Constructing an Ontology of Japanese Lexical Properties: Specifying its Property Structures and Lexical Entries

Terry Joyce
School of Global Studies,
Tama University,
Fujisawa, Japan
terry@tama.ac.jp

Bor Hodošček
School of Global Japanese Studies,
Meiji University,
Nakano, Japan
bor.hodoscek@gmail.com

Abstract

Regarding the construction of an ontology of Japanese lexical properties (JLP-O) as fundamental in terms of establishing a conceptual framework to guide and facilitate the construction of a large-scale lexical resource (LR) database of the Japanese lexicon, this paper primarily focuses on two major concerns for the construction of the JLP-O. The first is to map out and appropriately structure the numerous lexical and psycholinguistic properties, or variables, associated with the Japanese lexicon. The second concern is to specify an appropriate range of lexical entries classes within the JLP-O. Both concerns have far-reaching implications for effectively capturing the rich patterns of interconnections among lexical entries and lexical properties and thus for realizing a multifunctional LR. After discussing the solutions integrated into the current Resource Description Framework (RDF) representation of the JLP-O, the paper also briefly describes the extraction of a corpus-based lexicon from the recently released Balanced Corpus of Contemporary Written Japanese (BCCWJ; Maekawa et al., 2013), an authoritative sampling of the contemporary Japanese lexicon. Categorized according to the JLP-O's range of lexical entry classes, and supplemented with orthographic variant and decomposition information, the BCCWJ-based lexicon represents a key reference LR for constructing the large-scale LR.

1 Introduction

The overarching objective of our research project is to construct a large-scale lexical resource (LR) of Japanese lexical properties—interpreted inclusively as any characteristic or variable associated with words—which, as a comprehensive model of the Japanese lexicon, can potentially be beneficial for various researchers within the linguistic and cognitive sciences. Within that larger endeavor, we regard the task of constructing an ontology of Japanese lexical properties (JLP-O) as being absolutely foundational for two important reasons. The first is primarily pragmatic in nature. As reflected in the relatively recent trend towards merging LRs and ontologies (Huang et al., 2010; Oltramari et al., 2013), the formal specification of the ontology can unquestionably provide considerable advantages in terms of enhanced compatibility with natural language processing (NLP) and knowledge system tools for efficiently integrating data, checking for consistency, and realizing powerful query functionality. In contrast, however, the second reason is both more conceptual and more skeptical in nature. In many ways, ontology construction can be thought of as the very epitome of academic endeavor in seeking to clearly elucidate the phenomenon of interest, but it is also crucial to understand that natural systems, such as language, do not necessarily conform to the standards of ontological completeness. As outlined further in section 2, our approach to ontology construction particularly values the utility of the ontology as working conceptual framework. Reflecting this, our approach attempts to strike a reasonable balance between ontological rigor, on the one hand, and recognizing a number of other important cognitive criteria, on the other hand, such as theoretical description, consistencies, psychological reality, and preferences, in order to realize a core framework that can both guide the construction of the LR and, ultimately, facilitate multifunctional querying.

Against that larger background, this paper specifically focuses on two major concerns addressed in constructing the JLP-O. Given the extensive range of Japanese lexical properties that must be represented in a satisfactory manner within a large-scale LR for the Japanese lexicon, the first concern has been to map out and appropriately structure the many and varied lexical and psycholinguistic properties, or variables, associated with the Japanese lexicon into domains, or modules. The second major concern has been to specify an appropriate range of lexical entries as core entities of the JLP-O for a highly agglutinative language like Japanese. As outlined in more detail in section 3, these concerns have direct implications for implementing effective links among lexical entries and lexical properties. In section 4, we explain how both concerns have been resolved within the JLP-O in ways that simultaneously help to represent the rich patterns of interconnectivity between various lexical properties and facilitate the realization of a multifunctional LR that both possesses powerful search capabilities and can be utilized by a wide range of users. Section 4 also outlines the extraction and formal encoding of a major corpus-based lexicon essential for constructing the LR. Section 5 recaps the main points and briefly discusses future work for the larger LR project.

2 Ontology as conceptual framework

After some general comments about defining ontologies, this section also briefly introduces the two models for LRs that we specifically draw inspiration from in constructing the JLP-O; namely, the lemon model (lexical model for ontologies; http://lemon-model.net/) and Spohr's (2012) model for multifunctional LRs.

2.1 General comments

Although Gruber's (1993; 199) immensely influential pronouncement that an "ontology is an explicit specification of a conceptualization" continues to provide the basic template, following Guarino (1998) and Guarino et al. (2009), many subsequent definitions of ontologies tend to also emphasize the shared nature of the conceptualization (Guarino et al., 2009; Prévot et al., 2010). The elements of this ontology definition require a little unpacking. First, conceptualization refers to both the explicit and implicit knowledge about a system or entity, such as its component entities and their relations. Next, explicit, or formal, specification refers to a commitment to encode the body of knowledge in the form of some representation language, usually in a machine-readable format. And, finally, shared conceptualization refers to the criterion that, to have value, there should be a general consensus among interested parties about the target conceptualization (Guarino et al., 2009; Prévot et al., 2010).

As already suggested, our approach towards ontology construction is admittedly somewhat nuanced in nature, reflecting a basic tension at the conceptual level. While we fully concur with the laudable drive towards clearer descriptions of phenomena that ontology construction entails, we are equally cautious of seeing ontologies alone as some magical panacea for all knowledge representation problems. The sentiment is particularly visceral in the case of natural systems like language which abounds in various forms of redundancy and biases that are not readily represented by ontologies. Thus, in our efforts to construct a comprehensive model of the Japanese lexicon, we are endeavoring to incorporate vital aspects of linguistic and cognitive knowledge that are embedded within its diverse lexical properties. However, at the pragmatic level, we acutely recognize the numerous benefits of adopting the ontology as a conceptual framework for effectively realizing the overall research objective of constructing a large-scale LR database of Japanese lexical properties. In some ways, our qualified position on ontology construction is rather aptly captured in the following comments from Franconi, Kerhet, and Ngo (2013):

> An ontology provides a conceptual view of the database and it is composed by constraints on a vocabulary extending the basic vocabulary of the data. Querying a database using the terms in such a richer ontology allows for more flexibility than using only the basic vocabulary of the relational database directly.

2.2 Models for linguistic resources

For the sake of clarity (albeit at some risk of possibly overstating what may already be sufficiently obvious), our primary objective in constructing the JLP-O is to have a conceptual view, or framework, to aid the development of a large-scale LR database with multifunctional querying capabilities, which we hope will come to serve as a comprehensive model of the Japanese lexicon. That is to say, we are seeking to apply linguistic, psycholinguistic and cognitive conceptualizations about Japanese words in order to realize a formal specification of the Japanese lexicon. This naturally brings into focus the next vital piece in the puzzle; namely, the need for an ontology model that is particularly suitable for linguistic resources, where linguistic and psycholinguistic conceptualizations (lexical properties) are linked to the lexical entries (words) of the database. Although a number of candidate models exist, such as LexInfo (Cimiano, Buitelaar, McCrae, & Sintek, 2011), LIR (Linguistic Information Repository; Peters, Montiel-Ponsoda, & Cea, 2007) and LMF (Lexical Markup Framework; Francopoulo, 2013), the present work draws inspiration most directly from the lemon model and Spohr's (2012) model for multifunctional LRs.

Building directly on the LexInfo, LIR and LMF models, lemon has been specifically developed to be a standard for the exchange of lexical information on the semantic web, and so it has a number of advantages that are particularly appealing for JLP-O. These include the facts that lemon is based on RDF, a semantic web standard that can greatly facilitate the representation of links between parts of the LR, and that, reflecting its policy not to prescribe over linguistics definitions, lemon effectively delegates the burdens of constraining domain-specific information to external sources, such as WordNet and ontologies of linguistic descriptions such as GOLD (General Ontology for Linguistic Description; http://linguistics-ontology.org/). Other advantages are that lemon is relatively concise, because it requires few classes and relies on external definitions, and that it is organized in terms of a number of separate modules, which can be constructed independently for greater flexibility. In contrast, reflecting its emergence from the intersection between semantic web technology and lexicography, Spohr's (2012) model for multifunctional LRs is particularly concerned with the informational needs of diverse users, encompassing both humans (from monolinguals, bilinguals, novices, to linguistic experts) and NLP applications, and with realizing suitable query and display interfaces. As the goal of achieving a high degree of multifunctionality, in Spohr's sense of the notion, is also central to our LR project, we particularly take to heart Spohr's suggestion that one vital key for realizing multifunctionality is the incorporation of an appropriate typology, or range, of lexical entries.

3 Construction concerns

This section briefly sketches out the two major issues for constructing the JLP-O; namely, appropriately structuring the wide range of lexical properties associated with the Japanese lexicon and determining a suitable range of lexical entries. The solutions incorporated into JLP-O's RDF representation are discussed further in section 4.

3.1 Range of Japanese lexical properties

For researchers within the language and cognitive sciences to be able conduct significant research on various aspects of the Japanese language, such as developing more robust models and simulations of linguistic and cognitive abilities, obviously, access to a wide range of information about the contemporary Japanese lexicon is absolutely essential. Traditionally, available LRs have been limited to various kinds of dictionaries, such as language dictionaries like Shinmura's (2008) *Kōjien* and Kindaiichi et al.'s (2011) *Shinmeikai Kokugojiten* and character dictionaries like Morohashi's (2000) *Daikanwajiten*. However, as dictionaries rarely provide much summary information beyond headword counts, researchers have also had to rely on scarce sources of data summaries. Hayashi's (1982) *Zūsetsu Nihongo* is a classic example that included a lexical section, with some frequency, word class and formation information, an orthographic section, with some counts, usage and readings information for kanji in particular, as well as sections on phonology and accent, grammar, and style.

Setting aside genuine issues for keeping such data up-to-date, however, a particularly serious problem is the tremendous expansion in the range of lexical properties that researchers require data about today. One helpful way to understand the variety of lexical properties is in terms of Nation's (2001/2013) aspects of knowing a word. Highly influential in the areas of vocabulary research and second language acquisition, his framework consists of nine broad types of word knowledge grouped under three main categories of form (spoken, written, and word parts), meaning (form and meaning, concept and referent, and associations) and use (grammatical functions, collocations, and constraints on use (such as register and frequency)). In addition to the breadth dimension, it is also useful to think about the considerable range of lexical properties in terms of the depths of analyses conducted within these various domains. For instance, taking just the domain of visual word recognition experiments to illustrate, Adelman (2012) recently notes 14 kinds of potentially confounding lexical variables that should be controlled for, including frequency and contextual diversity data, various forms of neighborhoods (orthographic, phonological, phonographic, and Levenshtein-distance), spelling-sound regularities and consistencies, length, morphological properties as well as rating-based measures. While granting that some of these form-related lexical properties will undoubtedly also be of interest to researchers in other areas, still, equally undeniably, they will also require information about many other lexical properties, such as semantic properties of denotation and semantic groups of thesauri, or usage information, such as valencies, collocations and associations, and educational levels.

In addition to facilitating the integration and ongoing maintenance of the large-scale LR, we see the construction of the JLP-O as being especially valuable for helping to elucidate divergent interpretations about lexical properties. For instance, regardless of the markedly different theoretical motivations underlying orthographic neighborhoods and morphological families, LRs of kanji compound neighborhoods and LRs of morphological families yield identical data, at least, with respect to the possible two-compound word combinations for a given set of kanji. Awareness of such data equivalencies despite contrasting ontological perspectives is vital both for realizing robust queries of the LR and, in turn, for developing more robust simulations and models.

3.2 UniDic's short-unit words

For a research project aiming to construct a large-scale LR that can serve as a comprehensive model of the Japanese lexicon, one of the thorniest issues that must be addressed is surely just what to treat as the core entities within the LR database in the case of a highly agglutinative language like Japanese where word boundaries are often ambiguous. While doubting that an ideal solution exists, given that any decision is certain to have broad implications for implementing a LR, the issue must be taken seriously.

In this context, it is illustrative to look at UniDic; the electronic morphological dictionary for the Japanese language that was developed as part of the BCCWJ project. Reflecting its objectives to be a high-performance dictionary for wide coverage of contemporary written Japanese, UniDic adopted as its prime entity the so-called short-unit word (SUW), which roughly corresponds to the shortest meaningful unit, the morpheme. However, although that decision is certainly not without some justification, as Joyce, Hodošček and Nishina (2012) discuss, it is also fair to say that the SUW is far from convenient for human users, unless additional information about higher-order groupings is also readily available. Although the BCCWJ project's provision of supplementary information in the form of annotations about so-called long-unit words (LUWs)—groupings of verb and adjective agglutinations and compound nouns as single units—probably originates from this issue, still, the distinction is somewhat artificial and requires a certain degree of familiarity. Figure 1 highlights the basic relationships between UniDic's SUWs and LUWs, with an example sentence of 報告書を読み始める /hōkokusho o yomihajimeru/ 'to begin reading a report (document)' consisting of three LUWs. The two content LUWs of 報告書/hōkokusho/ 'report (document)' and 読み始める /yomihajimeru/ 'to begin reading' are both combinations of two SUWs, while case-marking particles, like the object marker を/o/, are simultaneously both SUWs and LUWs.

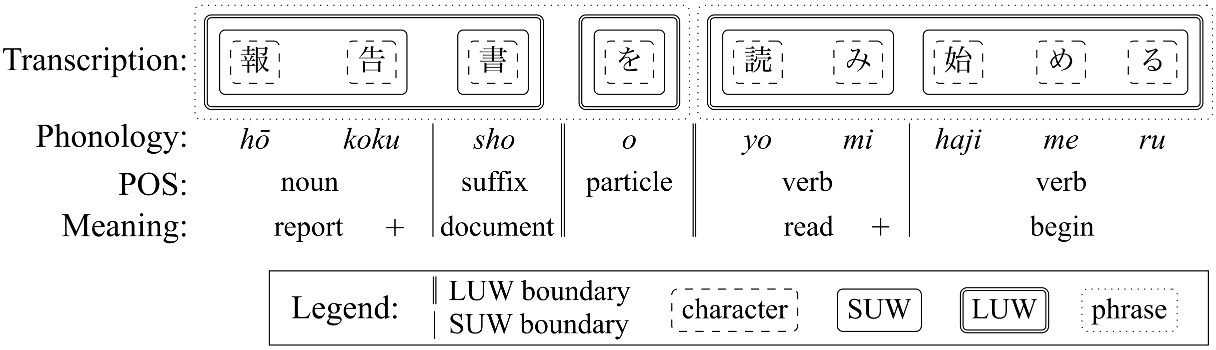

Figure 1: The relationships between characters, short-unit words, long-unit words and phrases in an example sentence of 報告書を読み始める /hōkokusho o yomihajimeru/ 'to begin reading a report (document)'.

While again acknowledging that any proposals about what to treat as core lexical entries within a LR of the Japanese lexicon are likely to involve some degree of compromise, still one obvious lesson to emerge from referring to the UniDic case is that a single lexical unit alone is insufficient in order to handle all items of the Japanese lexicon adequately for all kinds of users. As outlined further in section 4.2, our solution for JLP-O is to specify a wider range of lexical entries. However, despite the problems of SUWs and LUWs associated with the BCCWJ, it is an authoritative sampling of the contemporary Japanese lexicon, consisting of approximately 100 million words. Accordingly, it unquestionably remains the most valuable source from which to extract a corpus-based lexicon, with information about numerous core lexical properties, including lemma and orthographic specification, their respective frequencies, their phonological information and word classes, to establish a solid foundation to the construction of the large-scale LR project.

4 JLP-O construction

Having touched on the extensive range of Japanese lexical properties and the inherent tensions involved in selecting a range of lexical entry classes, this section turns to explain how these concerns have been handled in constructing the current version of the JLP-O. The section also outlines the extraction of the BCCWJ-based lexicon and the assignment of JLP-O's `LexicalEntry` subclasses as a foundation for the LR.

4.1 JLP-O modules

Although a number of separate Japanese LRs have already been created to address various lexical properties, given that they have been developed with different objectives and with diverging interpretations about the lexical properties themselves, their treatments of key properties, such as phonological, orthographic or semantic information, are not always consistent across resources, which can also vary greatly in terms of their levels of coverage. In order to help remedy this situation, the overall aim of our research project is to create a single comprehensive LR by constructing the JLP-O as its core framework to facilitate the integration of existing LRs.

Some of the initial groundwork for the LR project is outlined by Joyce, Masuda, and Ogawa (2014), within the context of discussing the revised jōyō kanji list as the core building block of the Japanese writing system. In addition to identifying and organizing a number of lexical properties at the jōyō kanji character level, they also describe a new analysis of the components of jōyō and JIS1 kanji, and apply an initial orthographic coding to the corpus word lists created in Joyce et al. (2012). Building directly from that, our continuing investigations of lexical properties have already identified 65 important properties; although, naturally, we fully expect the number to expand still further as additional LRs are consulted and examined for their particular merits. Reflecting both their natural mutual relationships and the need to structure their representations within the LR, these properties have also been organized under six modules of character, orthographic, phonological, morphological, semantic, and use. These are presented in Table 1 with a few examples of the relevant lexical properties.

Modules	Example properties
Character	type, configuration, internal structure, stroke counts, status, references, …
Orthographic	representation, variations, length (in characters), neighborhood data, …
Phonological	stress, length (in mora), CV structures, homophones, neighborhoods, consistency, …
Morphological	word structure, family data (size/frequency), constituent analysis, transparency, …
Semantic	denotation, connotations, sense range, lexical stratum, groups, concreteness, relations, …
Use	frequency/familiarity data, collocations, grammatical patterns, genre/register/style, …

Table 1: The six modules of the JLP-O with examples of relevant lexical properties.

This structuring of the lexical properties is highly consistent with lemon's modular design, which includes five modules in its core: linguistic description, variation, phrase structure, syntax and mapping, and morphology. We are, therefore, able to utilize a great deal of lemon's basic descriptive infrastructure pretty much intact, albeit with some relabeling of module names and some element reallocations to conform to JLP-O's modularization of lexical properties. For example, much of lemon's syntax and mapping module could be used with only minimal label changes in the integration of LRs such as Japanese versions of WordNet (Isahara et al., 2012) and FrameNet (Ohara et al., 2004) to JLP-O's semantic and use modules, respectively. In contrast, however, some basic characteristics of the Japanese lexicon, such as unit size issues and more extensive levels of orthographic variation, necessitate more expressive alternatives to both lemon's variation module and the decomposition property within the phrase structure module. Given these clear parallels, however, we believe that lemon's notion of modules is the most effective approach to structuring the lexical properties and to realizing their complex mappings to the range of lexical entries within our LR.

4.2 JLP-O's core range of lexical entries

As noted in subsection 3.2, one of the thorniest concerns to address in constructing a large-scale LR of the Japanese lexicon is to determine a suitable range of lexical entries to use as the core entities of the database. The issue is certainly far from straightforward, because it involves finding a workable compromise between a set of conflicting constraints. Naturally, these include the highly agglutinative nature of the Japanese language itself, which, understandably, encourages a focus towards the smallest components. However, these constraints also include the needs of diverse users of the LR, where, in contrast, as Spohr (2012) convincingly argues, a wider range is preferable for enhancing the search capabilities of an LR. At its heart, however, the issue is primarily about representation, or formal specification, given that the complex relationships between lexical entries themselves and between lexical entries and the modules of lexical properties must be efficiently captured within the JLP-O.

However, on consulting with our two reference models for LRs for guidance, we discover that they adopt radically different approaches to the specification of lexical entries. Consistent with its aspirations to be concise, lemon specifies just three classes of lexical entries; namely, `Part`, `Word` and `Phrase`. While this rather minimalist level of specification is, arguably, not so dissimilar to the de facto distinction that emerges with UniDic between SUWs and LUWs, comparisons quickly break down on closer inspection. For example, SUWs cover both bound morphemes (affixes and particles) and free morphemes (simple words), but these would correspond to lemon's part and word classes, respectively. The LUW concept also fails to fit nicely with lemon's tripartite division. Given that LUWs are either polymorphemic words or compound words formed by combining SUWs, the unit does not extend to phrases which are not marked by UniDic. In sharp contrast, but also consistent with his goal of multifunctionality, Spohr's MLR model incorporates a highly detailed typology of lexical entries (lexemes). Although the upper-level division into `BoundUnit`, `FreeUnit`, and `Clitic` may not, at first glance, appear so different, the subsequent divisions of `BoundUnit` into `BoundStem` and `Affix` (further divided into 9 kinds) and of `FreeUnit` into `Idiom`, `Syntactically-ComplexFreeUnit`, and `Syntactically-SimpleFreeUnit` (of which the final two are further divided eventually into 17 and 11 subclasses, respectively) clearly demonstrate very different theoretical motivations and objectives. That noted, however, the distinction between `Syntactically-SimpleFreeUnit` and `Syntactically-ComplexFreeUnit` parallels more closely to the contrast between SUWs and LUWs.

Aiming for a realistic balance between the constraints afforded by the characteristics of the Japanese lexicon, the LR's ambitions to realize a high degree of multifunctionality, and the need to achieve an acceptable degree of formal specification concerning the relationships among lexical entries and the six modules of lexical properties, the solution that we adopt for JLP-O is closer in spirit to the upper-levels of Spohr's (2012) typology of lexeme subclasses. More specifically, as illustrated in Table 2, we specify for JLP-O five classes of `LexicalEntry`, which are `Character`, `BoundUnit`, `SimpleWord`, `ComplexWord`, and `MultiWordExpression`. Thus, while the basic entities of the JLP-O draws inspiration more directly from Spohr's typology of lexeme classes, the range of JLP-O lexical entries has been increased in order to more faithfully represent the nature of the Japanese lexicon.

4.3 Extraction and RDF encoding of corpus lexicon

Having identified our practical solutions, this section briefly outlines their implementation in the RDF encoding. First, the current version of the JLP-O was specified by extending lemon's OWL specification using the Protégé ontology editor (http://protege.stanford.edu/). Second, a program was executed to simultaneously

Lexical entry type	Examples of units included
Character	kanji (仮), hiragana (か), katakana (カ), rōmaji (KA), …
BoundUnit	prefixes (御—), suffixes (—的), auxiliary verbs (—れる), …
SimpleWord	nouns (報告), verbs (読む), particles (を), adjectives (詳しい), …
ComplexWord	nouns (報告書), verbs (読み始める), adjectives (詳しくない), …
MultiWordExpression	collocations, idioms

Table 2: Examples of lexical entries.

extract the corpus lexicon from the BCCWJ corpus and assign the appropriate LexicalEntry subclasses.

In addition to replacing lemon's three classes of lexical entries with JLP-O's five LexicalEntry subclasses, two further minor extensions to the structure and facilities provided by the lemon model have also been necessary. The first minor extension relates to the high levels of orthographic variation that exists within the Japanese lexicon, as evidenced in Joyce et al. (2012), which is far beyond that envisioned by either lemon or Spohr's models. In seeking to be more consistent with UniDic's basic distinction between an abstract lemma form and all orthographic variations of a Japanese word, we have somewhat expanded upon lemon's distinction between canonicalForm and otherForm, by retaining the first label for the lemma form and changing the otherForm label to orthographicForm for each orthographic variant of a word. The second minor expansion is to more extensively utilize lemon's decomposition object property (which in lemon is limited to specifying the decomposition of phrases into parts). Thus, apart from the Character subclass within the current JLP-O (compositional analysis of radicals is not fully implemented at present), all lexical entry classes have a decomposition object property; such that both BoundUnit and SimpleWord entries are decomposed into one or more Characters, while the orthographicForms of ComplexWord and MultiWordExpression entries are decomposed into the relevant orthographicForms of BoundUnits and SimpleWords. By adopting this approach to linking structures, it is possible to search for complex lexical entries based on lower-level components, such as characters, by traversing the implemented hierarchical structure. Figure 2 shows a part of the lexical model with a focus on the lexical entries.

In order to extract the corpus lexicon, a program was written to convert the SUW and LUW information from the BCCWJ corpus into the JLP-O's RDF format, including the appropriate assignment of LexicalEntry subclasses. The M-XML format of the BCCWJ (version 1) includes basic structural encodings of LUWs in the form of lists of component SUWs. First, SUWs were assigned as either BoundUnit or SimpleWord lexical entries based on their unique identifier, which consists of the unique combination of the lemma and the POS category. Next, all orthographic variations of the lexical entry (based on their shared identifier) were recorded within the single entry specification using the orthographicForm object property, together with their decompositions into lists of characters. Finally, the frequency counts for orthographic variants were recorded using the use object property, allowing us to specify frequency counts together with their sources, which in the present case is a corpus identifier but could also specify a particular genre or style. The total frequency count, as sum of all orthographicForm variants, was also recorded under the canonicalForm property.

Figure 3 shows part of the RDF encoding in Turtle format for the SimpleWord lexical entry for the verb 読む /yomu/ 'to read'. For the sake of brevity, it is not possible to display all 12 orthographic variants, but the figure includes the three most frequent; the standard kanji-kana mixed orthography (of verbal stem and inflectional ending), the hiragana-orthography representation of よむ, and a kanji variation of 詠む with the nuances of 'to read or recite poetry; chant'.

Similarly, Figure 4 presents part of the RDF encoding in Turtle format for the SimpleWord lexical entry of the verb 始める /hajimeru/ 'to begin'. Interestingly, 始める is the verb2 element of many verb1-verb2 compounds that express senses of 'to begin V1'.

Finally, the extraction program also assigned LUWs to the ComplexWord subclass. The process essentially mirrored the extraction of BoundUnit or SimpleWord lexical entries, except that links under the decomposition object property link back to its constituent BoundUnit or SimpleWord lexical entries. Figure 5 presents part of the ComplexWord lexical entry for 読み始める /yomihajimeru/ 'to begin to read', which is

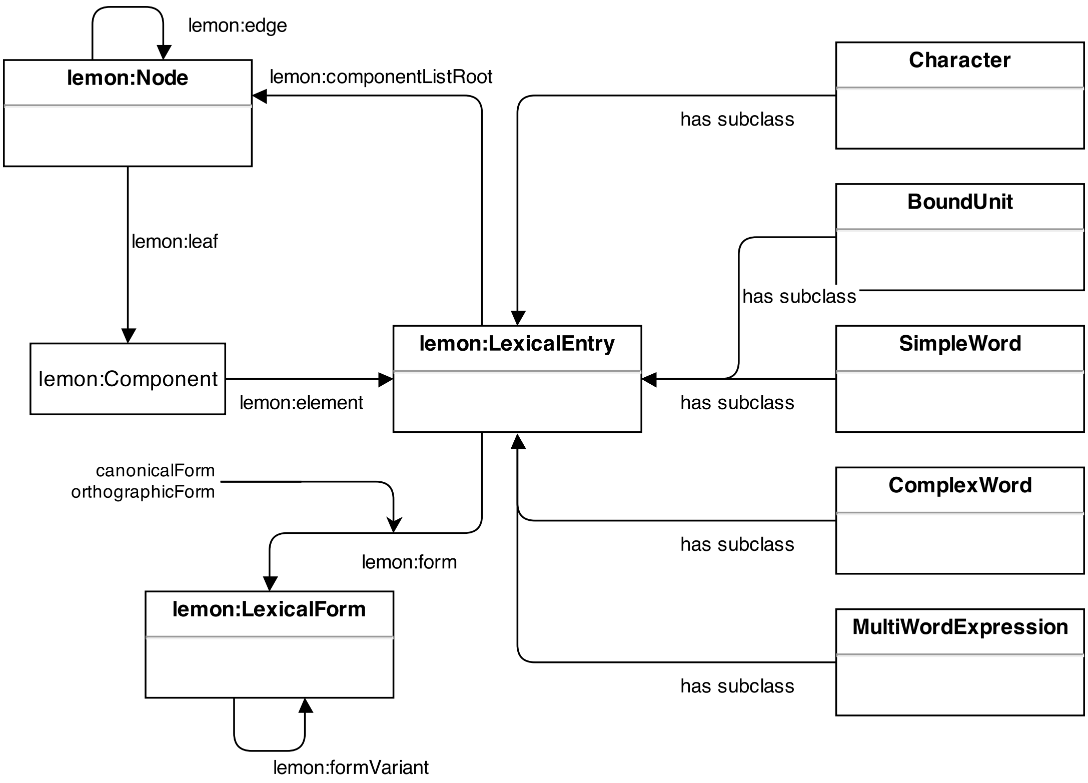

Figure 2: A subset of the JLP-O model.

```
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 ...]
  .
```

Figure 3: Part of the RDF representation for the SimpleWord lexical entry ‘読む’ in Turtle format.

```
jlpo:始める_動詞-非自立可能
  a jlpo:SimpleWord ;
  lemon:canonicalForm [
    lemon:writtenRep "始める"@ja ;
    jlpo:decomposition (
      [ jlpo:Character jlpo:始_character ]
      [ jlpo:Character jlpo:め_character ]
      [ jlpo:Character jlpo:る_character ] ) ;
    jlpo:use [ jlpo:frequency 30770 ; jlpo:corpus "BCCWJ" ] ] ;
  jlpo:orthographicForm [
    lemon:writtenRep "始める"@ja ;
    jlpo:decomposition (
      [ jlpo:Character jlpo:始_character ]
      [ jlpo:Character jlpo:め_character ]
      [ jlpo:Character jlpo:る_character ] ) ;
    jlpo:use [ jlpo:frequency 20591 ; jlpo:corpus "BCCWJ" ] ] ;
  jlpo:orthographicForm [
    lemon:writtenRep "はじめる"@ja ;
    jlpo:decomposition (
      [ jlpo:Character jlpo:は_character ]
      [ jlpo:Character jlpo:じ_character ]
      [ jlpo:Character jlpo:め_character ]
      [ jlpo:Character jlpo:る_character ] ) ;
    jlpo:use [ jlpo:frequency 10112 ; jlpo:corpus "BCCWJ" ] ] ;
  jlpo:orthographicForm [
    lemon:writtenRep "初める"@ja ;
    jlpo:decomposition (
      [ jlpo:Character jlpo:初_character ]
      [ jlpo:Character jlpo:め_character ]
      [ jlpo:Character jlpo:る_character ] ) ;
    jlpo:use [ jlpo:frequency 7 ; jlpo:corpus "BCCWJ" ] ] ;
  # [... 4 other orthographicForms ...]
  .
```

Figure 4: Part of the RDF representation for the `SimpleWord` lexical entry ‘始める’ in Turtle format.

```
jlpo:読み始める_動詞-一般
  a jlpo:ComplexWord ;
  lemon:canonicalForm [
    lemon:writtenRep "読み始める"@ja ;
    jlpo:decomposition (
      [ jlpo:SimpleWord jlpo:読む_動詞-一般 ]
      [ jlpo:SimpleWord jlpo:始める_動詞-非自立可能 ] ) ;
    jlpo:use [ jlpo:frequency 228 ; jlpo:corpus "BCCWJ" ] ] ;
  jlpo:orthographicForm [
    lemon:writtenRep "読み始める"@ja ;
    jlpo:decomposition (
      [ jlpo:SimpleWord jlpo:読む_動詞-一般 ]
      [ jlpo:SimpleWord jlpo:始める_動詞-非自立可能 ] ) ;
    jlpo:use [ jlpo:frequency 139 ; jlpo:corpus "BCCWJ" ] ] ;
  jlpo:orthographicForm [
    lemon:writtenRep "読みはじめる"@ja ;
    jlpo:decomposition (
      [ jlpo:SimpleWord jlpo:読む_動詞-一般 ]
      [ jlpo:SimpleWord jlpo:はじめる_動詞-非自立可能 ] ) ;
    jlpo:use [ jlpo:frequency 82 ; jlpo:corpus "BCCWJ" ] ] ;
  # [... 4 other orthographicForms ...]
  .
```

Figure 5: Part of the RDF representation for the `ComplexWord` lexical entry ‘読み始める’ in Turtle format.

a verb1-verb2 compound type consisting of 読み conjugation of 読む (verb1) together with 始める (verb2).

We executed the extraction and RDF encoding program to enumerate the BCCWJ-based corpus lexicon according to the JLP-O's core `LexicalEntry` classes. As summarized in Table 3, approximately 2.7 million lexical entries were assigned to the four core `LexicalEntry` classes.

Lexical entry	**Types**	**Tokens**
`Characters`	6,761	195,500,491
`BoundUnit`	433	11,327,729
`SimpleWord`	195,380	112,557,387
`ComplexWord`	2,438,506	101,684,786

Table 3: Type and token counts for the BCCWJ-based corpus lexicon.

The number of lexical entries assigned to the `Character` subclass is highly consistent with encoding specifications for Japanese characters. Similarly, the relatively smaller number of `BoundUnit` lexical entries is also consistent with the fact that this class consists of a small number of closed word classes, such as particles and the relatively limited sets of affixes. In contrast, the much higher counts for the `SimpleWord` and `ComplexWord` classes obviously reflect in large measure the fact that these cover the major open word classes, and, in particular, the noun class, which is extremely open. Another closely related factor is that these lexical entries also include vast numbers of proper nouns, which is a particular feature of large corpus data. The substantial difference between the `SimpleWord` and `ComplexWord` classes clearly illustrates the agglutinative nature of the Japanese language with rich verbal and adjectival conjugations and productive compounding. However, also a characteristic of large corpus data, it should also be noted that approximately 66% of the `ComplexWord` lexical entries occur only once within the BCCWJ corpus. And, a natural corollary is that while the `ComplexWord` lexical entries are on average decomposed into 3.1 `BoundUnit` and `SimpleWord` lexical entries, 94% of these have only one orthographic variant (because, for extremely low frequency words, one obviously requires even larger corpora to capture all possible orthographic variations). A final observation to make is that the corpus lexicon does not yield any lexical entries under the `MultiWordExpressions` subclass; although it would be feasible to extract collocational data from the BCCWJ, these will be identified for the large-scale LR in the future in the course of integrating other LRs.

5 Conclusion

As a principal component of a larger research project to construct a large-scale LR database concerned with the lexical and psycholinguistic properties associated with the Japanese lexicon, the paper has described the construction of the ontology of Japanese lexical properties (JLP-O) as its working conceptual framework. More specifically, the paper focused on two important issues. After outlining the first concern of mapping out and organizing the wide range of lexical and psycholinguistic properties that linguistic and cognitive science researchers require up-to-date information about in section 3.1, section 4.1 detailed how these are being structured under six modules, which mirrors the flexible approach towards construction employed by lemon. Similarly, after outlining the second difficult concern of what to treat as core entities of the LR database in section 3.2, section 4.2 explained the reason behind our solution to recognize five `LexicalEntry` classes, namely, that a wider range of classes is key to achieving the high degree of multifunctionality that our LR project aspires to (Spohr, 2012). Section 4.3 also outlined the extraction and RDF encoding of the BCCWJ-based corpus lexicon. Classified according to the JLP-O's range of lexical entries, and supplemented with information about orthographic variations, decompositions and frequencies, the corpus lexicon provides solid foundations for the large-scale project to construct the comprehensive LR of Japanese lexical properties.

Although we are fully aware that it will be necessary to further develop and refine the JLP-O as the larger LR project progresses, as the present working version has been constructed to specifically handle two fundamental aspects about the Japanese lexicon, we believe that it represents a sufficiently robust conceptual framework that can guide future work of integrating existing LRs. Thus, we approach the integration task not merely as a mechanical process of expanding the LR database by merging data, but as a dialectic one that requires ongoing consideration and investigation of the theoretical adequacies and psychological realities of candidate lexical properties; a reflective process that is exemplified by ontology construction. And, as a working conceptual framework for examining the LR database, the JLP-O represents the kind of architectural blueprint of the structural relationships within the LR database that is essential for realizing high degrees of multifunctionality in terms of developing various interfaces for search queries and data presentation. In this way, we hope to realize a comprehensive LR of Japanese lexical properties that will be beneficial as a systematic model of the Japanese lexicon that can be effectively mined in the pursuit of deeper insights into lexical knowledge.

Acknowledgments

This research was partially supported by a JSPS Grant-in-Aid for Foreign Postdoctoral Fellows (no. P13303).

References

Adelman, J. S. (2012). Methodological issues with words. In J. S. Adelman (Ed.), *Visual word recognition volume 1: Models and methods, orthography and phonology* (pp. 116–138). Current issues in the psychology of language. London: Psychology Press.

Cimiano, P., Buitelaar, P., McCrae, J., & Sintek, M. (2011). LexInfo: a declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, *9*(1), 29–51. doi:http://dx.doi.org/10.1016/j.websem.2010.11.001

Franconi, E., Kerhet, V., & Ngo, N. (2013). Exact query reformulation over databases with first-order and description logics ontologies. *Journal of Artifical Intelligence*, *48*, 885–922. doi:10.1613/jair.4058

Francopoulo, G. (2013). *LMF Lexical Markup Framework* (G. Francopoulo & P. Paroubek, Eds.). Wiley Online Library.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2), 199–220.

Guarino, N. (1998). Formal ontology in information systems. In *Proceedings of the first international conference on Formal Ontology in Information Systems (FOIS'98)* (Vol. 46). IOS Press.

Guarino, N., Oberle, D., & Staab, S. (2009). What is an ontology? In *Handbook on ontologies* (pp. 1–17). International handbooks on information systems (E2). Springer.

Hayashi, O., Miyajima, T., Nomura, M., Egawa, K., Nakano, H., Sanada, S., & Satake, H. (Eds.). (1982). *Zūsetsu nihongo: Gurafu de miru kotoba no sugata [Graphic Japanese: State of vocabulary seen in graphs]*. Tokyo: Kadokawa Shojiten.

Huang, C.-R., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., & Prévot, L. (2010). *Ontology and the lexicon: A natural language processing perspective* (C.-R. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari, & L. Prévot, Eds.). Studies in Natural Language Processing. Cambridge University Press.

Isahara, H., Bond, F., Kanzaki, K., Uchimoto, K., Kuroda, K., Kuribayashi, T., … Torisawa, K. (2012). Japanese WordNet. Retrieved May 30, 2013, from http://nlpwww.nict.go.jp/wn-ja/index.en.html

Joyce, T., Hodošček, B., & Nishina, K. (2012). Orthographic representation and variation within the Japanese writing system: Some corpus-based observations. *Written Language & Literacy*, *15*(2) Special Issue on Units of Language – Units of Writing, 254–278. doi:10.1075/wll.15.2.01rob

Joyce, T., Masuda, H., & Ogawa, T. (2014). Jōyō kanji as core building blocks of the Japanese writing system: Some observations from database construction. *Written Language & Literacy*, *17*(2), 173–194. doi:10.1075/wll.17.2.01joy

Kindaichi, K., Yamada, T., Shibata, T., Sakai, K., Kuramochi, Y., & Yamada, A. (2011). *Shinmeikai Kokugo Jiten (Shinmeikai Japanese-Japanese dictionary)* (7th edition). Tokyo: Sanseido.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., … Den, Y. (2013). Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, 1–27. doi:10.1007/s10579-013-9261-0

McCrae, J., Spohr, D., & Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. In *The semantic web: research and applications* (pp. 245–259). Springer.

Morohashi, T. (2000). *Daikanwajiten [Comprehensive Chinese-Japanese dictionary]* (Vols. 13). Tokyo: Taishukan.

Nation, I. (2001). *Learning vocabulary in another language*. Cambridge Applied Linguistics. UK: Cambridge University Press.

Nation, I. (2013). *Learning vocabulary in another language* (2nd edition). Cambridge Applied Linguistics. UK: Cambridge University Press.

Ohara, K. H., Fujii, S., Ohori, T., Suzuki, R., Saito, H., & Ishizaki, S. (2004). The Japanese FrameNet project: An introduction. In *Proceedings of LREC-04 Satellite Workshop "Building Lexical Resources from Semantically Annotated Corpora" (LREC 2004)* (pp. 9–11).

Oltramari, A., Vossen, P., Qin, L., & Hovy, E. (2013). *New trends of research in ontologies and lexical resources: Ideas, projects, systems*. Springer.

Peters, W., Montiel-Ponsoda, E., & Cea, G. A. D. (2007). Localizing Ontologies in OWL. In *Proceedings of the OntoLex07 Workshop (held in conjunction with ISWC'07*.

Prévot, L., Huang, C.-R., Calzolari, N., Gangemi, A., Lenci, A., & Oltramari, A. (2010). Ontology and the lexicon: a multidisciplinary perspective. In C.-R. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari, & L. Prévot (Eds.), *Ontology and the lexicon: a natural language processing perspective* (pp. 3–24). Studies in Natural Language Processing. Cambridge University Press.

Shinmura, I. (2008). *Kōjien (Japanese dictionary)* (6th edition). Tokyo: Iwanami Shoten.

Spohr, D. (2012). *Towards a multifunctional lexical resource: Design and implementation of a graph-based lexicon model*. Walter de Gruyter.