Constructing an ontology and database of Japanese lexical properties

Handling the orthographic complexity of the Japanese writing system

Terry Joyce, Bor Hodoščék and Hisashi Masuda
Tama University, Japan / Osaka University, Japan / Hiroshima Shudo University, Japan

As a significant milestone within ongoing efforts to construct a comprehensive database in the form of a lexical resource (LR) of Japanese Lexical Properties (JLP-LR), this paper outlines the initial construction of an Ontology of Japanese Lexical Properties (JLP-O) (Joyce & Hodoščék 2014), and, in particular, describes some of its key aspects specifically incorporated in order to satisfactorily handle the orthographic complexity of the Japanese writing system (Joyce 2013, 2016; Joyce, Hodoščék & Nishina 2012). While motivated primarily by issues of orthographic representation for the Japanese lexicon, these key features potentially have wider implications for the effective construction of integrated orthographic databases and lexicons.

Keywords: ontology, database, Japanese lexical properties, Japanese writing system, orthography, lexical entry (LE), decomposition

1. Introduction

As the papers of this special issue on *orthographic databases and lexicons* vividly illustrate, while there are sound justifications for attempting to realize integrated and comprehensive orthographic databases and lexicons, there are also formidable challenges. Central among the justifications are the laudable aspirations to create accurate and comprehensive lexical resources (LRs) that can aid a wide range of researchers across the diverse language-related disciplines to advance their research into understanding language and cognition. For languages employing phonographic writing systems, such as the Roman alphabet, the core challenges are to

Written Language & Literacy 20:1 (2017), **27–51**. DOI 10.1075/wll.20.1.03joy
ISSN 1387–6732 / E-ISSN 1570–6001 © John Benjamins Publishing Company

devise systematic methods to precisely capture the underlying principles of orthographic representation – to essentially map out the grapheme-to-phoneme correspondences for a given language. In the case of the Japanese language, however, the justifications and the challenges of constructing comprehensive orthographic databases and lexicons are, arguably, even more substantial. The multiple-script Japanese writing system – with its complementary utilization of morphographic kanji, the two syllabic scripts of hiragana and katakana, and rōmaji (the Roman alphabet) – has a largely uncontestable reputation for being the world's most complicated writing system (Joyce 2013, 2016). The complexity of the Japanese writing system alone is ample vindication for endeavors to construct reliable and comprehensive LRs that can potentially serve to illuminate its intricacies. The multiple-script nature of the Japanese writing system unquestionably affords remarkable degrees of orthographic variation (Backhouse 1984; Joyce, Hodošček & Nishina 2012), which stretches the very notion of orthographic representation to extremes, on the one hand, and seriously accentuates the challenges of constructing orthographic databases and lexicons for the Japanese language, on the other hand. Beyond the fundamental tasks of attempting to adequately encapsulate the full range of principles of orthographic representation – from the morphographic principle of kanji, the syllabographic principle of hiragana and katakana, to the phonemic principle of rōmaji – it is also essential to represent the multifarious connections between orthographic and lexical variants, as well as documenting other information such as corpus attested frequencies and contexts.

Against that background, this paper reports on a significant development within an ongoing project to create a comprehensive database in the form of a lexical resource of Japanese Lexical Properties (JLP-LR) as an integrated orthographic database and lexicon (see Joyce, Masuda & Hodošček (2016) for a general summary of the project).[1] More specifically, after cursorily noting some of the foundational work, within this introduction section, and after briefly outlining the initial construction of an Ontology of Japanese Lexical Properties (JLP-O) (Joyce & Hodošček 2014) in Part 2, the main locus of the paper, in Part 3, describes some of the JLP-O's key characteristics incorporated specifically in order to satisfactorily handle the orthographic complexity of the Japanese writing system. These key features of special relevance for an orthographic database and lexicon of the Japanese language include (1) the inclusion of both a character lexical entry (LE) class and a character module, (2) encoding the distinction between the lemma

[1] In contrast to this paper's specific emphasis on how the database copes with the complexity of the Japanese writing system (Joyce, Hodošček & Masuda 2014a), Joyce et al's (2016) overview summary focuses more on two database components not previously described in any detail in an English-language paper.

and its orthographic variants, and (3) utilizing decomposition strategies for orthographic, phonological and morphological information. As Part 2 describes in a little more detail, the JLP-O represents a pivotal transition for the large-scale database project because it establishes a valuable guiding framework for the database construction work. It is beneficial not only for its potential to facilitate the efficient integration of existing LRs using natural language processing (NLP) techniques, but also for its utility as a tool for evaluating the theoretical and psychological validities of lexical properties (Joyce & Hodošček 2014).

While sharing similar motivations with earlier endeavors at database construction, such as Joyce (2005) and Masuda and Joyce (2005), our more recent and current ventures in this direction can, to some extent, be traced back to Joyce et al's (2012) creation of corpus word lists extracted from the Balanced Corpus of Contemporary Written Japanese (BCCWJ; Maekawa, Yamazaki, Ogiso, Maruyama, Ogura, Kashino, Koiso, Yamaguchi & Den 2013).[2] Within their discussions, Joyce et al. (2012) first addressed some issues of far-reaching implications for orthographic databases and lexicons for Japanese. Of particular importance given the high incidences of homophones within Japanese, one multi-dimensional issue is the treatment of lemmas, word forms and orthographic forms, and their grouping within the BCCWJ, which have certainly been factors behind the JLP-O's distinction between `canonicalForms` and `orthographicForms`, as described in more detail in Section 3.2. Another crucial issue relates to BCCWJ's elusive demarcation between short-unit words (SUWs) and long-unit words (LUWs). As Joyce, Masuda & Ogawa (2014) observe, although the short-long contrast undoubtedly conjures up notions of length, the distinction is, in reality, essentially one of lexical status, as SUWs cover both bound morphemes and simple words (dictionary headwords) and LUWs are complex words and phrases. This conceptualization undeniably influenced the specification of the JLP-O's classes of LEs, as explained further in Part 2. The corpus word lists are effectively incorporated within the JLP-O as its primary corpus lexicon, although, as elaborated further in Part 2, all lexical information is now organized primarily according to the JLP-O's specification of five classes of LEs.[3]

[2] The BCCWJ is an authoritative sampling of contemporary Japanese written language, consisting of approximately 100 million word tokens compiled by the National Institute of Japanese Language and Linguistics.

[3] For both SUWs and LUWs, Joyce et al. (2012) compiled 14 corpus word lists based on the parts-of-speech (POS) tagging of UniDic (Den, Ogiso, Ogura, Yamada, Minematsu, Uchimoto & Koiso 2007) – the morphological analyzer dictionary developed for the BCCWJ. Including proper nouns, the lists comprised 175,708 SUW and 2,396,515 LUW word types. The corpus word lists also include information for a variety of extracted and computed lexical properties.

However, the core impetus behind the database project came with Joyce, Masuda et al. (2014) that sketched out foundational work on three key database components. The first database component consists of orthographic, phonological, morphological and semantic properties, as well as metadata (such as cross-references and pedagogical information) for the 2,136 kanji of the official jōyō kanji list; the main building blocks of orthographic representation of the Japanese lexicon (see Joyce, Masuda et al. (2014) for discussion of the 2010 revision of the jōyō kanji list (Bunkachō 2010)). The second database component resulted from analyzing the internal structures of the jōyō kanji and the 2,965 Japanese Industrial Standard (JIS)[4] level 1 kanji according to the three basic configurations of left-right, top-bottom, and enclosure-enclosed. The analysis revealed that 1,951 (91.3%) jōyō and 2,747 (92.6%) JIS1 kanji conform to these configurations and also identified 1,072 components for jōyō and 1,290 components for JIS1 kanji. The third database component, at the lexical level, appended orthographic codes, consisting of letter coding of script type (i.e., C = kanji, H = hiragana, K = katakana) for each character of words, to the corpus word lists and the headwords of Shinmura's (2008) *Kōjien* dictionary. In addition to revealing an immense variety of orthographic codes, summary data for both types and tokens confirmed that although single hiragana symbol words, such as the grammatical particle words, are the most common by token counts, the most frequent orthography types are kanji-orthography words, with two-kanji dominating within the SUW lists and four-kanji compound words being the most frequent for the LUW lists.

Notwithstanding the ambitious aspirations towards ultimately realizing a large-scale integrated JLP database tangible within Joyce, Masuda et al. (2014), still, major challenges for the project relate to how to develop the database by integrating existing LRs while maintaining the coherence essential for realizing multifunctional search capabilities. However, given that these challenges are deeply interconnected, Joyce and Hodošček's (2014) initial work on constructing an ontology of Japanese lexical properties (JLP-O) seems to represent an elegant solution to both challenges. Accordingly, Part 2 next turns to outline the JLP-O in a little more detail, as further vital background to Part 3 which describes the specific aspects incorporated to capture the complexity of the Japanese writing system.

[4] The Japanese Industrial Standard (JIS) defines the character sets used for electronic communication, and the core set consists of 2,965 level 1 (JIS1) and 3,390 level 2 (JIS2) kanji. Although most jōyō kanji are JIS1 kanji, under the 2010 jōyō kanji revision, 30 are JIS2 kanji.

2. Ontology of Japanese lexical properties (JLP-O)

As alluded to already, it is important to keep in mind that constructing comprehensive orthographic databases and lexicons is not really an end in and of itself. Rather, it is only a means – albeit an extremely promising one – of facilitating broader investigations of language and cognition. Thus, the direct consequences of regarding the target JLP database as primarily both lexical model and research tool are in directing attention to the somewhat oppositional desideratum of being as comprehensive as possible while delivering multifunctional search capabilities.

Thus, as a particularly appealing solution to both these criteria, Joyce and Hodošček (2014) commenced construction of the JLP-O; a step akin to recent NLP trends towards merging LRs, such as WordNet and FrameNet, with ontologies, such as SUMO (Huang, Calzolari, Gangemi, Lenci, Oltramari & Prévot 2010; Oltramari, Vossen, Qin & Hovy 2013). Consistent with general definitions within the areas of computer science and knowledge engineering (Guarino 1998; Guarino, Oberle & Staab 2009; Prévot, Huang, Calzolari, Gangemi, Lenci & Oltramari 2010), most concisely, an ontology is a formal specification of a shared conceptualization. Elaborating slightly, formal specification can be interpreted as a commitment to representing in a machine-readable format and shared conceptualization refers to knowledge about systems or entities, such as the components and their relations, for which there is a general consensus among the relevant community. Joyce and Hodošček cited two significant reasons for constructing the JLP-O. The first is highly pragmatic and straightforward and comes from recognizing that there are many advantages to adopting a formal specification. These substantial advantages include compatibility with NLP and knowledge tools, which can, in turn, greatly facilitate the integration of existing databases and provide means of checking and maintaining consistencies, and the fact that such formal specification is a prerequisite for being able to realize powerful query functionality. In contrast, the second reason is rather more nuanced, and more skeptical in nature, but it is very important nevertheless. As attempts to clearly elucidate phenomena such as complex systems and entities, it is not difficult to understand why ontology construction is an appealing research endeavor – it seems to be the very epitome of academic scholarship. It is, however, crucial to remember that natural systems, such as language in particular, rarely conform totally to ontological standards of completeness. In striving to balance these aspects of ontology construction, Joyce and Hodošček sought to, simultaneously, both fully leverage the JLP-O's utility for database construction and its potential to highlight issues of theoretical adequacy and psychological reality when considering candidate lexical properties for inclusion.

Given the explicit motives for constructing the JLP-O – to serve as a guiding framework for the large-scale database where lexical entries (words) are linked together according to various linguistic and psycholinguistic conceptualizations (lexical properties) – it was, naturally, crucial to have an ontology model that is appropriate for linguistic resources. Accordingly, Joyce and Hodošček (2014) drew inspiration most directly from two sources; the *lemon* model (Lexical Model for Ontologies)[5] and Spohr's (2012) model for multifunctional LRs. The lemon model has a number of advantages of relevance for JLP-O, including that it is based on the Resource Description Framework (RDF)[6] as a widely-used web standard that facilitates link representation, that domain-specific information is delegated to external sources, it is relatively concise with few classes, and that information is organized in terms of separate modules which allows for greater construction flexibility. In contrast, being more orientated towards lexicographical issues, Spohr's (2012) particular emphases on both the informational needs of diverse users and the development of appropriate query and display interfaces have also been highly instructive for the larger project's goal of realizing powerful search capabilities.

Joyce and Hodošček (2014) focused primarily on solving two foundational issues for the JLP-O; namely, specifying the structural organization of the lexical properties data and determining its range of LEs. Naturally, the first issue of property specification is crucial in order to realize a dynamic database rather than just a static lexicon list, and is, thus, clearly not a trivial matter. Traditionally, printed dictionaries have been the primary sources of lexical information, which for Japanese, include both language dictionaries, such as Shinmura's (2008) *Kōjien* and Yamada, Shibata, Sakai, Kuramochi, Yamada, Ueno, Ijima and Sasahara's (2011) *Shinmeikai Kokugo Jiten*, and character dictionaries, such as Morohashi's (2000) *Daikanwajiten*. However, in pursuing their research, linguists and psycholinguists require ready access to a steadily expanding range of contemporary lexical information beyond mere static dictionary word listings. One can start to gain some inkling of the sheer range of lexical properties of interest by firstly thinking of Nation's (2013) influential categorization of nine kinds of word knowledge, as effectively mapping out the broad domains, and, then, by imagining the kinds of detailed analyses being conducted within each area, like, for example, Adelman's (2012) listing of 14 potentially confounding lexical variables within the realm of visual word recognition research alone. While the issues of how to organize different categories of lexical properties are obviously of considerable significance, as already noted, one of the particularly appealing aspects of the lemon model

[5] http://lemon-model.net/

[6] https://www.w3.org/RDF/

is its notion of modules, which opens up flexible approaches to structuring the lexical properties and realizing the complex connections among the database's LEs. Therefore, Joyce and Hodošček adopted the same basic module strategy in specifying the JLP-O's six basic modules of character, orthographic, phonological, morphological, semantic and use related lexical properties, as illustrated in Table 1.[7] As later explanations clarify further, the term module here refers to both groupings of lexical properties and various forms of metadata about lexical properties. Although Joyce and Hodošček's initial JLP-O specification included 65 lexical properties, our subsequent discussions expand the range of lexical properties and sub-modules still further.

Table 1. JLP-O's six basic modules, with some examples of lexical properties and some initial sub-modules

Modules	Example properties
Character	type, JIS specifications, status, stroke counts, various radical classifications, various cross-references, …
Orthographic	representation, variations, length (in characters), neighborhood data, …
Phonological	stress, length, homophones, neighborhoods, consistency, … mora sub-module: code, type, C+V coding, transcriptions …
Morphological	constituent analysis, families (size/frequency), transparency, … conjugation sub-module: type …
Semantic	denotation, connotations, sense range, lexical stratum, groups, concreteness, relations, …
Use	frequency/familiarity data, collocations, grammatical patterns, genre/register/style, …

The second foundational issue addressed by Joyce and Hodošček (2014) was to determine an appropriate range of LEs for the JLP-O. Involving a number of conflicting constraints, this is understandably a thorny topic for any lexical database, but it is a particularly knotty problem for highly-agglutinative languages like Japanese with ambiguous word boundaries. Harking back to the BCCWJ's distinction between SUWs and LUWs, although their prioritizing of SUWs is not without some merits (Joyce et al. 2012), in terms of enhancing the search capabilities of the database, a wider range of LEs is undoubtedly preferable (Spohr 2012). Thus, in order for the JLP-O to more faithfully represent the Japanese lexicon and facilitate the eventual realization of powerful search capabilities, Joyce and Hodošček opted to specify five classes of `LexicalEntry`. As the first class of

[7] It may be noted, however, that JLP-O's six modules are not attempting to blindly replicate lemon's five modules or even Nation's (2013) nine aspects of word knowledge.

Character LEs is an essential aspect of realizing an orthographic database of the Japanese language, it will be discussed further in Section 2.1. The second class of `BoundUnit` includes prefixes (e.g., 御- /go-/ or /o-/ honorific prefix), affixes (e.g., -的 /-teki/ adjectival '-ic'), and auxiliary verbs (e.g., -れる /-reru/ passive conjugation). The third and fourth classes of `SimpleWord` and `ComplexWord`, respectively, encompass most of the main word classes, such as nouns (e.g., `SimpleWord`: 報告 /hō-koku/[8] 'report'; `ComplexWord`: 報告書 /hō-koku-sho/ 'report (document)'), verbs (e.g., `SimpleWord`: 読む /yo.mu/ 'read'; `ComplexWord`: 読み始める /yo.mi.haji.meru/ 'begin to read'), and adjectives (e.g., `SimpleWord`: 詳しい /kuwa.shii/ 'detailed'; `ComplexWord`: 詳しくない /kuwa.shikunai/ 'not detailed'). The fifth class is of `MultiWordExpression`.

Having developed satisfactory solutions to these two foundational issues for the JLP-O, Joyce and Hodošček (2014) also executed an extraction and RDF encoding program to create the BCCWJ-based corpus lexicon as a central pillar for constructing the comprehensive JLP database.[9] In total, the corpus lexicon consists of approximately 2.7 million LEs according to the JLP-O's core LE classes.[10]

As briefly explained already, within their initial work of establishing the JLP-O, Joyce and Hodošček (2014) draw inspiration from both the lemon model and Spohr's (2012) model for multifunction LRs. However, given that both have been developed primarily for alphabet-using languages, such as English and German, it has also been necessary to incorporate a number of key features in order to satisfactorily handle the orthographic complexity of the Japanese writing system (Joyce 2013, 2016; Joyce et al. 2012). Accordingly, Part 3 next turns to the description of those features.

3. Handling aspects of the Japanese writing system

As acknowledged already, the Japanese writing system has something of a notorious reputation for being the world's most complicated writing system (Joyce 2013).

[8] Whenever phonological glosses are provided, hyphens are used to mark kanji-kanji boundaries and periods are used to indicate kanji-hiragana boundaries.

[9] We realize that there are potentially a number of negative aspects to overly relying on the corpus lexicon within the construction of the comprehensive JLP-LR (such as treatment of proper nouns and long-tail distributions where the vast majority of LEs have very low token frequencies), but, equally, it is an important sampling of the contemporary written Japanese language that greatly enhances the database project.

[10] At present, the JLP database does not have LEs of the `MultiWordExpressions` subclass, because although it is possible to extract collocational and idiom data from the BCCWJ, Joyce and Hodošček (2014) elected to create these LEs in the future when integrating other appropriate LRs.

Most simply, the kernel of the Japanese writing system's complexity lies in the fact that the relationships between semantics, orthography and phonology are overwhelmingly many-to-many (Joyce et al. 2012). Given that such a bold assertion warrants a few words of support, at this juncture, it is worth expanding a little more on the issues of profound repercussions for Japanese orthographic databases and lexicons, noted briefly within the introduction. To some extent, Joyce et al's (2012) discussions of these issues represent reactions to policy revisions for UniDic, the morphological analyzer dictionary created within the BCCWJ project, concerning the treatment of lemmas, BCCWJ's distinction between SUWs and LUWs, and the three basic levels of lexical entries, as outlined in Ogura, Ogiso, Koiso, Hara and Miyauchi (2010). While fully accepting the importance of developing a consistent policy for accurately tagging the high instances of homophones within the BCCWJ corpus, by the same token, practical solutions for the needs of a morphological analyzer do not necessarily satisfy the requirements of other researchers, such as lexicographers, policy makers, and Japanese language instructors (Joyce et al. 2012). For instance, even Ogura et al's (2010) policy to treat あう /au/ homophones as belonging to just two separate lemmas involves some degree of blurring semantic nuances. While a basic distinction between 合う 'fit, match, agree with, be correct' and 会う 'meet, encounter' is reasonable, in also grouping 逢う 'meet, encounter (date or tryst nuance)', 遭う 'meet, encounter (undesirable nuance)' and 遇う 'meet, encounter (unexpected nuance)' as just separate orthographic forms of the 会う lemma, there is also, undeniably, some degree of sacrifice, or relegation, of the semantic distinctions that different kanji orthography representations convey, which may be less suitable for some research purposes.

Table 2. UniDic entry consisting of 3 basic levels

語彙素 /go-i-so/ **'lemma'**	**語形 /go-kei/** **'word forms'**	**書字形 /sho-ji-kei/** **'orthographic forms'**
矢張り (adverb) "also; as I thought; still; in spite of; absolutely; of course"	ヤハリ /yahari/	やはり 矢張り
	ヤッパリ /yappari/	やっぱり 矢っ張り
	ヤッパ /yappa/	やっぱ

Note. Taken from Joyce et al. (2012: 262) as adapted from Ogura et al. (2010: 86)

Directly related to UniDic's treatment of lemmas, its lexical entries are organized according to three basic levels; a lemma level, a word-forms level (distinguishing pronunciation variants), and an orthographic-forms level (distinguishing orthographic variants), as presented in Table 2. It is worth remarking that although Ogura et al's (2010) examples are nicely illustrative of the core issues – 会う as a case of subsuming sense distinctions traditionally differentiated by different kanji as sense nuances of a single lemma and 矢張り as a case of phonological and orthographic variation – neither are exceptionally extreme in nature. However, they suffice to evidence the substantial challenges for constructing orthographic databases and lexicons to adequately capture the complexity of the Japanese writing system. The following sub-sections introduce some of JLP-O's key features specifically incorporated in response to these challenges.

3.1 Character LEs and character module

As astute readers have undoubtedly already discerned from our earlier outline of the initial specifications of the JLP-O within Part 2, JLP-O has both a lexical entries class for characters and a character module, where, again, the notion of modules within JLP-O refers to a grouping of related lexical properties and various forms of metadata about those lexical properties. We readily acknowledge that the inclusion of a separate LE class for characters may, at first encounter, seem somewhat superfluous and, arguably, not totally consistent with the notion of a lexical database. However, given the significant role of kanji as the core orthographic building blocks for much of the Japanese lexicon (Joyce et al. 2012), we believe that this is well-motivated for realizing a complete orthographic database, and the incorporation of both a character sub-class and a character module provide flexible and complementary perspectives on the database.

In sharp contrast to the highly constrained character sets of phonographic writing systems, undeniably, a direct and profound consequence of the morphographic principle is that one requires a considerably larger inventory of characters. However, this immediately poses fundamental questions about how many kanji, as well as which kanji, should be included, as one seeks to construct a comprehensive orthographic database of the Japanese language. These questions are also far from straightforward and a number of criteria need to be carefully considered. On the one hand, as the culmination of a series of guidelines concerning kanji usage issued by the Japanese government since the mid-20th century, the official jōyō kanji list – revised in 2010 to include 2,136 jōyō kanji – effectively represents a de facto minimum standard for functional literacy within Japan (Joyce, Masuda et al. 2014), that must be covered by any orthographic database. On the other hand, the legacies of former kanji usage conventions involving larger character inventories have not simply disappeared because of the introduction of the jōyō kanji guidelines. Rather, many non-jōyō kanji continue to be regularly used in daily life because of their great cultural significance; not least of which is as the components of many Japanese family and place names. This means that educated Japanese people are expected to know considerably more kanji than just the 2,136 jōyō kanji, and, clearly, any orthographic database of the Japanese language must also reflect this reality by including more complete character sets.

The JIS organization has defined a number of character set standards for electronically-mediated Japanese language communication. Foundational among them is the JIS X 0208 character encoding that consists of 2,965 level 1 (JIS1), 3,390 level 2 (JIS2) and 524 other kanji for a total of 6,879 kanji. JIS have also defined further character encoding standards, including JIS X 0212 which adds 5,801 level 3 (JIS3) kanji to the JIS X 0208 set for a total of 12,156 and the JIS X 0213 2004 character encoding of 11,272 kanji proposed as a replacement for JIS X 0212. In terms of jōyō kanji coverage, while the majority are JIS1 kanji, 30 of the 196 kanji added under the 2010 revision are JIS2 kanji. Moreover, looking at type and token frequency data for JLP-O's corpus lexicon, given that newspapers and official documents generally conform to the jōyō kanji guidelines, it is not surprising that the jōyō kanji represent approximately 96% of all kanji tokens.[11] However, given the continuing significance of many non-jōyō kanji, it is also little surprise that the jōyō kanji only represent approximately 33% of the kanji types (Joyce, Masuda et al. 2014).

Nevertheless, the implications for attempting to construct a comprehensive orthographic database for the Japanese language could not be clearer. Although jōyō kanji represent the core components for the orthographic representation of contemporary vocabulary, they alone are far from sufficient, and in order to adequately cover all kanji of relevance for contemporary written Japanese, it is prudent to incorporate all the kanji of the JIS X 0213 2004 character encoding.[12] However, because the JIS character sets essentially only provide a listing of the kanji and their reference codes, we have also drawn on the KANJIDIC2[13] project's consolidated XML-format kanji database of 12,847 kanji to establish a number of extra data fields for subsequent verification and consolidation. Accordingly, based on the union of the 6,781 character LEs of the corpus lexicon, the JIS X 0213:2004

[11] It is worth stressing that, as the JLP-O's corpus lexicon was extracted from the BCCWJ corpus, the frequency data reflect the BCCWJ's five year sampling period of 2006–2011, and that the 2010 revision to the jōyō kanji list was implemented more towards the end of that sampling period. Thus, it is likely that token counts for newly added jōyō kanji will increase slightly as the revised list comes to exert more influence over contemporary written Japanese, but comparisons with the prior jōyō kanji list from 1981 also indicate that many of the additional kanji were already common, because of their usage in Japanese family and place names.

[12] Naturally, the JLP-LR will provide as much practical information about corpus usage and source coverage for the character LEs, to help database users discern for themselves the significance of any given kanji character.

[13] http://www.edrdg.org/kanjidic/kanjd2index.html

kanji and the KANJIDIC2 kanji, the JLP-LR contains 13,888 character LEs, as shown in Table 3 which presents both type and token counts for all the current lexical entries of the JLP-LR.

Table 3. The lexical entries currently distinguished within the JLP-LR

Lexical entry	Types	Tokens
`Characters`	13,888	195,500,491
`BoundUnit`	433	11,327,729
`SimpleWord`	195,380	112,557,387
`ComplexWord`	2,438,506	101,684,786

A natural corollary of having a large set of kanji characters is simply that there is also a larger body of metadata associated with kanji. Within the JLP-LR, the various links and inter-connections between characters LEs and between the various forms of metadata are mediated by the character module, which effectively serves as a reference inventory of relevant lexical properties and their candidate ranges of values. While not every lexical property associated with the character module will be relevant to every character LE, JLP-O's character module will eventually include specifications for all orthography-related lexical properties incorporated within the JLP-LR. For instance, starting from the most basic of lexical properties, all character LEs have a type specification (such as C for Chinese characters, H for hiragana, K for katakana, R for rōmaji and S for symbol), their JIS character encoding specifications (both JIS level and reference), as well as stroke count information (for both kana and kanji). Moreover, all character LEs for kanji also include information about the status of the kanji, such as whether it is included within the set of jōyō kanji, and, if so, at which instruction grade.[14] In addition to including a variety of cross-references to character standards, such as Unicode, and various dictionaries, all kanji LEs will also include a variety of classifications relating to their internal structures, from the traditional classification of the principal radical components to numerous alternative classification schemes that have been proposed, as well as all associated information, such as radical names.

While we accept that the inclusion of a separate LE class solely for characters may initially seems rather at odds with our ultimate objective of realizing a comprehensive lexical database, as this section has sought to explain, we believe the move is fully justified given both the extensive numbers of kanji still in

[14] The jōyō kanji list also consists of a sub-list of 1,006 kyōiku kanji 'education kanji' that are assigned for instruction across the six grades of elementary school, with the remaining 1,130 kanji being introduced across the remaining three grades of compulsory education at junior high-school.

contemporary use and the wealth of information associated at the level of Japanese orthographic symbols. Moreover, recognizing the substantial benefits in terms of utilizing more flexible construction strategies, we also regard the specification of the character module as a valuable way of integrating the complex network of interrelationships between the character LEs and the other LEs of the database.

3.2 `canonicalForm` and `orthographicForm`

As Joyce et al's (2012) quantitative analyses of their corpus word lists vividly demonstrate, orthographic variation is a pervasive characteristic of the Japanese writing system, particularly in terms of the orthographic representation of the most common words of the Japanese lexicon. For instance, the distributions of orthographic variations for the most frequent 100 lemmas across the four main word classes of nouns, verbs, adverbs and i-adjectives revealed the mean number of variations to be 8.44 for SUWs and 5.80 for LUWs, with ranges between 1–34 and between 1–28, respectively. Naturally, this malleable aspect of Japanese orthographic representation must be efficiently incorporated with the Japanese orthographic and lexicon database.

The JLP-O's strategy for capturing this feature of the Japanese writing system is essentially an amalgamation of the BCCWJ's distinction between the lemma and its orthographic forms, as illustrated in Table 2, and lemon's demarcation between its `canonicalForm` and `otherForm` sub-properties. Notwithstanding our reservations about UniDic's treatment of certain lemmas, which, as outlined earlier, may not always be totally consistent with traditional nuances of meaning,[15] the move to maintain a core distinction between the lemma, as the more abstract LE, and its orthographic variants, as a listing of its various orthographic representations, undoubtedly has much merit as a way of handling the high degrees of orthographic variation inherent within the Japanese writing system. Moreover, although stemming from a quite different motivation of encoding the syntactic variant relationship between 'animal' and 'animals', we regard lemon's approach to differentiating syntactic variants from preferred forms through its use of two sub-properties of form, namely, `canonicalForm` and `otherForm`, as providing a very viable way of formally specifying the phenomenon of Japanese orthographic variation within the JLP-O. Accordingly, JLP-O's solution is to adopt the sub-property of `canonicalForm` to refer to the standard orthographic representation of the lemma, while

[15] The issues of determining when word senses are sufficiently divergent to warrant separate headword entries or whether they can be assumed as nuances of a core sense are naturally fundamental to lexicographic endeavors. It is, accordingly, something that the larger database project will continue to consider carefully in further developing the JLP-LR.

changing the label of the second sub-property of `otherForm` to `orthographicForm` that records all orthographic variants of a LE. Figure 1 presents the relevant section for the `simpleWord` LE of 始める /haji.meru/ 'to begin', which includes eight `orthographicForm` variants of the `canonicalForm`.

```
jlpo:始める_動詞-非自立可能 a jlpo:SimpleWord;
   lemon:canonicalForm
       [lemon:writtenRep "始める"@ja];
   jlpo:orthographicForm [lemon:writtenRep "始める"@ja];
   jlpo:orthographicForm [lemon:writtenRep "はじめる"@ja];
   jlpo:orthographicForm [lemon:writtenRep "初める"@ja];
   jlpo:orthographicForm [lemon:writtenRep "肇む"@ja];
   jlpo:orthographicForm [lemon:writtenRep "創める"@ja];
   jlpo:orthographicForm [lemon:writtenRep "始む"@ja];
   jlpo:orthographicForm [lemon:writtenRep "はじめる"@ja];
   jlpo:orthographicForm [lemon:writtenRep "始める"@ja].
```

Figure 1. Section of JLP-O specification for the `simpleWord` LE of 始める /haji.meru/ 'to begin' highlighting its `canonicalForm` and `orthographicForm` properties.

3.3 Forms of decomposition

In order to both adequately handle the complexity of the Japanese writing system and to ultimately realize a comprehensive orthographic and lexicon database with powerful search capabilities, the third broad tactic being employed within the JLP-O is to utilize as fully as possible a strategy of decomposing properties into their components. The basic tactic is akin in spirit to lemon's decomposition object property that is used within lemon to mark morphological and phrase components. However, JLP-O is currently using the approach for three related purposes: namely, orthographic, phonological and morphological decompositions that are all key for elucidating the network of many-to-many interconnections that contribute to the complexity of the Japanese writing system. The following subsections briefly describe these three forms of decomposition.

3.3.1 *Orthographic decomposition*

The first of the three decompositions currently realized within the JLP-O is that of orthographic decomposition, which has been the simplest to implement as it draws directly on the sub-class of character LEs, which, as outlined earlier within Section 3.1, are part of the JLP-O's specifications together with a separate character module of metadata about the character LEs. Reflecting the morphographic nature of kanji (Joyce 2013), and the fact that jōyō kanji are the principal building blocks for the orthographic representation for much of the Japanese lexicon (Joyce, Masuda et al. 2014), clearly, it would be effectively impossible to conduct any meaningful analyses of the orthographic structures of the Japanese lexicon without recourse to some reference listing of their orthographic components. Within the JLP-O, the sub-class of character LEs essentially functions as a master reference listing of all the orthographic elements of the Japanese writing system, which is vital for both conducting orthographic analyses and for realizing powerful querying of the orthographic database.

Accordingly, all the `orthographicForms` for all LEs (apart from the character LE class itself) are decomposed into their component characters in terms of identification links to the respective character LEs. Figure 2 presents a section from the JLP-O's specification of the `simpleWord` LE of 始める which shows the `orthographicDecomposition` for two of its `orthographicForms` (the other variants are omitted for the sake of brevity).

```
jlpo:始める_動詞-非自立可能 a jlpo:SimpleWord;
    lemon:canonicalForm [lemon:writtenRep "始める"@ja];
    jlpo:orthographicForm [
      lemon:writtenRep "始める"@ja;
      jlpo:orthographicDecomposition (
        [jlpo:Character jlpo:始_character]
        [jlpo:Character jlpo:め_character]
        [jlpo:Character jlpo:る_character])];
      jlpo:orthographicForm [
      lemon:writtenRep "はじめる"@ja;
      jlpo:orthographicDecomposition (
        [jlpo:Character jlpo:は_character]
        [jlpo:Character jlpo:じ_character]
        [jlpo:Character jlpo:め_character]
        [jlpo:Character jlpo:る_character])]; […].
```

Figure 2. Section of JLP-O specification for the `simpleWord` LE of 始める highlighting the `orthographicDecomposition` of two of its `orthographicForms`

3.3.2 *Phonological decomposition*

The second kind of decomposition incorporated within the JLP-O is that of phonological decomposition. As the name implies, in addition to the orthographic decompositions, all the `orthographicForms` of the LEs also include a breakdown into their basic phonological components, which, in the case of Japanese phonology, is the mora unit that is generally defined as an equal-length syllable. However, in order to implement the phonological decomposition of all `orthographicForms`, it was first necessary to create, as a sub-module of JLP-O's phonological module, a mora module both to serve as a comprehensive listing of all Japanese mora and to facilitate the organization of all relevant metadata.

Even though the traditional range of Japanese sounds has been expanded considerably to cope with foreign loanwords and names that could not be readily transcribed with the historical mora inventory, still, contemporary Japanese phonology is comparatively straightforward. In total, the mora module contains 168 Japanese morae, which are classified according to three main types of 71 basic, 33 contracted and 64 extended morae. The basic mora group consists of the five Japanese vowels alone (/a/, /i/, /u/, /e/ and /o/),[16] the 39 contemporary consonant-vowel (CV) combinations formed with the nine unvoiced consonants (/k/, /s/, t/, /n/, /h/, /m/, /y/, /r/, /w/),[17] the 20 CV combinations formed with the four voiced consonants (/g/, /z/, /d/, /b/), the five CV combinations formed with the one semi-voiced consonant (/p/), as well as one moraic nasal (/N/)[18] and one glottal stop of consonant gemination (促音 /soku-on/). The contracted (拗音 /yō-on/) mora group consists of the CyV clusters formed by combining 11 of the basic CV mora, where the V corresponds to /i/, with either /ya/, /yu/ and /yo/, such as /ki/ + /ya/ → /kya/, /ki/ + /yu/ → /kyu/ and /ki/ + /yo/ → /kyo/. The third main type of extended mora relates to the expansion of the traditional Japanese sound inventory to cope with foreign loanwords and names.[19] Very similar in nature to

[16] Japanese long vowels are essentially phonologically decomposed as vowel repetitions, irrespective of the various orthographic strategies employed for their representation. These strategies include for hiragana じょうよう /jōyō/, where the lengthening of both /o/ vowels are indicated by the additions of う /u/, for katakana データベース /dētabēsu/ 'database', where the lengthening of both /e/ vowels are indicated by the additions of the ー symbol, and cases of adding small kana to indicate the elongation of the vowel sounds, often in imitation of interjections and cries, such as きゃぁぁ /kyaaa/.

[17] The simple multiplication of 5 times 9 yields 45, but the number is 39 because /yi/, /ye/ and /wu/ were never distinguished from /i/, /e/ and /u/, respectively, and because the historical distinctions between /wi/, /we/, and /wo/ and /i/, /u/, and /o/, respectively, are no longer maintained. Although the hiragana symbol of を /wo/ is still employed in contemporary Japanese writing, its usage is restricted to orthographically representing the grammatical object marker which is pronounced as /o/.

[18] The moraic nasal is conventionally glossed as /N/, because although it is usually pronounced as /n/, it also becomes /m/ before /b/ and /p/, such as 新聞 /shim-bun/ 'newspaper' and 万博 /bam-paku/ 'world fair; international exposition'.

[19] Notwithstanding the considerable degrees of orthographic variation demonstrated in Joyce et al. (2012), the basic orthographic convention is for foreign loanwords and names to be represented with the katakana syllabary. Accordingly, the extended mora group is usually represented in katakana.

the conventions for traditional contracted sounds, the extended morae all involve combinations of either a basic V or CV mora with either a V alone or one of the /ya/, /yu/ and /yo/ CV combinations, such as /ku/ + /wa/ → /kwa/, /ku/ + /i/ → /kwi/, /ku/ + /e/ → /kwe/ and /ku/ + /o/ → /kwo/. As summarized in Table 4, frequency data has been computed for both the number of times each mora is an element of the JLP-LR's LEs and their token counts within the BCCWJ. The most frequent mora element of the JLP-LR's LEs is the nasal /N/, which, presumably, reflects the fact that /N/ is the final element of many two-morae Sino-Japanese morphemes and is, thus, a very common mora of compound words, such as 簡単 /kan-tan/ 'simple' where it appears twice. Moreover, while the general distributions of mora types for both JLP-LR LE types and BCCWJ tokens are fairly similar, the shift to a somewhat higher proportion of the BCCWJ tokens being basic CV-unvoiced morae (59.2%) is consistent with fact that many high-frequency functional words, such as case-marking particles, are single mora native-Japanese words, such as /no/ 'possessive' marker, /ka/ 'question-sentence' marker, /to/ 'with; whenever', and /ni/ 'location' marker.

Table 4. Mora frequency data as functions of both number of JLP-LR's LEs and BCCWJ tokens

	JLP-LR's LE types		**BCCWJ tokens**	
Mora type	**Count**	**Percentage**	**Count**	**Percentage**
Basic [Vowels]	3,897,827	22.2	35,986,361	18.7
Basic [CV-unvoiced]	8,350,264	47.5	114,070,192	59.2
Basic [CV-voiced]	1,829,365	10.4	21,567,008	11.2
Basic [CV-semi-voiced]	178,952	1.0	999,206	0.5
Basic [Nasal]	1,673,452	9.5	10,237,960	5.3
Basic [Glottal stop]	249,812	1.4	2,073,375	1.1
Contracted [CyV]	1,300,570	7.4	7,234,435	3.8
Extended	97,647	0.6	355,391	0.2
Total	**17,577,889**	**100.0**	**192,523,928**	**100.0**

The mora module also contains various forms of metadata. Beyond their classification according to the three main types (i.e., basic, contracted and extended), all morae are further sub-categorized according to their phonological features (e.g., voicing), analysis of their structures (i.e., V, C, CV, CyV and CwV), and breakdowns of their components. The module also includes information concerning the correspondences between various transcription systems that have been proposed over time.

```
jlpo:始める_動詞-非自立可能 a jlpo:SimpleWord;
    lemon:canonicalForm [lemon:writtenRep "始める"@ja];
    jlpo:orthographicForm [
      lemon:writtenRep "始める"@ja;
      jlpo:orthographicDecomposition (
        [jlpo:Character jlpo:始_character]
        [jlpo:Character jlpo:め_character]
        [jlpo:Character jlpo:る_character])];
      jlpo:phonologicalDecomposition (
        [jlpo:Mora jlpo:ha_mora]
        [jlpo:Mora jlpo:ji_mora]
        [jlpo:Mora jlpo:me_mora]
        [jlpo:Mora jlpo:ru_mora ])];
      jlpo:orthographicForm [
      lemon:writtenRep "はじめる"@ja;
      jlpo:orthographicDecomposition (
        [jlpo:Character jlpo:は_character]
        [jlpo:Character jlpo:じ_character]
        [jlpo:Character jlpo:め_character]
        [jlpo:Character jlpo:る_character])];
      jlpo:phonologicalDecomposition (
        [jlpo:Mora jlpo:ha_mora]
        [jlpo:Mora jlpo:ji_mora]
        [jlpo:Mora jlpo:me_mora]
        [jlpo:Mora jlpo:ru_mora ])] […].
```

Figure 3. Section of JLP-O specification for the `simpleWord` LE of 始める highlighting the `phonologicalDecomposition` of two of its `orthographicForms`

Figure 3 presents a section of the JLP-O specification for the `simpleWord` LE of 始める highlighting the `phonologicalDecomposition` of two of its `orthographicForms`. Although the basic character to phonology relationship is one-to-one, and generally highly consistent, in the case of the two Japanese syllabaries of hiragana and katakana, the largely arbitrary relationships between kanji characters and their phonological values are undeniably a major factor contributing to the complexity of the Japanese writing system (Joyce 2013, 2016; Joyce et al. 2012).[20] Thus, as Figure 3 illustrates, although the four hiragana characters of はじめる correspond

[20] The relationships are not perfect in the case of hiragana, because は /ha/ is pronounced as /wa/ when representing the mono-mora subject particle, へ /he/ is pronounced as /e/ when representing the mono-mora destination particle, and, as noted already, を /wo/ is pronounced as /o/ when representing the mono-mora object particle (which it is restricted to in contemporary usage). On the other hand, it should also be pointed out that the relationships between kanji and

one-to-one with the four mora of ha-ji-me-ru, the standard mixed-kanji-kana orthographic representation of 始める only consists of three graphemes, where the two mora verbal root of /haji/ is represented by the kanji 始.

3.3.3 *Morphological decomposition*

The final kind of decomposition being implemented within the current version of the JLP-LR is an initial and partial form of morphological decomposition, where all `complexWord` LEs include information about their component `boundUnit` and `simpleWord` LEs. The current level of `morphologicalDecomposition` draws primarily on the BCCWJ's annotations of their LUWs as concatenations of component SUWs. However, even to implement this preliminary level of `morphologicalDecomposition`, it has also been necessary to develop a `conjugationParadigm` module, as a sub-component of the morphology module, as a practical solution to explicitly incorporating knowledge about Japanese verb conjugations within the JLP-O. Because the BCCWJ annotations only refer to SUW lemmas, which, as explained earlier, are the basis for JLP-O's `canonicalForm` property, the annotations alone provide no explanation about the verb1 conjugations that occur within verb1+verb2 compound verbs, like 読み始める /yo.mi.haji.meru/ 'to begin to read'. While it is essentially true that 読み始める is a combination of the SUW lemma 読む /yo.mu/ 'to read' (where the SUW lemma corresponds to what is also known as the citation, or plain, form of verbs) with the SUW lemma 始める /haji.meru/ 'to begin', rather than merely decomposing 読み始める into 読む + 始める, the initial implementations of the JLP-O's `morphologicalDecomposition` also include reference to its `conjugationParadigm` module in order to explain that the verb1 element has undergone the obligatory conjugation into its conjunctive form of 読み /yo.mi/. While the impetus for creating the `conjugationParadigm` module has been to facilitate the initial `morphologicalDecomposition` of compound verbs, the module will also be particularly invaluable for representing and exploring the important lexical properties related to Japanese verbs. Figure 4 presents a section of JLP-O's specification for the `complexWord` LE of 読み始める, which highlights, as sub-properties of the `canonicalForm` property (abstract lemma representation), its `morphologicalDecomposition` into its two component `simpleWords` together with reference to the `conjugationParadigm` module.

phonology are somewhat less arbitrary in the case of 音読み /on-yo.mi/ (Sino-Japanese pronunciations) than for 訓読み /kun-yo.mi/ (Native-Japanese pronunciations).

```
jlpo:読み始める_動詞-一般 a jlpo:ComplexWord;
    lemon:canonicalForm [
      lemon:writtenRep "読み始める"@ja;
      jlpo:morphologicalDecomposition (
        [jlpo:SimpleWord [ jlpo:読む_動詞-一般; jlpo:conju….]]
        [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:orthographicDecomposition (
        [jlpo:Character jlpo:読_character]
        [jlpo:Character jlpo:み_character]
        [jlpo:Character jlpo:始_character]
        [jlpo:Character jlpo:め_character]
        [jlpo:Character jlpo:る_character])];
      jlpo:use [jlpo:frequency 139;
                 jlpo:corpus "BCCWJ"]]; […].
```

Figure 4. Section of JLP-O specification for the `complexWord` LE of 読み始める /yo.mi.haji.meru/ 'to begin to read' highlighting its `morphologicalDecomposition` into its two component `simpleWords` together with reference to the `conjugationParadigm` module.

Although the present `morphologicalDecomposition` of all `complexWord` LEs into their component `boundUnit` and `simpleWord` LEs establishes a firm foundation for investigating the morphological structures of the Japanese lexicon, we fully recognize that more work will be required to further develop the morphology module of the JLP-LR, and one of the first tasks in that direction will be to analyze all polymorphemic `simpleWord` LEs into their morphological compounds. To that aim, Joyce, Hodošček and Masuda (2014b) (see also Joyce et al. 2016) have already conducted an initial quantitative study of the three- and four-kanji compounds words within the JLP-LR. More specifically, they executed automatic analyses of 226,850 three-kanji and 337,270 four-kanji compound words to identify their word structures through recursive matching. For instance, although the four-kanji compound word of 大学入試 /dai-gaku-nyū-shi/ 'university entrance examination' has a 2+2 structure, being the product of combining 大学 'university' and 入試 'entrance examination', the four-kanji compound word of 決定論的 /ket-tei-ron-teki/ 'deterministic' has a [2+1]+1 structure, being the product of adding the adjective forming suffix 的 '-ic' to the three-kanji compound word 決定論 'determinism', which, in turn, is the product of 決定 'determine' and 論 'theory' combined. These analysis results are also being included within the JLP-LR.

4. Conclusion

The work of constructing comprehensive orthographic databases and lexicons is rather analogous to that of the toolsmith seeking to develop a unique tool that is vital for accomplishing some particular endeavor. Within this simple metaphor, the larger endeavor is, unquestionably, to aid researchers in advancing their understanding of language, but the true value of the analogy lies in highlighting the indispensible nature of the tool itself – the orthographic and lexicon database.

As an important transition within an ongoing project to construct a comprehensive and integrated orthographic database and lexicon that focuses on the lexical properties of the Japanese lexicon (JLP-LR), this paper has outlined the initial construction of the Ontology of Japanese Lexical Properties (JLP-O) (Joyce & Hodošček 2014), with particular emphasis on some of its key features incorporated specifically in order to satisfactorily handle the orthographic complexity of the Japanese writing system. As described in Part 3, the first key characteristic is to incorporate both a separate sub-class of LEs for characters and a character module. In part, reflecting the large character inventory of Japanese, with its use of morphographic kanji, the separate sub-class of character LEs greatly facilitates both the orthographic analyses of the two main sub-classes of `simpleWord` and `complexWord` LEs and mapping out the complex network of interrelationships between LEs and their lexical properties. The second key feature of providing separate specifications for `canonicalForms` and `orthographicForms` for all LEs is essential, given that orthographic variation is such a pervasive characteristic of the Japanese writing system, where the most common words of the Japanese lexicon generally have multiple orthographic representations. The third key aspect is to utilize the strategy of decomposition as extensively as possible, and the current version of the JLP-LR includes three kinds of decomposition for orthographic, phonological and morphological information. The first kind is of `orthographicDecomposition`, that relies on the existence of the sub-class of character LEs, making it possible decompose all the `orthographicForms` of all LEs into their component characters. The second kind is `phonologicalDecomposition`, which depends on the separate specification of a mora module – as part of the larger phonology module – that lists the complete set of Japanese morae and appends several kinds of metadata about their frequencies and properties. The third kind is of `morphologicalDecomposition`, where currently all `complexWord` LEs are decomposed into their component `boundUnit` and `simpleWord` LEs. While these key characteristics are all crucial for adequately capturing the complexity of the Japanese writing system, all also greatly enhance the potential for developing sophisticated querying of the JLP-LR to aid researchers in their investigations into the rich interconnections among Japanese lexical properties.

In addition to extending the present level of `morphologicalDecomposition` by also analyzing all polymorphemic `simpleWord` LEs into their morphological compounds, another future task of high priority will be to fully integrate the second database component developed in Joyce, Masuda et al. (2014) relating to the internal structures of jōyō and JIS1 kanji. Although their analysis of these kanji sets were in terms of the three basic configurations of left-right, top-bottom, and enclosure-enclosed, the task of integrating their results within the JPL-O will inevitably also involve developing and specifying a new radical module, as a subcomponent of the orthographic module, to support `radicalDecomposition` properties for all kanji character LEs. Yet another task will be to integrate the database of semantic transparency ratings for approximately 10,000 two-kanji compound words (Joyce et al. 2016; Masuda 2014; Masuda, Joyce, Ogawa, Kawakami & Fujita 2014), which, interestingly, reveals a skewed distribution in the semantic transparency ratings, where 94.4% of the survey compound words have strong relationships between the constituent meanings and the compounds meanings as indicated by ratings of 4–6 (on 6-point scale).

As outlined in Part 2, the JLP-O functions as a valuable framework to both guide and facilitate the larger project of constructing the JLP-LR as a comprehensive database of Japanese lexical properties. The degree of formal specification that an ontology like the JLP-O requires is particularly valuable both for assessing the theoretical and psychological validities of candidate lexical properties and for facilitating the efficient integration of existing LRs using NLP techniques. However, in developing the JLP-O to serve as formal specification of the lexical properties of the Japanese lexicon, it has also been absolutely essential to carefully specify effective methods of handling the highly complicated nature of the Japanese writing system.

References

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

Backhouse, A. E. (1984). Aspects of the graphological structure of Japanese. *Visible Language* 18: 219–228.

Bunkachō [Agency for Cultural Affairs]. (2010). *Jōyōkanjihyō* [Jōyō kanji list]. Available at <http://kokugo.bunka.go.jp/kokugo_nihongo/joho/kijun/naikaku/pdf/joyokanjihyo_20101130.pdf> (13 November 2016).

Den, Yasuharu, Toshinobu Ogiso, Hideki Ogura, Atsushi Yamada, Nobuaki Minematsu, Kiyotaka Uchimoto & Hanae Koiso (2007). Kōpasu nihongogaku no tame no gengo shigen: Keitaisokaisekiyō denshijisho no kaihatsu to ōyō [The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics]. *Nihongo Kagaku* [Japanese Linguistics], 22: 101–122.

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

Guarino, Nicola, Daniel Oberle & Steffen Staab. (2009). What is an ontology? In Steffen Staab & Rudi Studer (Eds.), *Handbook on ontologies* (Second edition; International handbooks on information systems) (pp. 1–17). Springer.

Huang, Chu-Ren, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, Alessandro Oltramari & Laurent Prévot (Eds.). (2010). *Ontology and the lexicon: A natural language processing perspective*. (Studies in Natural Language Processing). Cambridge: Cambridge University Press.

Joyce, Terry. (2005). Constructing a large-scale database of Japanese word associations. In Katsuo Tamaoka, (Ed.). *Corpus Studies on Japanese Kanji* (Glottometrics 10) (pp. 82–98). Hituzi Syobo: Tokyo, Japan and RAM-Verlag: Lüdenschied, Germany.

Joyce, Terry. (2013). The significance of the morphographic principle for the classification of writing systems. In Susanne R. Borgwaldt & Terry Joyce (Eds.), *Typology of writing systems* (Benjamins Current Topics 51) (pp. 61–84). Amsterdam: John Benjamins.

Joyce, Terry. (2016). Writing systems and scripts. In Andrea Rocci & Louis de Saussure (Eds.), *Verbal communication* (Handbooks of Communication Science 3) (pp. 287–308). Berlin/ Boston: De Gruyter Mouton.

Joyce, Terry, & Bor Hodošček. (2014). Constructing an ontology of Japanese lexical properties: Specifying its property structures and lexical entries. In Michael Zock, Reinhard Rapp, & Chu-Ren Huang (Eds.), *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex4)* (pp. 174–185). 23 August, 2014. Dublin, Ireland.

Joyce, Terry, Bor Hodošček & Hisashi Masuda. (2014a). Constructing an ontology and database of Japanese lexical properties: Handling the orthographic complexity of the Japanese writing system. *'Orthographic Databases and Lexicons': 9th International Workshop on Writing Systems and Literacy*, 4–5 September. University of Sussex, Brighton, UK.

Joyce, Terry, Bor Hodošček & Hisashi Masuda. (2014b). Quantitative study of 3- and 4-kanji Japanese compound words: Database extraction and automatic analysis of word structures. *15th International Conference on the Processing of East Asian Languages*, 24–26 October, 2014. Korean University, Seoul, Korea.

Joyce, Terry, Bor Hodošček & Kikuko Nishina. (2012). Orthographic representation and variation within the Japanese writing system: Some corpus-based observations [Special issue: Units of language – units of writing, edited by Terry Joyce and David Roberts]. *Written Language and Literacy*, 15(2), 254–278. doi: 10.1075/wll.15.2.07joy

Joyce, Terry, Hisashi Masuda & Bor Hodošček. (2016). Constructing a database of Japanese lexical properties: Outlining its basic framework and initial components. *Tama University School of Global Studies Bulletin*, 8, 35–60.

Joyce, Terry, Hisashi Masuda & Taeko Ogawa. (2014). Jōyō kanji as core building blocks of the Japanese writing system: Some observations from database construction [Special issue: The architecture of writing systems, edited by Kristian Berg, Nanna Fuhrhop, & Franziska Buchmann]. *Written Language and Literacy*, 17(2), 173–194. doi: 10.1075/wll.17.2.01joy

Masuda, Hisashi. (2014). *Kanjinijihyōkigokan no imitekikankeisei ni kansuru dētabēsu no kōchiku* [Constructing a database of the semantic relationships within two-kanji orthographic words]. Kagaku Kenkyū Hijo Seijigyō Kenkyū Seika Hōkokusho [Research Report for Grant-in-Aid for Scientific Research from the Japanese Society for the Promotion of Science].

Masuda, Hisashi, & Terry Joyce. (2005). A database of two-kanji compound words featuring morphological family, morphological structure, and semantic category data. In Katsuo Tamaoka, (Ed.). *Corpus Studies on Japanese Kanji* (Glottometrics 10) (pp. 30–44). Hituzi Syobo: Tokyo, Japan and RAM-Verlag: Lüdenschied, Germany.

Masuda, Hisashi, Terry Joyce, Taeko Ogawa, Masahiro Kawakami & Chikako Fujita. (2014). A database of semantic transparency ratings for two-kanji Japanese compound words. Poster presentation given at *'Orthographic Databases and Lexicons': 9th International Workshop on Writing Systems and Literacy*, 4–5 September, 2014. University of Sussex, Brighton, UK.

Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takeiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi & Yasuharu Den. (2013). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 1–27. doi: 10.1007/s10579-013-9261-0

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

Nation, I. S. P. (2013). *Learning vocabulary in another language* (Second edition) (Cambridge applied linguistics). Cambridge: Cambridge University Press.

Ogura, Hideki, Toshinobu Ogiso, Hanae Koiso, Yutaka Hara, & Sayaka Miyauchi. (2010, March). Keitaiso kaiseki jisho UniDic ni okeru goiso midashi no rikkō hōshin [Criteria for the lemmatization of UniDic], *Tokuteiryōiki kenkyū "nihongo kōpasu" heisei 21 nendo kōkai waakushoppu (Kenkyū seika hōkokukai) yokōshū* [Priority-Area Research "Japanese Corpus": Proceedings of the 2010 public workshop]. Tokyo: General Headquarters, Priority-Area Research "Japanese Corpus".

Oltramari, Alessandro, Piek Vossen, Lu Qin & Hovy, Eduard. (2013). *New trends of research in ontologies and lexical resources: Ideas, projects, systems*. Springer.

Prévot, Laurent, Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci & Alessandro Oltramari. (2010). Ontology and the lexicon: a multidisciplinary perspective. In Chu-Ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, Alessandro Oltramari & Laurent Prévot (Eds.), *Ontology and the lexicon: a natural language processing perspective* (Studies in Natural Language Processing) (pp. 3–24). Cambridge: Cambridge University Press.

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

Spohr, Dennis. (2012). *Towards a multifunctional lexical resource: Design and implementation of a graph-based lexicon model* (Lexicographica, series major). Berlin/Boston: Walter De Gruyter.

Yamada, Tadao, Takeshi Shibata, Kenji Sakai, Yasuo Kuramochi, Akio Yamada, Zendō Ueno, Masahiro Ijima & Hiroyuki Sasahara. (2011). *Shinmeikai Kokugo Jiten* [Shinmeikai Japanese-Japanese dictionary] (7th edition). Tokyo: Sanseido.

Authors' addresses

Terry Joyce
School of Global Studies
Tama University
802 Engyo, Fujisawa-shi, Kanagawa-ken
252-0805
Japan

terry@tama.ac.jp

Bor Hodošček
Department of Language and Culture,
Graduate School of Language and Culture
Osaka University
Machikaneyama-cho 1–8, Toyonaka, Osaka
560-0043
Japan

bor@lang.osaka-u.ac.jp

Hisashi Masuda
Hiroshima Shudo University
1-1-1 Ozuka-Higashi
Asaminami-ku, Hiroshima, 731-3195
Japan

hmasuda@shudo-u.ac.jp