Preliminary Thoughts on Issues of Modeling Japanese Dictionaries Using the OntoLex Model

Louis Lecailliez

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
louis.lecailliez@outlook.fr

Abstract. Recent works aiming at making Linked Data dictionaries make use of the Lemon or OntoLex models. Application to existing dictionaries revealed the need for extensions to the model to properly deal with lexicographic data without loss of information. These works however focus on languages found in Europe, and thus let the issue of Est-Asian lexicography for future exploration. This paper provides a small typology of existing dictionaries in Japan and exposes issues in existing related works that could form the ground of new modules for OntoLex.

Key words: Linked Data, Lemon model, dictionary, e-lexicography, Japanese

1 Introduction

After a first wave of dictionary computerization, where they have been structured in a way close to their existing paper embodiment [13], the next step appears to be graph-based dictionaries [16]. In this trend, dictionaries are made or being converted using formalisms and the technology stack of the Linked Data [5]. In particular, the OntoLex model [12] – based on the Lemon model – initially created to lexicalize ontologies is under development to support more information coming from traditional (electronic) dictionaries. However, as Bosque-Gil [1] note it: "Future steps include the analysis of dictionaries in languages that are underrepresented in the LLOD cloud (e.g. Japanese) to identify further representation challenges".

The present work aims to start identifying a few of these issues. At first, we expose a concise typology of existing Japanese dictionaries in order to pin down the lexicographic landscape of the Japanese language. In this paper, we focus on the Japanese language but some parts are equally applicable to other languages as well. That's why we will indicate when a class of dictionary is relevant to the East-Asia area as a whole. Secondly, existing works related to the problem of modelling dictionaries and encoding lexical knowledge of Chinese characters and the Japanese language will be presented. These works are not directly incorporable in OntoLex [18] as such because of an initial divergent goal but form a very viable base of reflexion for a dedicated module.

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2017, pp. 11–19, 2017.

2 Concise Typology of Japanese Dictionaries

In addition to unilingual and bilingual dictionaries, Japanese features some original kind of dictionaries: some are specific to Japan while others are also found in the whole Sinosphere. Most of these dictionaries emerged because of the characteristics of the Chinese writing system, which was imported in Japan and subsequently derived to write the autochthon language. Others, such as the accent and katakana dictionaries arose respectively from features of Japanese and its vernacular writing system.

Table 1 lists the most common dictionaries found in Japan. For each class, it is specified if it also exists in other countries of the Sinosphere, or if it can be found in Japan only. A dash means it is not specific to either regions. The last column describes the form (writing system and number of characters) of headwords compiled in a given dictionary type, if applicable. Kana encompass hiragana and katakana.

2.1 Chinese Characters[1] Related Dictionaries

The first class of specific dictionary to be found in Japan is the *kanwa-jiten* (漢和辞典). Literally a "Chinese-Japanese Dictionary", it focusses on explaining Chinese characters and compound words made from Chinese morphemes that were borrowed in Japanese. This is different from the *ch-unichi-jiten* (中日辞典) that are bilingual dictionary of contemporary Mandarin to modern Japanese language. The class may be further split between *kanji-jiten* and *kango-jiten*. The former lists and describes sinograms and the afferent readings, meanings and compounds while the latter focusses on providing lists of compound words that use a given Chinese character. Chinese character dictionaries can also be found in Korea under the generic term of *hanja-sajeon* (漢字辭典).

Another type of kanji dictionary exists that is targeted at non-Japanese people: the bilingual sinogram dictionary. In them, characters are headwords, meaning is given in a foreign language and readings are written in a romanization. The *New Nelson Japanese-English Character Dictionary* [7] for the Japanese-English pair and the *Dictionnaire Ricci de caractères chinois* [17] for Mandarin-French are example of such dictionaries.

2.2 Proverb Dictionaries

Two types of dictionary exist for proverbs. A *kotowaza-jiten* is a book listing idioms: any proverb used in Japanese can fit in this kind of dictionary, regardless of its form. The *yojijukugo-jiten* on the other hand only lists idiotisms made of four Chinese characters. Most of the time entries in such a dictionary feature

[1] In the rest of the paper, the terms "Chinese character" and "sinogram" are used in an interchangeable way to denote the Chinese characters and their use in the whole Sinosphere. The Japanese term "kanji" (漢字) is used only when speaking of sinograms in a Japanese context.

a Japanese reading using Japanese native words and have an idiomatic value, hence the compilation in a different kind of dictionary than a *kango* or *kotowaza* dictionary. Proverb dictionaries are also found in China under the name of *chengyu-cidian* (成語詞典). As most proverbs of Mandarin are made of four characters, there is only one type of dictionary for them in Chinese.

2.3 Specialized Dictionaries Specific to Japan

The written language of Japan was for centuries modeled on literature of ancient times. The diglossia between the spoken and the written languages made it so that dictionaries are well needed for understanding the classical language. This is the raison d'être of the classical language dictionary: *kogo-jiten* (literally old language dictionary).

The contemporary Japanese language features a pitch accent that is not uniform between locations. There is a class of dictionary (accent dictionary) made to indicate the "proper" accentuation of words, modelled on the Tokyo dialect. One of them is available on the web [15].

Foreign words of non-Chinese origin are written with the Japanese katakana syllabary. Some dictionaries compile such words. Contrary to other dictionary types, these dictionaries make use of the latin alphabet in the definitions in order to display word in their original writing.

Finally, the various styles that were used to write sinograms through the ages give birth to the need for *jitai-jiten* (style dictionary) that compile writing of characters in different styles. That kind of work is different from the others in that the information it encodes cannot be stored in text form. These dictionaries are used by calligraphers or readers of literary work in manuscript form.

Table 1. Most common and specialized type of dictionaries found in Japan

Dictionary type	Japanese name	Romanized name	Specific to	Headword
Unilingual Dictionary	国語辞典	Kokugo jiten	—	Kana
Bilingual Dictionary	XY辞典	XY jiten	—	Kana or Alphabet
Etymology Dictionary	原語辞典	Gengo jiten	—	Mixed
Sinogram Dictionary	漢字辞典	Kanji jiten	Sinosphere	One kanji
Chinese Compound Dictionary	漢語辞典	Kango jiten	Sinosphere	Kanji
Proverb Dictionary	ことわざ辞典	Kotowaza jiten	—	Mixed
"Four Character Compound Dictionary"	四字熟語辞典	Yoji-jukugo jiten	Sinosphere	Four kanji
Accent Dictionary	アクセント辞典	Akusento jiten	Japan	Kana
Classical Language Dictionary	古語辞典	Kogo jiten	Japan	Kana
Dictionary of Words in Katakana	カタカナ語辞典	Katakanago jiten	Japan	Katakana
Style Dictionary	字体辞典	Jitai jiten	Sinosphere	Kanji

2.4 Modeling Problematics

It is clear from the list of dictionaries listed in Table 1 that one module cannot encode by itself all information required by dictionaries that address such a variety of concerns. As a starting point, the main dictionaries to be modeled are unilingual and *kanwa* dictionaries. Moreover, the different dictionaries target distinct demographics and needs. In particular, the native speaker population and the non-native one have different concerns when it comes to searched entries. A native is probably more interested in checking the meaning of a word or the way to write it in *kanji*, while a foreign learner may be more concerned about the reading of a character or a word and its translation in his native language. Both unilingual and *kanji* dictionaries are used in conjunction even by natives to find the meaning of an unknown word written in *kanji* [2].

This duality is expressed in the dictionaries themselves: unilingual dictionaries will compile entries listed in a phonetic way and sub-entries may be distinguished by graphical forms. One the other hand, *kanji* dictionaries headwords are characters for which multiples readings are listed in a syllabary (often either in hiragana or katakana given the origin of the reading). This particular setting means, as we will see in section 4.2, that a graphical form (if represented by the *lemon:Form* class) cannot be considered as depending only of a lexical entry but needs to be linked someway to a concept, otherwise some information is lost.

Using the translation module mentioned by Gracia [6], bilingual (characters) dictionaries can be constructed by linking unilingual Japanese to other language lexicons. With this framework in mind, issues can be categorized in three sets related to: (1) Chinese character modeling, (2) representing unilingual Japanese dictionaries without loss of information, and (3) interaction between Chinese characters and other lexical entries of any other kind of dictionary, which is also needed to solve (2).

3 Modeling Chinese Characters

In the cultures that use Chinese characters as their main script or as part of their writing, sinograms are a lexicographic object *per se*. As such, a way to model them is needed.

The complete modeling of Chinese characters, the relation between them and to the language that use them is complex because it requires representing various phenomena and multiple many-to-many relationships. Modeling the sinograms for the Japanese language causes almost all the same problematics as modeling them for the Mandarin language, but additional phenomena needs to be taken in account. For example, additional information exists about the distinction between the type of reading (Sino-Japanese or pure Japanese) and the historical periods from which the readings were borrowed.

Related works on the matter include Hantology [3], an ontology derived from the Chinese writing. It was later expanded to include *kanji* as well [9].

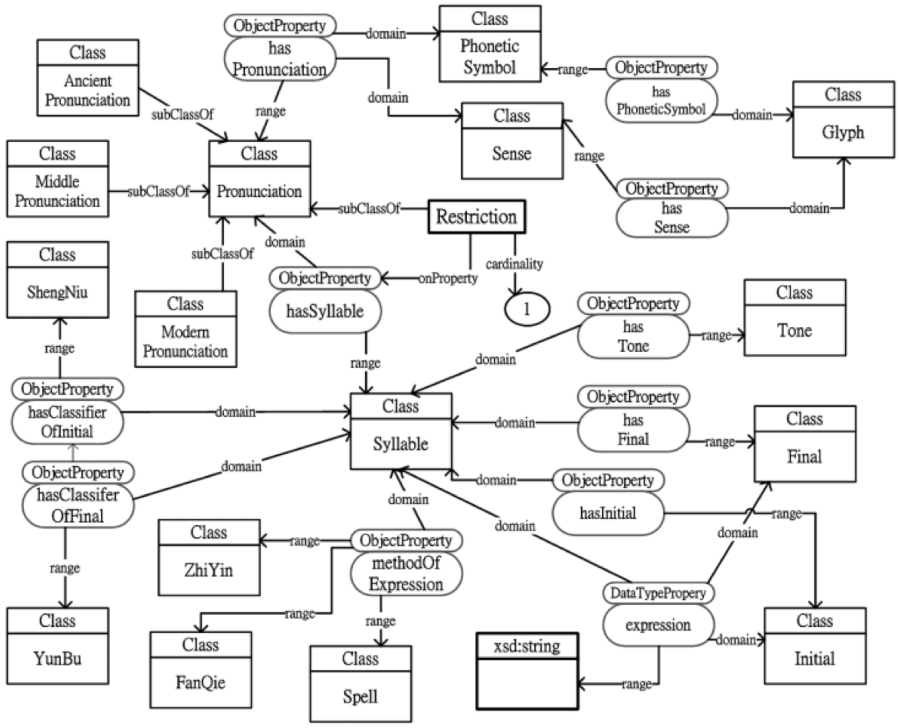

Fig. 1. Sinogram pronunciation description modeled in OWL from [3].

The ontology is based on the character decompositions and definitions given by the *Shuowen Jiezi* (說文解字) dictionary [19]. While their goal was not to encode the content of this dictionary nor any others, the meticulousness of their modelling and its implementation in Web Ontology Language [18] are a suitable base for later work. The whole ontology is detailed, schematized and explained in [3]. The sub-graph of the model dedicated to pronunciation description is reproduced as Figure 1.

A great attention has been made to the problem of character variants in Hantology. While there exist a sheer amount of characters – more than 50,000 referenced in the *Daikanwa Jiten* (大漢和辞典) – most are actually variants that were used at a given time and place. This information is important for automatic processing where the number of entries matters but at the same time introduce noise when they are not needed.

Link between variants is also an important point for linking sinograms between languages that may use different glyphs for the same character: it can act as a link between data resources of various languages and thus increase the connectivity in the Linguistic Linked Open Data (LLOD) cloud.

4 Modeling the Japanese Lexicon

4.1 The JLP-O Model

Linked Data modelling of the Japanese lexicon already went under scrutiny in various work of Joyce and Hodošček [10,11] about building an ontology of lexical properties of Japanese: the JLP-O model. Although the aim of these works is not to represent dictionaries *per se*, advance on this front is particularly interesting as the chosen model is built on Lemon. Joyce and Hodošček [10] derive five classes from lemon's *LexicalEntry* class to fit the perceived needs for representing the Japanese lexicon. Beside the complexity of the Japanese writing system, the Japanese lexicon feature a great deal of compound words and three of the classes (*BoundUnit*, *SimpleWord*, *ComplexWord*) were created specifically to deal with it.

The Figure 2 reproduces the figure 3 from [10]. It illustrates how the word "yomu" (to read) is handled in the JLP-O model. The model regroups the various graphical forms of a word under the *canonicalForm* and *orthographicForm* properties of a main lexical entry. It raises two issues. First, forms that carry a slightly different meaning are not separated in their own lexical entry. While this is not a problem in a tradition paper dictionary, it may cause wrong inference while using the graph (see section 4.2 below). In this example, both "読む" and "詠む" are grouped in the same entry despite the latter being only used in the context of reading poetry and song [8]. It means that the compound verb "読み始める" (to start reading) that links the "yomu" verb is also indirectly linked to forms that are actually not be used to write it. The second issue is related to the first: each form is wrapped in a blank node. RDF blank nodes are anonymous and thus cannot be reused as the object or subject of other properties. The first stated issue thus arises and cannot be solved because the precise forms that needs to be linked are not directly accessible.

4.2 Making Wrong Inferences

The *yomu* example shows how the pronunciation of a word and its graphical forms are intertwined in relation their meaning. A given word may feature very different senses if written with different characters. Reciprocally, a given written form can be used to write word of very different meaning. For example, "十分" read as *juubun* means enough; read as *juppun* or *jippun* it means ten minutes.

It is thus essential to link a couple of (form, reading) to a meaning. Failure to do so would allow a program using the graph to make inferences that are not true. For example, let have a *LexicalEntry* instance with form F_1 (meaning S_1) and F_2 (meaning S_2) which is linked to ontology entities denoting concepts S_1 and S_2. A program reading the graph could make the respective inferences by transitivity: F_1/F_2 can be used to express concept S_1/S_2 but also that F_2 can denote S_1 and F_1 means S_2. The two later predicates are false.

The way the class *Form* is linked to a *LexicalEntry* independently of a the *LexicalConcept* and *LexicalSense* class is a problem in modelling Japanese dictionaries. In addition, a similar problem exists in modelling *kanji*.

```
jlpo:読む_動詞-一般
  a jlpo:SimpleWord ;
  lemon:canonicalForm [
    lemon:writtenRep "読む"@ja ;
    jlpo:decomposition (
      [ jlpo:Character jlpo:読_character ]
      [ jlpo:Character jlpo:む_character ] ) ;
    jlpo:use [ jlpo:frequency 23324 ; jlpo:corpus "BCCWJ" ] ] ;
  jlpo:orthographicForm [
    lemon:writtenRep "読む"@ja ;
    jlpo:decomposition (
      [ jlpo:Character jlpo:読_character ]
      [ jlpo:Character jlpo:む_character ] ) ;
    jlpo:use [ jlpo:frequency 20382 ; jlpo:corpus "BCCWJ" ] ] ;
  jlpo:orthographicForm [
    lemon:writtenRep "よむ"@ja ;
    jlpo:decomposition (
      [ jlpo:Character jlpo:よ_character ]
      [ jlpo:Character jlpo:む_character ] ) ;
    jlpo:use [ jlpo:frequency 322 ; jlpo:corpus "BCCWJ" ] ] ;
  jlpo:orthographicForm [
    lemon:writtenRep "詠む"@ja ;
    jlpo:decomposition (
      [ jlpo:Character jlpo:詠_character ]
      [ jlpo:Character jlpo:む_character ] ) ;
    jlpo:use [ jlpo:frequency 653 ; jlpo:corpus "BCCWJ" ] ] ;
  # [... 9 other orthographicForms ...]
```

Fig. 2. Reproduction of figure 3 from [10]: Part of the RDF representation for the SimpleWord lexical entry '読む' in Turtle format.

5 Silex: Towards a Lemon Module for Chinese Characters

Chinese characters are an important lexicographic object for the Japanese language processing. It must be tackled first because almost every other kind of dictionary will link them from other lexical entries. Unilingual dictionaries typically feature annotations that need such an atomic decomposition to be fully encoded.

Sinograms also pose an issue of many-to-many relationships between readings and meanings of the same character. A similar problem exists for the Japanese orthography as a whole, thus solving the problem at the character level will provide a template to solve it a higher level.

Finally, as sinograms are used or were used in a variety of East-Asian language, it makes sense to model them in the most language agnostic way. That allows reuse of entities in a multilingual context, increasing the connectivity within the Linguistic Linked Open Data cloud. From this point of view, a language agnostic term for Chinese character should be chosen for the main entity. The term sinogram answers this problem elegantly by burying the reference to the Chinese culture in a root from latin and avoiding the use of

localized term such as *kanji* (Japanese), *hanzi* (Chinese), *hanja* (Korean), *hán tự* (Vietnamese). And it abbreviated nicely as Silex, the Sinogram Lexical module.

Acknowledgments. This paper was written with the support of the MSMT-10925/2017-64-002 scholarship grant from the Czech Ministry of Education, Youth and Sports.

References

1. Bosque-Gil, J., Gracia, J., & Montiel-Ponsoda, E. (2017). Towards a Module for Lexicography in OntoLex. *DICTIONARY News*, 7.
2. Breen, J. (2004). Multiple Indexing in an Electronic Kanji Dictionary. In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries* (pp. 1-7). Association for Computational Linguistics.
3. Chou, Y. M., & Huang, C. R. (2005). Hantology: An ontology based on conventionalized conceptualization. In *Proceedings of the Fourth OntoLex Workshop. A workshop held in conjunction with the second IJCNLP*. October (Vol. 15).
4. Chou, Y. M., & Huang, C. R. (2013). Hanzi zhishi de xingshi biaoda [汉字知识的形式表达]. *Dangdai yuyanxue* [当代语言学], 15(2), 142-161.
5. Gracia, J., Kernerman, I., & Bosque-Gil, J. (2017). Toward Linked Data-native Dictionaries. In Kosem, I., Tiberius, C., Jakubíček, M., Kallas, J., Krek, S., & Baisa, V. (eds) *Proceedings of eLex 2017 conference*, September 19-21, Leiden, Netherlands. Lexical Computing CZ s.R.O, Brno, Czech Republic.
6. Gracia, J., Villegas, M., Gómez-Pérez, A., & Bel, N. (2016). The apertium bilingual dictionaries on the web of data. *Semantic Web*, (Preprint), 1-10.
7. Haig, J., Nelson, A. (1997). The new Nelson Japanese-English character dictionary. C.E. Tuttle Co.
8. Hayashi, S., Nomoto, K., Minami, F., & Kunimatsu, A. (1992). Sanseid-o Reikai Shinkokugo Jiten, 3rd edition [三省堂例解新国語辞典　第三版].
9. Huang, Y. M., Huang, C. R., & Hong, J. F. (2008). The Extended Architecture of Hantology for Kanji. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Tapias, D. (eds) *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, May 28-30, Marrakech, Morocco. European Language Resources Association (ELRA).
10. Joyce, T., & Hodošček, B. (2014). Constructing an ontology of Japanese lexical properties: Specifying its property structures and lexical entries. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)* (pp. 174-185).
11. Joyce, T., Masuda, H., & Hodošček, B. (2016). Constructing a Database of Japanese Lexical Properties: Outlining its Basic Framework and Initial Components. *Tama University Global Studies Department Bulletin Paper*, 8, 35-60.
12. McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In Kosem, I., Tiberius, C., Jakubíček, M., Kallas, J., Krek, S., & Baisa, V. (eds) *Proceedings of eLex 2017 conference*, September 19-21, Leiden, Netherlands. Lexical Computing CZ s.R.O, Brno, Czech Republic.
13. Měchura, M. (2016). Data Structures in Lexicography: from Trees to Graphs in Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016*, pp. 97–104, 2016.
14. Nagasawa, K. (1991). Sanseid-o Jiten 4th edition [三省堂漢和辞典　第四版].

15. Nakamura, I., Minematsu, N., Suzuki, M., Hirano, H., Nakagawa, C., Nakamura, N., Tagawa, Y., Hirose, K., Hashimoto, H. (2013). Development of a web framework for teaching and learning Japanese prosody: OJAD (Online Japanese Accent Dictionary). *Proceedings of INTERSPEECH,* pp.2254-2258.
16. Polguère, A. (2014). From writing dictionaries to weaving lexical networks. *International Journal of Lexicography*, *27*(4), 396-418.
17. Ricci Institute. (2001). Grand dictionnaire Ricci de la langue chinoise. Desclée de Brouwer, Paris.
18. W3C Ontology Lexicon Community Group. (2016). Final Model Specification. `https://www.w3.org/community/ontolex/wiki/Final_Model_Specification`
19. Xu, S. (121). Shuowen Jiezi [說文解字].