

Project

Cameron

June 9, 2021

Contents

1	Introduction	1
2	Data Definitions	1
3	Use Cases	6
4	Script Usage	7

1 Introduction

This document will serve as the recordings of my work with parsing the incredibly expansive Japanese to English dictionary, JMDICT. The purpose of this parser is to a) generate a JSON version of the file to be used with other projects, namely my Yosh!eru project. I am writing this document both for myself and for demonstration purposes. The dictionary is defined by a grammar, which I detail below.

2 Data Definitions

- JMDict (entry*)
- entry (ent_seq, k_ele*, r_ele+, sense+)
 - ent_seq - Entry sequence number
 - k_ele - kanji elements
 - r_ele - reading elements
 - sense - sense element

- ent_seq (#PCDATA)
 - Unique numeric sequence number for each entry
- k_ele (keb, ke_inf*, ke_pri*)
 - The defining element of each entry. In its absence, the reading element is such.
 - Majority of entries have a single kanji element associated with a word in Japanese
 - Where there are multiple kanji elements within an entry, they will be orthographical variants (alternate spellings mostly)
 - Common "mis-spellings" may be included
 - Synonyms are not included
- keb (#PCDATA)
 - This element will contain a word or short phrase in Japanese which is written using at least one non-kana character.
 - Valid characters are kanji, kana, related characters such as chouon and kurikaeshi, etc.
- ke_inf (#PCDATA)
 - This is a coded information field related specifically to the orthography of the keb
 - Will typically indicate some unusual aspect, such as okurigana irregularity.
- ke_pri (#PCDATA)
 - Along with re_pri field, provided to record information about the relative priority of the entry
 - Consist of codes indicating the word appears in various references which can be taken as an indication of frequency of usage.
 - Current values in this field are:
 - * news1/2: appears in the "wordfreq" file compiled by Alexandre Girardi from the Mainichi Shimbun
 - * ichi1/2: appears in the "Ichimango goi bunruishuu", Senmon Kyouiku Publishing, Tokyo, 1998

- * spec1/2: a small number of words use this marker when they are detected as being common, but are not included in other lists.
 - * gail/2: common loanwords, based on the wordfreq file.
 - * nfxx: this is an indicator of frequency-of-use ranking in the wordfreq file. "xx" is the number of the set of 500 words in which the entry can be found, with "01" assigned to the first 500, "02" to the second, and so on.
- r_ele (reb, re_nokanji?, re_restr*, re_inf*, re_pri*)
 - The reading element typically contains the valid readings of the word(s) in the kanji element using modern kanadzukai
 - Where there are multiple reading elements, they will typically be alternative readings of the kanji element
 - In the absence of kanji, these elements will define the entry
 - reb (#PCDATA)
 - This is restricted to kana and related characters such as chouon and kurikaeshi.
 - Kana usage will be consistent between keb and reb elements (if one contains katakana, both will)
 - re_nokanji (#PCDATA)
 - This element, which will usually have a null value, indicates that the reb, while associated with the keb, cannot be regarded as a true reading of the kanji.
 - It is typically used for words such as foreign place names, gairaigo which can be in kanji or katakana, etc.
 - re_restr (#PCDATA)
 - This element is used to indicate when the reading only applies to a subset of the keb elements in the entry.
 - In its absence, all readings apply to all kanji elements.
 - The contents of this element must exactly match those of one of the keb elements.
 - re_inf (#PCDATA)

- General coded information pertaining to the specific reading.
 - Typically it will be used to indicate some unusual aspect of the reading.
- re_pri (#PCDATA)
 - See the comment on ke_pri
- sense (stagk*, stagr*, pos*, xref*, ant*, field*, misc*, s_inf*, lsource*, dial*, gloss*)
 - The sense element will record the translational equivalent of the Japanese word, plus other related information.
 - Where there are several distinctly different meanings of the word, multiple sense elements will be employed.
- stagk (#PCDATA)
- stagr (#PCDATA)
 - These elements, if present, indicate that the sense is restricted to the lexeme represented by the keb and/or reb.
- xref (#PCDATA)
 - This element is used to indicate a cross-reference to another entry with a similar or related meaning or sense.
 - The content of this element is typically a keb or reb element in another entry.
 - In some cases a keb will be followed by a reb and/or a sense number to provide a precise target for the cross-reference.
 - * Where this happens, a JIS "centre-dot" (0x2126) is placed between the components of the cross-reference. The target keb or reb must not contain a centre-dot.
- ant (#PCDATA)
 - This element is used to indicate another entry which is an antonym of the current entry/sense.
 - The content of this element must exactly match that of a keb or reb element in another entry.

- pos (#PCDATA)
 - Part-of-speech information about the entry/sense.
 - Should use appropriate entity codes.
 - In general where there are multiple senses in an entry, the part-of-speech of an earlier sense will apply to later senses unless there is a new part-of-speech indicated.
- field (#PCDATA)
 - Information about the field of application of the entry/sense.
 - When absent, general application is implied.
 - Entity coding for specific fields of application.
- misc (#PCDATA)
 - This element is used for other relevant information about the entry/sense.
 - As with part-of-speech, information will usually apply to several senses.
- lsource (#PCDATA)
 - This element records the information about the source language(s) of a loan-word/gairaigo.
 - If the source language is other than English, the language is indicated by the xml:lang attribute.
 - The element value (if any) is the source word or phrase.
 - * lsource xml:lang - Defines the language(s) from which a loan-word is drawn
 - * lsource ls_type - Indicates whether the lsource elemnt fully or partially describes the source word of phrase of the loadword. If absent, it will have the implied value of "full". Otherwise it will contain "part".
 - * lsource ls_wasei - Indicates that the Japanese word has been constructed from words in the source language
- gloss (#PCDATA | pri)*

- Within each sense will be one or more "glosses", i.e. target-language words or phrases which are equivalents to the Japanese word.
- This element would normally be present, however it may be omitted in entries which are purely for a cross-reference.
 - * gloss xml:lang - defines the target language of the gloss
 - * gloss g_type - Specifies that the gloss is of a particular type, e.g. "lit" (literal), "fig" (figurative), "expl" (explanation)
- pri (#PCDATA)
 - These elements highlight particular target-language words which are strongly associated with the Japanese word.
 - The purpose is to establish a set of target-language words which can effectively be used as head-words in a reverse target-language/Japanese relationship.
- s_inf (#PCDATA)
 - The sense-information elements provided for additional information to be recorded about a sense.
 - Typical usage would be to indicate such things as level of currency of a sense, the regional variations, etc.

3 Use Cases

This section will outline the use cases for this project, which will help in defining the database schema and algorithms needed. The web application will be referred to from here as 'the app', and any possible user simply as 'the user.'

1. The app will deliver an audio file to the user, along with a corresponding audio transcription of the file. The user will listen to the clip and 1) simply reveal the transcription, or 2) attempt to answer with the correct transcription (seeing the answer upon submission).
2. The user will be able to create and edit 'decks' of 'flashcards' and test themselves with the flashcards.
3. The user will be able to look up any word displayed on the app and see an entry for the word. The entries will include:

- The definition of the word
 - Alternative words (Same definition)
 - Words used within the word
 - Kanji used within the word
 - Examples of the word
 - JLPT level and other tags
4. The user will be able to convert words and example sentences into flashcards and add them to a deck

4 Script Usage

The script can be ran with the following command:

```
python3 JMDictToJSON.py [--option]
```

The following options are available for use with the script.

- indent
 1. By supplying `-indent=#`, will cause each nested level in the JSON file to have `#` number of indents inserted at the front of the line
- low memory
 - With `-low-memory`, tells the parser to keep a low memory profile. The output is written in chunks (currently every 10,000 entries), and entries are not kept in memory. In the future when analytical capabilities are added to this project, some with likely not be available with this mode enabled.