 麩質 is a relatively uncommon term for gluten in Japanese. It can also be written as 麩質 - the 麩 *kanji* being an alternative for 麩. (Gluten is usually written using the loanword グルテン.)

Another source of variation is the use of *kanji* which are very similar in appearance. The 柿 character, which is read *kaki* and means persimmon is often used as a substitute for the 柿 character, which is read *kokera* and means wood shaving or shingle (the two *kanji* are virtually identical in appearance). In fact the early computer code standards had 柿 filling both roles and most dictionaries used 柿 alone. The 柿 form only became available for computer use with Unicode, and although recent dictionary editions have begun to use it, terms such as *kokeraita* (thin shingles) are still more commonly written with the "incorrect" 柿板 *kanji* form, or the mixed こけら板 form.

- e. *Kanji* form variation also occurs with the use of *ateji* (当て字) forms, where combinations of characters are used for their pronunciations instead of the core meanings. Traditionally *ateji* have been used for such things as long-standing loanwords but on occasions they can be used as alternative forms for Japanese terms. One example is the word *surume* (dried squid or cuttlefish), which is usually written only in *kana*. The *kanji* form (鰯) is included in many dictionaries but is rarely used. An *ateji* version, 寿留女, which is rarely included in dictionaries, is often used as a substitute, especially in contexts such as food labelling and packaging.
- f. A common source of *kanji* form variation comes from the writer selecting an unrelated *kanji* which has a similar appearance. There are many examples of this, such as using 氣慨 instead of 氣概, 奇遇性 instead of 奇偶性, and 蕃主 instead of 藩主. The mechanism is often associated with keyboard input conversion systems and is variously known as 誤変換 (*gohenkan*) or 変換ミス (*henkanmisu*: *conversion mistake*).

## 5.2 Mixed Kanji/Kana Forms

Possibly the largest single category of multiple surface forms in terms containing *kanji* arises from the 交ぜ書き (*mazekaki*: mixed writing) process, in which a complex or less-common *kanji* is replaced by its *kana* reading. An example is the popular *sukiyaki* dish, which appears in many dictionaries as 鋤焼き but is almost always written as すき焼き.

交ぜ書き forms are particularly common with compound verbs. For example, *toroiritsuku* (to finally arrive [at]), is usually shown in dictionaries as 辿り着く however usage of the たどり着く form is more than twice as common, and the 辿りつく version is also in use.

In many cases, particularly in technical or medical terminology, the *kana* part of 交ぜ書き forms is written in *katakana*. For example, the term *erakokyū* (branchial respiration) is more often written as エラ呼吸 than えら呼吸, with the 鰓呼吸 *kanji* form being comparatively rare.

## 5.3 Okurigana Variations

A large number of Japanese morphemes consist of a *kanji* in association with one or more *kana* as a suffix. These *kana* are known as *okurigana* and are usually associated with the inflecting part of a verb or adjective. For example, the term 交ぜ書き (above) is derived from the verbs 交ぜる (*mazeru*: to mix, to blend) and 書く (*kaku*: to write) with the ぜ and き being the *okurigana* parts.

Often forms can vary as to the use of *okurigana*, for example the term 向こう脛 (*mukōzune*: shin) is often written 向う脛 and occasionally as just 向脛. Some published dictionaries will indicate such variations in *okurigana* with parentheses, for example the noun 取り扱い (*toriatsukai*: treatment, service), which is often written as 取扱い, will be found in some dictionaries as 取(り)扱い.

## 5.4 Irregular Kana Usage

A number of variant forms arise from irregular use of *kana* sequences which have the same pronunciations. A common example of this arises from the adjective 遠い (*tōi*: far, distant) which has the *kana* form of とおい. The use of the とお for "tō" is much less common than とう, and many people use the latter despite it being irregular. The expression お待ちどおさま (I'm sorry to have kept you waiting) is more commonly written お待ちどうさま as a result of this.

Another source of variants is from the voicing of the つ (*tsu*) kana to form *zu*. This should be written using the usual voiced form づ, but often it is written with as ず, which is the voiced form of す (*su*) and has a similar pronunciation.

## 5.5 Loanword Transcription Issues

Japanese has adopted many loanwords, which are typically transcribed into the *katakana* script. Most have been taken from English, but there are also many from French, German, Dutch and Portuguese in particular. The mapping of the pronunciation of the source words into the syllabary is not a precise process, and many variant transcriptions are in use leading to multiple forms for many terms.

An example of the variations can be seen in several words beginning with "dia", where the "a" is often transcribed as a ヤ (*ya*) and a ア (*a*). Thus the word "diamond", which is usually transcribed as ダイヤモンド (*daiyamondo*), is often rendered as ダイアモンド (*daiamondo*). Similarly the word "diagram" is transcribed as ダイアグラム (*daiaguramu*) and ダイアグラム (*daiyaguramu*) in about equal proportions.

Some words have more than two transcriptions. For example, "syndicalism" is usually transcribed as サンジカリズム or サンディカリズム reflecting two approaches to the "di" syllable (ジ vs デイ), and on occasions サンディカリズム is used where the "ism" is not voiced (all three forms can be found in published dictionaries.)

Some terms have large numbers of variant transcriptions. The main transcription of "flavor" is フレーバー which accounts for about 90% of the usage, but there are seven other variants in regular use. Similarly the word "vibraphone" has two common transcriptions (ビブラフォン and ヴィブラフォン) covering about 80% of the use but the other 20% is spread across ten other versions.

Loanwords containing the letter "v" often have variant transcriptions as many people choose a voiced "u" (ヴ) instead of the more standard "ba" (バ) or "bi" (ビ). According to the n-gram data, ヴァイオリン is more common as the term for "violin" than the standard バイオリン.

In some cases erroneous transcriptions have become quite common. The correct version of "assignment" is アサインメント (*asainmento*), which appears in many dictionaries, but an incorrect transcription アサイメント (*asaimento*) is equally common.

A common cause of variant transcriptions is the different handling of lengthened vowels. The standard with loanwords is to use the ー (*chōon*) character but many people use vowels such as イ, ウ, etc. instead. Thus "mate" is often written as メイト instead of メート, and "day" as デイ or デイ instead of デー. In most cases only the standard transcriptions occur in dictionaries.

A further form of variation with loanwords occurs with multi-word compounds or phrases where the "・" middle dot (*nakaguro*) character is often inserted between the component words. Thus the transcription of the golfing term "long iron" can be ロングアイアン or ロング・アイアン. Many dictionary entries use the "・" form, but in general writing such forms are less common.

## 6. Recording and Presentation of Multiple Surface Forms

As explained above, many Japanese terms can be written with varying surface forms. This sets some significant challenges for the compilers of decoding dictionaries for learners and non-native speakers. Having identified entries with multiple surface forms, the questions arise as to:

- which forms should be recorded
- which forms should be presented to dictionary users in the regular display of an entry;
- what order should be used to present the multiple forms;
- what additional information should be provided, e.g. sources, accuracy, etc.

The approach that has been adopted in the JMdict project is:

- record all the forms of a term which can be shown to be in use, or are in published reference sources;
- provide appropriate meta-information with the forms to assist the developers of systems using the database to decide whether forms should be displayed in entries, and the ordering of the display. A key goal here is avoid "clutter" of the entry display with forms which can be found in use but are not really common enough to be exposed;

- provide meta-information with the forms to indicate the nature and source of the forms.

An overview of these approaches is presented below. The full details of the meta-information tags can be found in the online project documentation.

### 6.1 Recording of Variant Surface Forms

Surface forms of terms are recorded if they can be shown to be in use (e.g. in corpora) or are included in references such as dictionaries (not all forms in dictionaries actually appear in corpora.) The Google n-gram corpus has an inclusion threshold of 20 occurrences, so the presence of a form in that resource is an indication that it is in use. Actual inclusion of a form in the database is decided on a case-by-case basis, and will typically take into account the proportions as well as absolute counts. Of course in the case of inflected verbs and adjectives the sum of the counts of the various inflected forms needs to be considered as well.

### 6.2 Visible and Hidden Forms

The arrival of electronic and online editions and forms of dictionaries has provided additional opportunities for identifying entries which go beyond traditional methods such as alphabetic or *kana* order. These alternatives are seeing some use, e.g. in the online system provided by the Kenkyūsha company for its collection of dictionaries the entry for *tekozuru* (mentioned above), for which it uses the headword form "手古摺る", can be found by searching for "手こずる" even though this form does not appear in the entry itself.

This approach of having effectively hidden or search-only forms is being supported by the JMdict project by associating meta-information with those forms to indicate to systems using the data that they should not (or need not) display the forms by default. The decision is, of course, up to the developers of the systems but in most cases the recommendations are being followed. Some systems allow optional viewing of these forms.

Of the entries in the JMdict database with multiple *kanji* forms, about 15% currently have one or more search-only forms.

In the case of loanwords with multiple variant versions the less common forms are usually tagged as search-only. For example, in the case of フレーバー (flavor) mentioned earlier, the other seven variants have been tagged as search-only.

### 6.3 Entry Display Order

The approach being followed in the JMdict project is to order the forms according to frequency of occurrence, largely based on the n-gram corpus. In most cases this is the same ordering seen in published dictionaries, however on occasions it results in a mixed 交ぜ書き form leading where it the most frequently used form.

An issue which possibly affects the ordering is the status of the *kanji* in the terms. For example, in the 2010 revision of the 常用漢字 list the 叱 *kanji* form was effectively replaced by 叱, which had not been in earlier character standards. The previous 叱 form was classified as an "official variant". Consideration was given as to whether terms containing the *kanji*, e.g. the verb 叱る/叱る (to scold), should lead with the new "official" *kanji* form. As usage of the "new" form is still comparatively rare, it was decided to leave the frequency-based ranking in place for now.

Another possible ordering issue relates to terms which have one or more *kanji* forms but in fact the term is more commonly written only in *kana*. An example of this is the verb 萎れる (*shioreru*: to droop, wilt, wither), which is usually written only in *kana*. This practice is not usually mentioned in dictionaries, however the main Kenkyūsha Japanese-English dictionary uses the *kana* form alone in its example sentences.

In JMdict a meta tag is included at the sense level when a term is primarily written in *kana* alone. Most user systems reflect this in the ordering of their entry displays by having the *kana* form(s) first.

In the case of loanword sequences containing the "." middle dot the usual ordering is to have both versions visible with the plain version first followed by the middle dot version. An exception to this is the case of initialisms such as DVD where the dotted reading (ディー・ブイ・ディー) is made the leading one and the plain reading (ディーブイディー) is made search-only.

## 6.4 Status Tagging of Forms

In addition to the meta-information mentioned above relating to search-only forms and ordering, a number of other meta-information items are in use to advise users of the dictionary of the status of the forms. They are:

### a. Rarely Used Forms

These tags are to indicate the forms, which have low levels of use, are recorded in major dictionaries and for this reason have not been treated as search-only. An example is 併記 (*heiki*: side-by-side writing), for which 並記 is a rarely used alternative which occurs in several dictionaries.

### b. Incorrect *Kanji* Usage

This is applied to forms which use an incorrect *kanji* but are common enough not to be made search-only. The purpose of the tag is to alert dictionary users of the status of the form.

### c. Incorrect *Kana* Usage

As with the *kanji* case above, the tag is used with relatively common but incorrect *kana* forms. These are usually either incorrect readings of *kanji* or incorrect transcriptions of loanwords. Examples include the お待ち どうさま expression and the アサイメント loanword mentioned earlier.

### d. Outdated *Kanji* or *Kana* Usage

These typically apply to forms containing 旧字体 *kanji*, or outdated or historical *kana* readings of terms.

## 7. Conclusion

In this paper the issue of variant surface forms of Japanese terms has been explored in the context of their inclusion in dictionaries, both as visible headwords and as search-only lookup keys. The source of the variations has been identified and analyzed, and the degree to which variants are used "in the wild" has been investigated using large corpora. It has been noted that on occasions variant forms are used far more often than those occurring in published dictionaries, which is particularly challenging to language learners.

Within in the JMdict dictionary project an approach has been developed and implemented for recording and structuring variant surface forms in a way which maximizes the usefulness of the collection of terms to the target audience. The approach involves assembling a wide variety of forms, and attaching meta-information to indicate the recommended display status of the terms, including their source and status.

## Acknowledgements

The author wishes to acknowledge the significant role of the members of the editorial team of the JMdict project in the formulation and development of the approach to the handling of multiple surface forms within the dictionary entries.

## References

- Breen J. 2004. 'JMdict: a Japanese-Multilingual Dictionary', *Multilingual Linguistic Resources Workshop*, COLING 2004, Geneva
- Breen J. 2009. 'WWWJDIC - a feature-rich WWW-based Japanese dictionary' *eLexicography in the 21st century conference*, Louvain-la-Neuve, Belgium, October 2009. pp. 381-386
- Breen J., McGraw S. 2013. 'JMdictDB - Online Development and Maintenance System for the Japanese-Multilingual Dictionary' ASIALEX 2013 - *Lexicography and Dictionaries in the Information Age*, August 2013, Bali, Indonesia.
- Breen J. 2021. 'Kanji and the Computer: A Brief History of Japanese Character Set Standards' *Joy O'Kanji* (<https://www.edrdg.org/~jwb/paperdir/kanjicomp.html>)
- Breen J. 2024. *Japanese Word Frequencies - Google N-Gram Corpus* (<https://www.edrdg.org/~jwb/paperdir/JGNGWordFreq.html>)
- Keller F, Lapata M, 2003 'Using the Web to Obtain Frequencies for Unseen Bigrams', *Computational Linguistics* 29(3), pp. 459-484
- Kudo T., Kazawa H., 2009. *Japanese Web N-gram Version 1* (<https://japan.googleblog.com/2007/11/n-gram.html>) (<https://catalog.ldc.upenn.edu/LDC2009T08>)

NINJAL. 2014. *BCCWJ: Balanced Corpus of Contemporary Written Japanese*, National Institute for Japanese Language and Linguistics (<https://clrd.ninjal.ac.jp/bccwj/en/freq-list.html>)

Seeley, C. 1991. *A History of Writing in Japan*. University of Hawai'i Press.

Twine, N. 1991. *Language and the Modern State: The Reform of Written Japanese*. Routledge.

Unger, J.M. 1996. *Literacy and Script Reform in Occupation Japan. Reading between the Lines*. Oxford University Press.

### **JMdict Documentation**

Project Documentation and Links ([https://www.edrdg.org/wiki/index.php/Main\\_Page#Documentation\\_and\\_Links](https://www.edrdg.org/wiki/index.php/Main_Page#Documentation_and_Links))

Kanji and Reading Information Fields Documentation

([https://www.edrdg.org/wiki/index.php/Kanji\\_and\\_Reading\\_Information\\_Fields](https://www.edrdg.org/wiki/index.php/Kanji_and_Reading_Information_Fields))