Constructing a Database of Japanese Lexical Properties: Outlining its Basic Framework and Initial Components[1]

日本語語彙特性のデータベースの構築
―その基礎枠組み及び主要中核要素の概観―

Terry Joyce, Hisashi Masuda, Bor Hodošček[2]
ジョイス・テリー、 増田尚史、 ホドシチェク・ボル

Abstract. In order to be able to conduct meaningful research into all aspects of language, it is essential for language science and cognitive science researchers to have practical access to an increasingly wider range of detailed and contemporary information about their target languages. Against that background, this paper presents a short overview summary of an ongoing project to construct a large-scale database of Japanese lexical properties (JLP). More specifically, after outlining the concurrent construction of the ontology of Japanese lexical properties (JLP-O; Joyce & Hodošček, 2014), which provides the basic guiding framework for the JLP database construction project, the paper also outlines the initial core components of the JLP database, with particular emphasis on two of those components; namely, a database of semantic transparency (ST) ratings for approximately 10,000 two-kanji compound words and some initial results for the extraction and automatic analyses of the word structures of both three- and four-kanji compound words.

Keywords: Japanese lexical properties (JLP), database construction, ontology, semantic transparency, automatic analyses of compound word structure

要旨： 言語科学者や認知科学者にとって，言語のあらゆる側面について有意義な研究を企図するためには，目的とする言語に関する詳細かつ現代的な幅広い情報に実用可能なレベルでアクセスできることが必要不可欠である。このことを背景として，本稿では，日本語の語彙特性に関する大規模データベースの構築を目指して現在進行中のプロジェクトについての概要を説明する。具体的には，この日本語語彙特性データベース構築プロジェクトに対して基本的な枠組みを提供する，日本語語彙特性に関するオントロジー（Joyce & Hodošček, 2014）の構築について概観したのちに，日本語語彙特性データベースの主要中核要素について略述する。特に，約 10,000 の漢字二字熟語に対する意味的透明性の評定データベースと，漢字三字および四字の熟語の抽出とその語構造に対する自動分析に関する主要な結果という 2 種類の中核要素を取り上げて論じる。

キーワード： 日本語語彙特性, データベース構築, オントロジー, 意味的透明性, 熟語構成の自動的分析

1. Introduction

Language science and cognitive science researchers seek to investigate the nature of language as both a complex phenomenon in its own right and as an integral component

of human cognition. While it is fair to say that a great deal has already been discovered, it is also equally true that still much about how language functions and how it contributes to cognition remains unknown and subject to academic speculation and debate. For instance, although the closely-related areas of visual word recognition and reading benefit greatly from the impressive breadth of expertise and knowledge amassed within Adelman's (2012b) two-volume edited collection on visual word recognition and, more recently, within Pollatsek and Treiman's (2015) edited handbook of reading, arguably, naïve assumptions and enduring misconceptions surrounding the typology of writing systems (Joyce, in press), coupled with the dominating influence of models conceived of primarily to account for the idiosyncratic nature of English orthography (Share, 2008), often only serve to seriously undermine the value of some studies and cross-linguistic comparisons in terms of advancing our understandings of language and cognition. Notwithstanding such theoretical considerations, however, in order to be able to conduct meaningful research into all aspects of language—from cognitive issues, such as tracing the time courses of orthographic, phonological and semantic activation within visual word recognition, to applied issues, such as creating effective instruction drills—it is absolutely fundamental for researchers to have practical access to an increasingly wide range of detailed and contemporary information about the target languages.

Traditionally, the sources of such information have been mainly limited to various kinds of dictionaries. For the Japanese language, for instance, authoritative language dictionaries include Shinmura's (2008) 広辞苑 /Kōjien/ and Kindaichi, Yamada, Shibata, Sakai, Kuramochi, and Yamada's (2011) 新明解国語辞典 /Shinmeikai Kokugojiten/, as well as kanji character dictionaries, such as Morohashi's (2000) 大漢和辞典 /Daikanwajiten/. However, as dictionaries rarely provide summaries of the lexical information that they contain, typically, researchers have to either independently undertake the tremendously time-consuming and often highly complicated tasks of extracting and summarizing target information or to simply suffice with whatever incomplete samplings and partial estimates that may already exist.[3] Moreover, the viabilities of pursuing either option are increasingly becoming severely challenged as the range of interrelated lexical properties that researchers require data about continues to expand rapidly in terms of both the scope and depth of analyses. For instance, it is possible to discern some sense of the extensive scope dimension from Nation's (2013) classification of lexical knowledge—particularly influential in the areas of second language acquisition and vocabulary instruction—which consists of nine broad kinds of

knowledge about words organized under three groupings of form (spoken, written and word parts), meaning (form and meaning, concept and referent, and associations) and use (grammatical functions, collocation, and constraints on use). Similarly, one can gain some sense of the depth of analysis dimension from momentarily reflecting on the single domain of word recognition research, where, for example, Balota, Yap, Hutchinson and Cortese (2012) acknowledge 15 influencing variables, including word frequency, familiarity, imageability, number of meanings, letter length, phoneme length, syllable length, number of morphemes, and various forms of neighborhoods.[4]

This paper reports on an ongoing research project that is seeking to address this crucial research problem by working towards the construction of a large-scale database of Japanese lexical properties (JLP). Our ultimate goal is to create a comprehensive JLP database that can serve as both a versatile research tool and powerful model for all areas of linguistic and psycholinguistic research on the Japanese lexicon and lexical knowledge. However, given that some aspects of the project have already been described in varying degrees of detail elsewhere (Joyce, 2014; Joyce & Hodošček, 2014; Joyce, Hodošček, & Masuda, 2014a, 2014b, under review; Joyce, Hodošček, & Nishina, 2010, 2012; Joyce, Masuda, & Ogawa, 2012, 2014; Masuda, 2014a, 2014b; Masuda, Fujita, Ogawa, Joyce, & Kawakami, 2013; Masuda, Joyce, Ogawa, Fujita, & Kawakami, 2012; Masuda, Joyce, Ogawa, Kawakami, & Fujita, 2014), the primary objective of this paper is to offer more of a coherent, albeit, of necessity, still relatively brief, overview summary of the project to construct the JLP database in terms of its basic framework, within Section 2, and some of its initial core components, within Section 3. In particular, this paper briefly outlines two database components that have not been described in any detail within our English-language papers to date; namely, a database of semantic transparency (ST) ratings for approximately 10,000 two-kanji compound words and initial automatic extraction and analyses of the word structures of three- and four-kanji compound words.

2. Basic Framework

While the initiative to construct a large-scale JLP database essentially emerged quite naturally and spontaneously out of a convergence of related studies on various orthographic aspects of the Japanese writing system (Joyce, Hodošček, & Nishina, 2010, 2012; Joyce, Masuda, & Ogawa, 2012, 2014; Masuda, Joyce, Ogawa, Fujita, & Kawakami, 2012), Joyce and Hodošček's (2014) proposal to concurrently construct an

ontology of Japanese lexical properties (JLP-O) marked a particularly pivotal development. As the JLP-O is now very much at the heart of the project, effectively providing a basic guiding conceptual framework for the construction work, it is particularly germane to continue this overview summary by briefly describing its initial specifications (Joyce & Hodošček, 2014), in Section 2.1, and some subsequent expansions made to satisfactorily handle the complexity of the Japanese writing system (Joyce, Hodošček, & Masuda, 2014a, under review), in Section 2.2.

2.1 Ontology of Japanese lexical properties (JLP-O)

Within the natural language processing (NLP) and knowledge engineering communities, there has been a trend recently towards the merging of lexical resources with ontologies (Huang, Calzolari, Gangemi, Lenci, Oltramari, & Prévot, 2010; Oltramari, Vossen, Qin, & Hovy, 2013), where an ontology is commonly defined as a formal specification of a shared conceptualization (Guarino, Oberle, & Staab, 2009; Prévot, Huang, Calzolari, Gangemi, Lenci, & Oltramari, 2010), with formal specification indicating a basic commitment to represent an ontology in a machine-readable format. Drawing inspiration from this trend, Joyce and Hodošček's (2014) proposition to simultaneously construct the JLP-O undoubtedly yields three extremely tangible benefits.

The first key advantage is that in representing a basic guiding framework for the project, the JLP-O can greatly facilitate the actual construction work by making it possible to utilize NLP techniques to integrate existing lexical resources. A second important benefit lies in the fact that, by their very nature, ontologies are particularly valuable tools for reflecting on the component entities of a domain and their interconnectivity. Thus, in simultaneously constructing the JLP-O, the construction project is also building a powerful tool for constantly evaluating both the theoretical and psychological validities of various candidate lexical properties inherent within existing lexical resources.[5] The third major merit is that the high degree of formal specification that an ontology entails is also a prerequisite for subsequently realizing powerful search and query capabilities for the JLP database. To be useful, modern databases must possess increasingly higher degrees of multifunctionality in order to be able to realize the easy extraction of relevant information about target lexical properties.

In setting out the initial JLP-O specifications, Joyce and Hodošček (2014) addressed two fundamental concerns; namely, to specify its properties modules, as a core structural issue, and to specify an appropriate range of lexical entry (LE) classes, as a

particularly thorny issue for lexical resources. Naturally, the two concerns are closely interrelated as the database's ability to accurately map out the rich patterns of interconnectivity both between various lexical properties and between the LEs that possess such properties crucially depends on having an effective synergism between the JLP-O's property modules and its LE classes. Thus, as an appealing approach towards the structuring of Japanese lexical properties and realizing their complex mappings across LEs within the JLP database,[6] the JLP-O is utilizing a notion of modules—referring to groupings of related lexical properties and various forms of related metadata—that is similar in spirit to that employed within lemon; lexical model for ontologies (http://lemon-model.net/). The initial JLP-O specification includes six main modules of character, orthographic, phonological, morphological, semantic, and use.

The second basic specification issue of selecting an appropriate range of LE classes also has far-reaching implications for constructing the JLP database. As Joyce and Hodošček (2014) discuss in some detail, the challenge was to find a reasonable compromise between a set of conflicting constraints, which include the nature of the Japanese language itself—where for a highly agglutinative language with ambiguous word boundaries, the strongest desideratum would seem to be for the smallest components of morphemes—the needs of diverse users of the eventual JLP database—where a wider range of LE classes would be preferable to developing powerful search capabilities (Spohr, 2012)—and issues of representation—where the formal specification employed must be adequate to capture the complex relationships between different classes of LEs and the property modules. Moreover, although UniDic, which is the electronic morphological dictionary developed as part of the Balanced Corpus of Contemporary Written Japanese project (BCCWJ; Maekawa, Yamazaki, Ogiso, Maruyama, Ogura, Kashino, Koiso, Yamaguchi, & Den, 2013), takes the so-called short-unit word (SUW), which roughly equates to morphemes, as its primary entity, as Joyce, Hodošček, and Nishina (2012) also discuss in some detail, one considerable downside of that decision is that users, both machines and humans, require a great deal of supplementary information about the permissible concatenations of Japanese SUWs, which the BCCWJ provides in terms of its annotations about so-called long-unit words (LUW). Given the drawbacks associated with defining just two LE classes (i.e., SUWs and LUWs), as a more realistic solution to the inherent constraints, Joyce and Hodošček elected to adopt five LE classes; a range that is more consistent in nature to the upper-levels of Spohr's (2012) typology of lexeme subclasses. Thus, the JLP-O's five LE classes are `Character`, `BoundUnit`, `SimpleWord`, `ComplexWord` and

`MultiWordExpression`. Having resolved these two crucial specification issues, Joyce and Hodošček were able to generate the first Resource Description Framework (RDF) representation of the JLP-O (using the Turtle format), as the small section of the `SimpleWord` LE for 読む /yomu/ 'to read' in Figure 1 illustrates.

```
jlpo:読む_動詞-一般
  a jlpo:SimpleWord ;
  lemon:canonicalForm [
    lemon:writtenRep "読む"@ja ;
    jlpo:orthographicDecomposition (
      [ 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:orthographicDecomposition (
      [ 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 1. Section of RDF representation for `SimpleWord` LE of 読む /yomu/ 'to read' using the Turtle format (Joyce & Hodošček, 2014; 181).

2.2 Handling the complexity of the Japanese writing system

With the core JLP-O specifications established in terms of a suitable range of property modules and an appropriate range of LE classes, Joyce, Hodošček, & Masuda (2014a; under review) next turned to tackle another aspect of fundamental importance for the construction project; namely, how to adequately handle the special issues that arise out of the sheer complexity of the Japanese writing system. While Joyce (2013) and Joyce, Hodošček, and Nishina (2012) describe and illustrate the complexity of the Japanese writing system in some detail, in a nutshell, it is a natural consequence of the fact that the relationships between semantic, orthographic and phonological information are overwhelmingly many-to-many. That is, in addition to the very high incidences of homophone relations between words, the vast majority of Japanese lemmas have either multiple word forms or multiple orthographic forms, or both multiple word and orthographic forms. For example, the adverb lemma of 矢張り, covering meanings of 'also; as I thought; still, in spite of; absolutely; of course", has, in descending order of discourse formality, the three word forms of /yahari/, /yappari/ and /yappa/, and multiple orthographic forms with 矢張り, やはり, ヤハリ, やっぱり, ヤッパリ, やっぱ and ヤッパ.[7] Given that the complexity of the Japanese writing system is a multifaceted issue in itself emerging from an amalgamation of a number of different factors, Joyce, Hodošček, and Masuda have adopted a combination of strategies to ensure that all aspects are accurately captured within the JLP-O.

Actually, the basic mechanisms for the first strategy were already incorporated within the initial JLP-O specifications with the intentional provision, for just this purpose, of both a character module and a `Character` LE class,[8] but Joyce, Hodošček, and Masuda (2014a; under review) have greatly extended the initial implementation. More specifically, they combined the initial set of 6,781 `Character` LEs extracted from a BCCWJ-based corpus lexicon (described further in Section 3.1 below) with the 11,272 characters of the Japanese Industrial Standard (JIS) character set (JIS X 0214 2004), which is the official character set standard for electronically-mediated communication, and the 12,847 kanji of the KANJIDIC2 project, which is a well-known consolidated XML-format kanji database, to yield the current JLP-O's 13,888 `Character` LEs. In parallel with the work of approximately doubling the number of `Character` LEs, Joyce, Hodošček, and Masuda also greatly expanded the range of lexical properties and metadata covered by the character module. Basic lexical properties associated with every `Character` LE include a basic type specification (i.e.,

C for Chinese characters, H for hiragana, K for katakana, R for rōmaji and S for symbol), JIS specifications (both reference number and JIS level category), and stroke counts (for both kana and kanji). Moreover, `Character` LEs for kanji also have a wide range of other lexical properties, including status (i.e., whether jōyō kanji, and, if so, instruction grade), various information relating to internal structure and components (from traditional radical classification systems to numerous alternative proposals), and various cross-references (such as Unicode and major dictionaries).

The second vital strategy for handling the complexity of the Japanese writing system is directly related to its inherent potential for orthographic variation, which, as the quantitative analyses of the BCCWJ-based corpus word lists conducted by Joyce, Hodošček, and Nishina (2012) vividly demonstrate, is a highly prevalent characteristic of written Japanese.[9] For instance, five orthographic variations are attested within the BCCWJ corpus for the `SimpleWord` LE of 玉葱 /tamanegi/ 'onions', where 49% of the total instances are represented as 玉ねぎ, 21% as タマネギ, 17% as たまねぎ, 8% as 玉葱 and 5% as 玉ネギ. The second strategy is really quite straightforward in nature, and basically hinges on maintaining a systematic distinction between the lemma of each LE and its orthographic variants.[10] It is also relatively simple to achieve with the JLP-O by defining a `canonicalForm` sub-property for the standard orthographic representation of the lemma and a second `orthographicForm` sub-property to record all the orthographic variants of a given LE.

The third complementary way of dealing with the complexity of the Japanese writing system is actually just one appealing application of a more general approach being employed in further developing the JLP-O beyond its initial specifications; namely, of exploiting the decomposition method as fully as possible. There are three kinds of decomposition implemented within the current JLP-O and, while these applications were conceived of primarily in terms of capturing the many-to-many relationships underlying the complexity of the Japanese writing system, unquestionably, they and other similar applications will also be immensely valuable in subsequently realizing powerful search capabilities for the JLP database. Of most immediate relevance to the complexity of the Japanese writing system, the first kind of decomposition is orthographic decomposition. Given that the total set of JLP-O's `Character` LEs together represent a de facto master list of all the orthographic elements used within the Japanese writing system, all the `orthographicForm`s included within every LE (apart from the `Character` LEs themselves, naturally) are decomposed into their component characters in terms of reference links to the

corresponding `Character` LEs. The second kind of decomposition is phonological decomposition, where, closely paralleling the basic implementation for orthographic decomposition, all the `orthographicForm`s included within every LE are also decomposed into their phonological components. In order to realize this phonological decomposition strategy, however, Joyce, Hodošček, and Masuda (2014a; under review) had to first create a new sub-module within JLP-O's phonological module for the basic units of Japanese phonology, which are equal-length syllable units known as mora. In addition to constituting a master list of all Japanese mora, including both native mora and extensions for foreign sounds, the mora module also contains various kinds of metadata, including unique identification codes, classification codes (i.e., whether basic, voiced, combination or extended), structure codes (i.e., C for consonants, V for vowels, CV for consonant + vowel combinations, etc), as well as correspondences between different transcription systems. Although the third kind of decomposition under development for the JLP-O is morphological decomposition, given both the agglutinative nature of the Japanese language and that compounding is a highly productive process of word formation, further work will be required to completely implement this information. However, as a substantial first step in that direction, all `ComplexWord` LEs now have information about their component `BoundUnit` and `SimpleWord` LEs, based on BCCWJ's LUW annotations about their SUW elements. Even for this initial partial implementation, it has also been necessary to develop a new `conjugationParadigm` module, as a sub-module of the morphology module, in order to be able to refer to the appropriate conjugation of the first verb in verb1+verb2 compound verbs, such as 読み始める /yomihajimeru/ 'to begin to read',[11] but the sub-module will undoubtedly also prove to be extremely useful for subsequently integrating other aspects of verb morphology.

3. Initial JLP Database Components

Following on from Section 2's brief description of the JLP-O, as the basic guiding framework for constructing the large-scale JLP database, this section turns to briefly outline some of the database components that are either already integrated within the JLP database or will be incorporated in the near future. As mentioned earlier, the impetus to construct the JLP database materialized from a number of studies on various aspects of Japanese orthography, which yielded a number of database components, albeit, in some cases, in more embryonic forms.

3.1 BCCWJ-based corpus lexicon

Although Joyce (2005) and Masuda and Joyce (2005) have both previously created smaller-scale lexical databases, the current project to construct the large-scale JLP database partially originates out of Joyce, Hodošček, and Nishina (2010, 2012). In addition to discussing some basic concerns with the BCCWJ—such as its treatment of lemmas (sense discriminations) and the distinction between SUWs and LUWs—and to conducting quantitative analyses of orthographic variation, Joyce, Hodošček, and Nishina also compiled a number of word lists extracted from the BCCWJ.

More specifically, Joyce, Hodošček, and Nishina compiled 14 word lists for both SUWs and LUWs, based on UniDic's parts-of-speech (POS) tags, and which also included a variety of extracted and computed lexical properties. However, after Joyce and Hodošček (2014) established the JLP-O's range of LE classes, they effectively superseded those corpus word lists by newly extracting the corpus lexicon from the BCCWJ to serve as the key component relating to corpus usage within the JLP database.[12] That was achieved by executing a program that used BCCWJ's annotations of SUW and LUW to simultaneously extract all the word types of the BCCWJ and assign them to the appropriate LE subclasses represented in the RDF format. The corpus lexicon consists of approximately 2.7 million LEs across the main LE classes.[13] As already noted, the `Character` LEs of the corpus lexicon have been subsequently supplemented by both the full JIS listings and the KANJIDIC2 database (Joyce, Hodošček, & Masuda, 2014a; under review). Consistent with their closed-class nature, there are relatively few `boundUnit` LEs (such as particles and some affixes), and, obviously, the vast majority of the corpus lexicon consists of `SimpleWord` and `ComplexWord` LEs. Moreover, the substantial difference in the number of `SimpleWord` LEs compared to the more numerous `ComplexWord` LEs is naturally a reflection of both Japanese's rich verb and adjective conjugations and its productive compounding.

3.2 Database components relating to jōyō kanji, radicals, and orthographic codes

Our aspirations towards constructing the large-scale JLP database crystallized substantially with Joyce, Masuda, and Ogawa's (2012; 2014) work in developing three key components of the JLP database; namely, a database of the 2010 revised jōyō kanji list; a radical database based on reanalysis of the internal structures of jōyō and JIS level 1 kanji; and the generation and analysis of orthographic codes for two lexicons.

The first component focused on the 2,136 kanji of the revised jōyō kanji list as they are the core building blocks in the orthographic representation of contemporary written Japanese.[14] A particularly significant aspect of that database component was its organization of the various kinds of data under five broad groupings of metadata, orthographic properties, phonological properties, semantic properties and morphological properties, in foreshadowing Joyce and Hodošček's (2014) specification of the JLP-O's six property modules. The second component database was the product of a new analysis of the internal structures of the jōyō and the JIS level 1 kanji (2,965 kanji) in terms of three basic configurations; namely, left-right, top-bottom, and enclosure-enclosed (as well as a non-divisible category to capture the remainder). In addition to discovering that 91.3% (1,951) of jōyō and 92.6% (2,747) of JIS1 kanji possess these basic configurations, the analysis also identified the 1,072 and 1,290 component elements of jōyō and JIS1 kanji, respectively. This information about the radical elements is particularly valuable both for capturing the orthographic and phonological relationships between kanji that share radical elements and for developing more powerful search capabilities that could utilize information at that level of structure. The third component database of orthographic codes is the product of coding and analyzing both all the headwords of Kōjien (Shinmura, 2008) and all the words of the corpus word lists (Joyce, Hodošček, & Nishina, 2012) in terms of their orthographic representations. The orthographic codes consist of a script code (i.e., C for kanji, H for hiragana, and K for katakana, etc.) for each character of an `orthographicForm`, such that 日本語 /nihongo/ 'Japanese language', for example, has an orthographic code of CCC, while the corresponding hiragana representation of にほんご is assigned an orthographic code of HHHH. Highlighting a more systemic dimension of orthographic variation that greatly contributes to the complexity of the Japanese writing system, one particularly noteworthy finding from the analyses of the orthographic codes was their sheer variety, with at least 9,358 distinct codes identified.[15] Although rather more predictable in nature, the results of analyzing the distributions of the orthographic codes across both SUW and LUW types and tokens also provide interesting insights into the nature of the Japanese writing system. For instance, for SUW tokens, the two most common orthographic codes are H (34.6%)—as many grammatical SUWs, such as particles, are single hiragana-orthography words—and CC (18.7%), while for SUW types, the two most common codes are CC (35.4%) and KKKK (15.2%). Moreover, for LUW tokens, the two most common codes are H (36.4%) and CC (12.0%), while for LUW types, the two most common codes are CCCC (15.4%) and CCC (9.3%). Given

that many three- and four-kanji compound words are complex words consisting of affixes appended to two-kanji words and combinations of two-kanji compound words, respectively, these results also underscore both the importance of two-kanji compound words as the most common orthographic representation of content `SimpleWord` LEs and the significance of kanji-orthography words for Japanese written language, with two-, three- and four-kanji compound words being three of the four most frequent orthographic codes for both `SimpleWord` and `ComplexWord` LEs by BCCWJ corpus type counts.

One of the construction project's most important tasks in the short-term future will be to fully integrate these components within the larger JLP database to ensure that all relevant connections between lexical properties and between LEs are consistently implemented, which will, in turn, contribute to the subsequent realization of search capabilities utilizing these lexical properties. In the case of the jōyō kanji database, the integration work will be greatly facilitated by the recent developments of both the character module and `Character` LEs (Joyce, Hodošček, & Masuda, 2014a; under review). As stressed by Joyce and Hodošček (2014) in outlining their rather nuanced approach to developing the JLP-O as a guiding framework, a fundamental aspect of the work of integrating existing lexical resources lies in the valuable opportunities that such work affords to constantly examine the consistency of the JLP database. The integration of the radical database will also benefit from the recent developments of the character module and `Character` LEs, particularly the incorporation of the KANJIDIC2 database for contrasting Joyce, Masuda and Ogawa's (2012; 2014) analyses of kanji in terms of three basic configurations with other radical classification schemes. Similarly, the recent deployment of decomposition strategies (Joyce, Hodošček, & Masuda, 2014a; under review) and, in particular, the orthographic decomposition of all `orthographicForm`s for all relevant LE classes makes it extremely straightforward to fully realize orthographic codes within the JLP database.

3.3 Database of semantic transparency (ST) ratings for two-kanji compound words

This sub-section provides a concise summary of a series of recent studies that have been constructing a database of ST ratings for two-kanji compound words (Masuda, 2014a, 2014b; Masuda, Fujita, Ogawa, Joyce, & Kawakami, 2013; Masuda, Joyce, Ogawa, Fujita, & Kawakami, 2012; Masuda, Joyce, Ogawa, Kawakami, & Fujita, 2014), which will also be integrated within the JLP database soon. This component database neatly

illustrates the pressing needs of researchers for data about an ever-increasing range of lexical properties, as the impetus to construct this database component stems directly from research evidence (i.e., Libben, 2006) that the semantic transparency of compound words is an important factor that influences visual word recognition and intentions to further investigate this lexical property for Japanese within the constituent-morpheme priming paradigm (i.e., Joyce, 2002; Joyce & Masuda, 2013).

As already noted, two-kanji compound words represent 37.5% of the word types within the corpus word lists (Joyce, Masuda, & Ogawa, 2012, 2014; Joyce, Hodošček, & Nishina, 2012), and, as Joyce (2013) has argued, the principal word formation processes underlying Japanese kanji-orthography words are morphologically-motivated in nature. For example, the word 漢字 /kanji/ 'Chinese characters' itself is a combination of 漢 'Han dynasty' and 字 'character, letter' in a relationship of modifier and modified, respectively. However, due to historical shifts in semantics and varying degrees of polysemy, there are considerable numbers of opaque Japanese compound words, where the overall meaning of the compound word is not readily interpretable in terms of the component meanings, such as 泥棒 /dorobō/ 'thief' which is a combination of 泥 'mud' and 棒 'stick'.

Accordingly, Masuda and colleagues have conducted a large-scale survey to gather ST ratings for a selection of two-kanji compound words. As described in Masuda, Fujita, et al (2013), a list of the most frequent 10,015 two-kanji compound words (excluding obvious proper nouns) was created from the corpus word lists (Joyce, Hodošček, & Nishina, 2012).[16] In total, 1,710 undergraduate students responded to sets of survey questionnaires (mode number of words rated was 1,000; min, 400; max, 10,015). While only requested when a respondent answered 'know' for both a compound word's meaning and pronunciation, the survey questionnaires also included separate ST ratings for the degree of semantic similarity between the meanings of the constituent kanji and the meaning of the whole compound word on 6-point scales (where 5 indicates a high degree of similarity). Accordingly, 151,237 ST ratings (88.6% of the 170,750 survey presentations) were obtained, with the mean number of ratings for each word being 15.1 (min, 13).

Table 1 presents the distribution of the two-kanji compound words across a matrix formed from the ST ratings for the left and right kanji components (Masuda, 2014a; Masuda, Joyce et al 2014). Clearly, the vast majority (94.4%) of the surveyed two-kanji compound words have high ST ratings (3-5 range) for both constituents. Of the remainder, 4.9% have high ratings for only one constituent, and only 0.7% of the two-

kanji compound words have low ratings (0-3 range) for both constituents.

Table 1. Distribution of two-kanji compound words as a function of the ST ratings for left and right components

	Right kanji				
Left kanji	5~4	4~3	3~2	2~1	1~0
5~4	6,213	1,583	179	2	0
4~3	1,082	577	102	6	0
3~2	107	88	49	5	0
2~1	2	3	5	9	0
1~0	0	0	0	0	1

As presented in Table 2, an additional analysis of the mean ST ratings (and standard deviations) was conducted as a function of the constituent's pronunciation; either 音読み /onyomi/ 'Sino-Japanese pronunciation' or 訓読み /kunyomi/ 'native-Japanese pronunciation. The results indicate that two-kanji compound words with mixed pronunciations (i.e., either On+Kun or Kun+On) received slightly lower mean ratings. For instance, for the mixed On+Kun compound word of 役目 /yakume/ 'duty, role', the ST ratings were 4.9 and 2.9 for the left and right constituents, respectively, while for the mixed Kun+On compound word of 弱味 /yowami/ 'weakness', the ST ratings were 4.8 and 2.4 for the left and right constituents, respectively.[17]

Table 2. Mean (and standard deviations) for ST ratings as a function of pronunciation type

		ST ratings	
Pronunciation type		Left kanji	Right kanji
On+On	(N = 8,690)	4.3 (0.50)	4.2 (0.53)
Kun+Kun	(N = 930)	4.3 (0.56)	4.2 (0.60)
Kun+On	(N = 157)	4.3 (0.53)	4.0 (0.64)
On+Kun	(N = 130)	4.1 (0.65)	4.0 (0.74)

Note: On stands for onyomi and Kun stands for kunyomi

Another finding of particular interest for research employing the constituent-morpheme priming paradigm (i.e., Joyce 2002; Joyce & Masuda, 2013), which contrasts

stimuli according to word-formation principle, comes from a further analysis of mean ST ratings (and standard deviations) as a function of the four main word-formation principles, based on word-formation classification data from Masuda and Joyce (2005). As shown in Table 3, the highest ST ratings for both left and right constituents are for synonymous-pair compound words, where both compounds are semantically related to the meaning of the whole compound word, such as 採取 /saishu/ 'pick, collect', where both constituents mean 'take', which has ST ratings of 5.0 and 4.9 for the left and right constituents, respectively. However, ST ratings appear to be independent of what might be referred to as the head element of other word-formation principles, such as the right position noun of modifier+modified compounds or the verbal constituents in either verb+complement or complement+verb compound words. For instance, for the modifier+modified compound word of 山腹 /sanpuku/ 'mountainside', the ST ratings are 5.0 and 3.4 for the left and right constituents, respectively. Similarly, for the verb+complement compound word of 投薬 /tōyaku/ 'give medicine', the ST ratings are 3.1 and 4.9 for the left and right constituents, while for the reversed ordering of the complement+verb compound word of 速達 /sokutatsu/ 'express delivery', the ST ratings are 4.8 and 3.5 for the left and right constituents, respectively.

Table 3. Mean (and standard deviations) for ST ratings as a function of word-formation principle (Masuda & Joyce, 2005)

		ST rating	
Word-formation		Left kanji	Right kanji
Modifier + modified	(*N* = 3,548)	4.4 (0.47)	4.3 (0.49)
Complement + verb	(*N* = 686)	4.3 (0.45)	4.3 (0.43)
Verb + complement	(*N* = 601)	4.4 (0.39)	4.3 (0.51)
Synonymous pairs	(*N* = 83)	4.7 (0.24)	4.6 (0.31)

As already noted, against a background of previous psycholinguistic research that suggests that semantic transparency is an important factor within the lexical processing of compound words (i.e., Libben 2006), Masuda and colleagues have constructed the database of ST ratings for two-kanji compound words to be a potentially valuable resource for various researchers investigating the involvement of morphological information within the lexical processing of Japanese two-kanji words. Moreover, although Masuda and colleagues have analyzed the ST ratings from the perspectives of

other important lexical properties, such as pronunciation types and word-formation principles, such work would have been greatly facilitated if the database of ST ratings was fully integrated within the large-scale JLP database. Accordingly, one of the next tasks for the construction project will be to incorporate the ST ratings as another valuable component.

3.4 Quantitative study of three- and four-kanji Japanese compound words

The most recent aspect of the JLP database under development, as initially introduced in Joyce, Hodošček, and Masuda (2014b), has been to augment all the LEs where the `canonicalForm`s have orthographic codes of either CCC or CCCC—that is, three- and four-kanji compound words—with information about their word structures. This sub-section presents a short outline of that development, moving from a brief framing of the core issue to noting the solution presented in Joyce, Hodošček, and Masuda (2014b).

As alluded to a number of times already, compounding is an extremely productive process of Japanese word formation, but a fundamental problem for both humans and machines in processing long compound words is how to segment them into their appropriate word structures (see, for example, papers in Verhoeven, Daelemans, van Zaanen, & van Huyssteen, 2014). Even with relatively short three-kanji compound words, there are three possible underlying word structures. For instance, 七五三 /shichigosan/ ‘festival (shrine visit) by children aged 7, 5, and 3’ has a 1+1+1 structure consisting of three `SimpleWord`s. In terms of frequency, however, the other two possible word structures of 1+[2] and [2]+1 are far more common, which also involve a variety of permutations in terms of the lexical status of the components, including `BoundUnits`, verbal and adjectival stems and `SimpleWords`.[18] Examples of the 1+[2] word structure include 不自由 /fujiyū/ ‘restricted; impaired’ as a combination of a `BoundUnit` and a `SimpleWord` meaning ‘not’ + ‘free’, respectively, 古美術 /kobijutsu/ ‘antiques’ as a combination of an adjective stem and a `SimpleWord` meaning ‘old’ + ‘art’, respectively, and 腕時計 /udedokei/ ‘wristwatch’ as a combination of two `SimpleWord`s meaning ‘arm’ + ‘watch’, respectively. Similarly, examples of the [2]+1 word structure include 感情的 /kanjōteki/ ‘emotional’ as a combination of a `SimpleWord` and a `BoundUnit` meaning ‘emotion’ + ‘al’, 決定論 /ketteiron/ ‘determinism’ as a combination of a `SimpleWord` and a stem meaning ‘determine’ + ‘theory’, respectively, and 農業者 /nōgyōsha/ ‘agricultural worker’ as a combination of two `SimpleWord`s meaning ‘agriculture’ + ‘person’, respectively.

Generally, as the length of a compound word increases, the number of possible word structures also increases. Accordingly, there are even more possible word structures underlying four-kanji compound words. For instance, an example of a 1+1+1+1 word structure is 関関同立 /kankandōritsu/ referring to 'four famous universities of Kansai (west Japan)', based on the first kanji for the full university names of 関西大学, 関西学院大学, 同志社大学, 立命館大学. An example of a 1+[3] word structure would be 非農業者 /hinōgyōsha/ 'non-agricultural worker' as a combination of a `BoundUnit` and a `ComplexWord` meaning 'non' + 'agricultural worker', while an example of a [3]+1 word structure would be 決定論的 /ketteironteki/ 'deterministic' as a combination of a `ComplexWord` and a `BoundUnit` meaning 'determinism' + 'ic', respectively.[19] Moreover, given that the two-kanji compound word is the most frequent orthographic code for content `SimpleWord`s, unsurprisingly, many four-kanji compound words are combinations of two-kanji `SimpleWord`s, such as 大学入試 /daigakunyūshi/ 'university entrance examination' as a combination of the two `SimpleWord`s of 'university' + 'entrance examination' and 単語認知 /tangoninchi/ 'word recognition' as a combination of the two `SimpleWord`s of 'word' + 'recognition', respectively.

As a first step towards augmenting all `ComplexWord` LEs with information about their word structures, Joyce, Hodošček and Masuda (2014b) have focused only on three- and four-kanji compound words. The first stage was to extract all three- and four kanji compound words from the corpus lexicon within the JLP database; the numbers of which are presented in Table 4.

Table 4. Type counts for both three- and four-kanji `SimpleWord` and `ComplexWord` LEs within the JLP database (Joyce, Hodošček, and Masuda, 2014b)

JLP database LEs	Three-kanji	Four-kanji
`SimpleWords`	6,489	655
`ComplexWords`	220,361	336,615
Totals	226,850	337,270

The second stage was to execute a program to automatically analyze the word structures of these compound words, by effectively referring back to the BCCWJ annotations for LUWs, which, as explained already are treated as `ComplexWord` LEs

within the JLP-O and JLP database. The resultant analyses of word structure have been added to all relevant LEs. While this work of automatic extraction and word-structure analysis has only become feasible because the JLP-O possesses the requisite information about all LE classes and the BCCWJ's analyses of compounds in the form of their LUW annotations, the database construction project also plans to conduct more detailed analyses of these initial word-structure results, such as appropriately coding the word class of stem elements, and to extend the approach to automatically analyzing the word structures of all `ComplexWord` LEs (i.e., compounds of five-kanji and even longer). The word-structure analyses will be invaluable for psycholinguistic research into the visual word recognition processes of longer compounds, which would also conduct supplementary rating surveys to obtain native-speaker evaluations concerning the psychological reality of such word-structure analyses.

4. Conclusion

Even though the ongoing research project to construct a large-scale JLP database is still in its relatively early stages, it has already tackled a number of fundamental specification issues (Joyce, 2014; Joyce & Hodošček, 2014; Joyce, Hodošček, & Masuda, 2014a, under review) and has developed a number of component databases (Joyce, Hodošček, & Masuda, 2014b; Joyce, Hodošček, & Nishina, 2010, 2012; Joyce, Masuda, & Ogawa, 2012, 2014; Masuda, 2014a, 2014b; Masuda, Fujita, Ogawa, Joyce, & Kawakami, 2013; Masuda, Joyce, Ogawa, Fujita, & Kawakami, 2012; Masuda, Joyce, Ogawa, Kawakami, & Fujita, 2014). Given that some of the details have already been provided elsewhere and that a full explication of the JLP-O and the JLP database would obviously be far beyond the scope of this paper, our main intention here has been to tender a summary overview of much of the JLP database construction work to date, which, hopefully, goes some way towards elucidating how the JLP-O is particularly central to the entire endeavor, how the initial core database components complement each other and establish a solid foundation for the JLP database, and how the project will progress in the near future.

To that aim, Section 2 focused on providing an outline of the JLP-O, which Joyce and Hodošček (2014) have justifiably advocated for the considerable advantages that can be leveraged by simultaneously constructing the JLP-O and the large-scale JLP database. Unquestionably, the most important merit is the JLP-O's value in effectively functioning as a guiding conceptual framework that supports the use of NLP techniques

to efficiently integrate existing lexical resources in constructing the JLP database. Closely related, a second major advantage stems directly from the inherent nature of ontology construction itself as an exercise in identifying the entities of a domain and capturing their interrelationships. From the perspective of constructing the JLP-O, that entails the continual assessment of candidate lexical properties in terms of their theoretical and psychological validities. Also intimately related to these first two merits, the third important benefit from concurrently constructing the JLP-O is that it naturally imbues the JLP database with the high degree of formal specification that is indispensable for realizing powerful database search and extraction capabilities.

More specifically, Section 2.1 outlined the initial JLP-O specifications of six main modules, or groupings, of lexical properties and five classes of LEs (Joyce & Hodošček, 2014). In combination, these two core specifications open up extremely promising and flexible approaches to structuring the multiple kinds of Japanese lexical properties and mapping out their complex patterns of interconnectivity both between lexical properties and between various LEs. Section 2.2 then briefly introduced some subsequent developments that have been explicitly implemented to ensure that both the JLP-O and the JLP database can satisfactorily handle the complexity of the Japanese writing system (Joyce, Hodošček, & Masuda, 2014a; under review). The first of these developments was to greatly expand both the number of `Character` LEs and the range of lexical properties encompassed by the character module. The second important development was to establish a distinction within the basic LEs between the standard orthographic representation of the lemma and all the corpus-attested orthographic variants, which was achieved by creating two sub-properties of `canonicalForm` and `orthographicForm` for lemmas and orthographic variants, respectively. The third significant area of development was in three initial deployments of the decomposition strategy, which, given its wide applicability and versatility, is an approach that the construction project will certainly extend to other lexical properties in the future. Thus, the current version of the JLP database has full orthographic decomposition, where all `orthographicForm`s are decomposed into their component `Character` LEs, full phonological decomposition, where all `orthographicForm`s are decomposed into their mora components, and partial morphological decomposition, where all `ComplexWord` LEs are decomposed into their component `BoundUnit` and `SimpleWord` LEs.

In Section 3, the focus shifted to outlining some of the JLP database's initial core components. As introduced in Section 3.1, a central component of the JLP database is

the BCCWJ-based corpus lexicon consisting of approximately 2.7 million LEs (Joyce & Hodošček 2014). It was compiled by extracting word types from the BCCWJ and assigning them to the appropriate JLP-O LE subclasses. As described in Section 3.2, Joyce, Masuda, and Ogawa's (2012; 2014) creation of three database components was especially significant in terms of kindling our ambitions to pursue the project of constructing a large-scale JLP database. The first of their components was a database of information relating to the 2010 revised jōyō kanji list, with its five broad groups of information as early precursors for the JLP-O's property modules. The second of their database components relates to the internal structures of jōyō and JIS1 kanji, based on analysis results in terms of three basic configurations and their elements. Their third component developed the assignment of orthographic codes for the JLP database by applying and summarizing orthographic codes to both all Kōjien headwords and the corpus word lists. Priority future tasks for the JLP database project will be to fully integrate these separate component databases within the larger JLP database with appropriate developments of the JLP-O to faithfully represent the additional information about various lexical properties that these components encompass.

As described in some detail in Section 3.3, another database component developed recently, which will soon be integrated within the JLP database, is a database of ST ratings for approximately 10,000 two-kanji compound words obtained by conducting a large-scale survey. The integration of this database component exemplifies one of the main approaches that the project will continue to employ in constructing the JLP database to be both as comprehensive and as beneficial to researchers as possible. Complementary to the other main approach of integrating important existing lexical resources—which can be characterized as being essentially top-down in nature—the approach to supplementing the database with newly compiled data is driven primarily by researcher needs for an ever-expanding range information about lexical properties—which can be regarded as being more bottom-up in nature. The brief outline of the ST rating data also sought to illustrate how the JLP database can facilitate investigations into how any given lexical property, such as ST ratings, is related to other lexical properties, such as component kanji pronunciations (i.e., onyomi or kunyomi) and word-formation principle. As presented in Section 3.4, another database component under recent development is to augment the LEs for three- and four-kanji compound words with information about their word structures through methods of automatic extraction and analysis. The pilot work with three- and four-kanji compound words will subsequently be extended to generate word-structure information for all

`ComplexWord` LEs,[20] but it already stands as simple testament to the kinds of valuable analytical investigations that really only become feasible by constructing the JLP-O and JLP database simultaneously.

In summary, this paper has presented a summary overview of an ongoing project to create the JLP-O—by outlining its considerable merits as guiding framework, assessment tool for theoretical and psychological validities, and formal specification vital to realize search capabilities—in parallel with constructing the large-scale JLP database itself—by outlining some of the initial core components of the JLP database. Central among the motivations driving the project is the genuine needs for large-scale and contemporary information about a wide range of Japanese lexical properties, and the aspiration that the JLP-O and JLP database can become a truly comprehensive model of the Japanese lexicon that could be used by language science and cognitive science researchers in advancing our understanding of the amazing phenomenon that is language.

Notes

[1] During the 2014 academic year, this research was partially supported by Tama University Funding for Joint Research Projects awarded to the first author and by JSPS Grant-in-Aid for Foreign Postdoctoral Fellows (no. P13303) awarded to the third author.

[2] Terry Joyce is the professor of psychology at Tama University's School of Global Studies; Hisashi Masuda is a dean and professor of psychology at Hiroshima Shudo University; Bor Hodošček is an assistant professor at Osaka University. The authors wish to acknowledge significant academic contributions to the larger research project outlined within this paper from both Dr. Chikako Fujita of Nanzan University and Dr. Taeko Ogawa of Tokai Gakuin University.

[3] Hayashi, Miyajima, Nomura, Egawa, Nakano, Sanada, and Satake's (1982) 図説日本語 /Zūsetsu Nihongo/ is a relatively rare source of partial summary data, organized under a lexical section, with some frequency, word class and formation information, and an orthographic section, with some counts, usage, readings information particularly for kanji, as well as sections on phonology, accents, and grammar. However, in addition to being somewhat fragmented in its coverage, clearly, it is no longer a reliable source of information concerning contemporary usages.

[4] Adelman (2012a) also presents a list of 14 important variables that need to be controlled for when conducting visual word recognition experiments. While Adelman's listing is generally consistent with Balota et al's (2012) variable list, there are a number of variables that are only mentioned on one list and not the other.

[5] At the same time, we remain acutely aware that natural systems, like language, do not necessarily confirm to the standards of ontological completeness and logic. Thus, our approach to constructing the JLP-O can perhaps be characterized as one of skeptical pragmatism in seeking to strike a realistic balance between the practical merits obtainable from fully leveraging the formal specifications of an ontology and constantly assessing the JLP-O's psychological validity.

[6] As construction of the JLP database advances, the range of lexical properties continues to expand steadily, but the JLP-O already covers more than 70 Japanese lexical properties.

[7] Ogura, Ogiso, Koiso, Hara, and Miyauchi (2010) use 矢張り as an effective example within a table that illustrates the three basic levels of UniDic entries; the lemma (語彙素 /goiso/), word forms (語形 /gokei/) and orthographic forms (書字形 /shojikei/), respectively.

[8] Although the inclusion of the `Character` LE class is, arguably, not totally consistent with the ultimate goal of the project to realize a database of the Japanese lexicon and lexical properties, Joyce and Hodošček (2014) anticipated how it would be key to handling the complexities of Japanese orthography.

[9] For instance, taking the 100 most frequent lemmas from the four word classes of nouns, verbs, adverbs and i-adjectives, the average number of orthographic variants and ranges were found to be 8.44 and 1-34 for SUWs and 5.80 and 1-28 for LUWs, respectively.

[10] The demarcation is conceptually similar to the BCCWJ's distinction between the lemma and its orthographic forms (語彙素 and 書字形, respectively) and it is also methodologically similar to lemon's distinction between its `canonicalForm` and `otherForm` sub-properties, even though the motivation there is quite different.

[11] This has been necessary, as the BCCWJ annotations only refer to SUW lemmas which correspond to the JLP-O's `canonicalForm` property.

[12] While acknowledging that there are inherent issues with using corpus data for constructing a comprehensive database, such as treatments of proper nouns and extremely low frequency words, still, as the BCCWJ unquestionably represents the most authoritative sampling of contemporary written Japanese language currently available, it is extremely valuable for the JLP database construction project.

[13] It should be noted that `MultiWordExpression`s LEs are not included in the present version of the JLP database, for, although it would have been feasible to also extract collocational and idiom data when creating the corpus lexicon, Joyce and Hodošček (2014) decided it would be more prudent to create those LEs in the future when integrating other lexical resources with suitable information.

[14] Joyce, Masuda, and Ogawa (2012) examined the coverage rates for jōyō kanji within the corpus word lists (Joyce, Hodošček, & Nishina 2012) and found that, while they only account for 33.03% of all types, they represent the vast majority of kanji tokens at 96.12%. Moreover, while the additional JIS1 and JIS2 kanji (i.e., excluding the jōyō kanji) account for an extra 63.30% of types, they only represent an extra 3.60% of tokens.

[15] Actually, Joyce, Masuda and Ogawa (2012) counted 242 and 42,226 different orthographic codes for lemma types for SUWs and LUWs, respectively, but, given that 37 and 33,073 of those, respectively, were unique orthographic codes associated with just one word within the corpus word lists, it seems more prudent, pending further analyses, to acknowledge the number of orthographic codes shared by at least two words, even though it, admittedly, only provides an extremely conservative estimate of the phenomenon.

[16] However, two of the items (群落, 馬丁) were subsequently excluded from analyses, as all the respondents indicated that they were unknown words.

[17] It should be noted, however, that ST ratings are independent of pronunciation type and constituent position, for although the degree of semantic similarity was rated higher for left kanji of both 役目 and 弱味, the mixed On+Kun compound word of 蛇口 /jyaguchi/ 'tap, faucet' has ST ratings of 2.8 and 4.0 for the left and right constituents, respectively, while the mixed Kun+On compound of 手帳 /techō/ 'notebook' has ST ratings of 3.3 and 4.5 for the left and right constituents, respectively.

[18] Setting aside debates over the appropriateness of extending the notion of derivational morphology to Japanese, a number of the `BoundUnit` LEs are certainly affix-like in behavior.

[19] Admittedly, we are omitting the additional analyses of the `ComplexWord` LEs into their word structures from the presentations of examples within the paper, but these additional analyses are included within the JLP database.

[20] The basic methodology of matching to component LEs within the JLP database will also be extended to realize morphological analysis information for all appropriate LEs.

References

Adelman, J. S. (2012a). Methodological issues with words. In J. Adelman (Ed.), *Visual word recognition Volume 1: Models and methods, orthography and phonology* (pp. 116-138). London: Psychology Press.

Adelman, J. (Ed.). (2012b). *Visual word recognition [Volume 1: Models and methods, orthography and phonology; Volume 2: Meaning and context, individuals and development]*. London: Psychology Press.

Balota, D. A., Yap, M. J., Hutchinson, K. A., & Cortese, M. J. (2012). What do millions (or so) of trials tell us about lexical processing. In J. S. Adelman (Ed.), *Visual word recognition volume 1: Models and methods, orthography and phonology* (pp. 90-115). London: Psychology Press.

Guarino, N., Oberle, D., & Staab, S. (2009). What is an ontology? In S. Staab & R. Studer (Eds.), *Handbook on ontologies* (second edition) (pp. 1−17). Berlin: 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. (Eds). (2010). *Ontology and the lexicon: A natural language processing perspective*. Cambridge: Cambridge University Press.

Joyce, T. (2002). Constituent-morpheme priming: Implications from the morphology of two-kanji compound words. *Japanese Psychological Research*, 44(2), 79-90.

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

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

Joyce, T. (2014). Constructing an ontology-based lexical database of Japanese lexical properties. Invited talk at "Contributions of large-scale lexical databases to psycholinguistic research" Symposum of *15th International Conference on the Processing of East Asian Languages*, 24-26 October. Korean University, Seoul, Korea.

Joyce, T. (in press). Writing systems and scripts. In L. de Saussure & A. Rocci (Eds.), *Verbal communication* (Handbooks of Communication Science 3). Berlin: Mouton de Gruyter.

Joyce, T., & Hodošček, B. (2014). Constructing an ontology of Japanese lexical properties: Specifying its property structures and lexical entries. In M. Zock, R. Rapp, & C.-R. Huang (Eds.), *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex4)* (pp. 174-185). 23 August. Dublin, Ireland.

Joyce, T., Hodošček, B., & Masuda, H. (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, T., Hodošček, B., & Masuda, H. (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. Korean University, Seoul, Korea.

Joyce, T., Hodošček, B., Masuda, H. (under review). Constructing an ontology and database of Japanese lexical properties: Handling the orthographic complexity of the Japanese writing system [Special issue: Orthographic databases and lexicons, edited by L. Cahill & T. Joyce] *Written Language and Literacy*.

Joyce, T. Hodoscek, B., & Nishina, K. (2010). Orthographic representation within the Japanese writing system. *'Units of Language – Units of Writing' 7th International Workshop on Writing Systems*, 30 September – 1 October, Université Paris Descartes – Sorbonne, Paris, France.

Joyce, T., Hodošček, B., & Nishina, K. (2012). Orthographic representation and variation within the Japanese writing system: Some corpus-based observations [Special issue: Units of language − units of writing, edited by T. Joyce & D. Roberts]. *Written Language & Literacy*, 15(2), 254−278. doi:10.1075/wll.15.2.01rob

Joyce, T. & Masuda, H. (2013). Constituent-morpheme priming study of Sino-Japanese and Native-Japanese compound words. *8th International Morphological Processing Conference*, 20-22 June, Cambridge, UK.

Joyce, T., Masuda, H., & Ogawa, T. (2012). Jōyō kanji: Recent revision, characteristics, and role as core component of the Japanese writing system. *'The Architecture of Writing Systems': 8th International Workshop on Writing Systems and Literacy*, 4-5 October. Institut für Germanistik, Oldenburg, Germany.

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

KANJIDIC2: http://www.edrdg.org/kanjidic/kanjd2index.html

Kindaichi, K., Yamada, T., Shibata, T., Sakai, K., Kuramochi, Y., & Yamada, A. (2011). *Shinmeikai Kokugojiten* [Shinmeikai Japanese-Japanese dictionary] (seventh edition). Tokyo: Sanseido.

Libben, G. (2006). Why study compound processing? An overview of the issues. In G. Libben & G. Jarema (Eds.), *The representation and processing of compound words* (pp. 1-22). Oxford: Oxford University Press.

Masuda, H. (2014a). Fukugōgo to shite no kanjinijihyōkigo no imikōzō: Imitekitōmeisei no chōsa to kōseikeitaiso puraimingu jiken o moto ni. *17th Ninchi Shinkeishinrigaku Kenkyūkai*, Okayama Prefectural Library.

Masuda, H. (2014b). Kanjinijihyōkigokan no imitekikankeisei ni kansuru dētabēsu no kōchiku. *Kagaku Kenkyū Hijo Seijigyō Kenkyū Seika Hōkokusho*.

Masuda, H., Fujita, C., Ogawa, T., Joyce, T., & Kawakami, M. (2013). Kanjinijijukugo no imitekitōmeisei ni kansuru chōsa. *32nd Annual Meeting of the Nihon Kiso Shinrigakkai*, Kanezawa Cultural Hall.

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

Masuda, H., Joyce, T., Ogawa, T., Fujita, C., & Kawakami, M. (2012). Mental lexicon (XVI): Kanjinijijukugo no imitekitōmeisei no dētabēsu no kōchiku ni mukete (1). *76th Meeting of the Japanese Psychological Association*, Senshū University.

Masuda, H., Joyce, T., Ogawa, T., Kawakami, M., & Fujita, C. (2014). A database of semantic transparency ratings for two-kanji Japanese compound words. *'Orthographic Databases and Lexicons': 9th International Workshop on Writing Systems and Literacy*, 4-5 September. University of Sussex, Brighton, UK.

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

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

Nation, I. (2013). *Learning vocabulary in another language* (second edition). Cambridge: Cambridge University Press.

Ogura, H., Ogiso, T., Koiso, H., Hara, Y., & Miyauchi, S. (2010). Keitaiso kaiseki jisho UniDic

ni okeru goiso midashi no rikkō hōshin. In *Proceedings of the 2010 Tokuteiryōiki Kenkyū "Nihongo Kōpasu" Public Workshop*. Tokyo: General Headquarters, Priority-Area Research "Japanese Corpus".

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

Pollatsek, Alexander, & Treiman, Rebecca (Eds.). (2015). *The Oxford handbook of reading*. Oxford: Oxford University Press.

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). Cambridge: Cambridge University Press.

Share, D. L. (2008). On the anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin*, 134(4), 584-615.

Shinmura, I. (2008). *Kōjien* [Japanese dictionary] (sixth edition). Tokyo: Iwanami Shoten.

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

Verhoeven, B., Daelemans, W., van Zaanen, M., & van Huyssteen, G. (Eds.). (2014). *Proceedings of the First Workshop on Computational Approaches to Compound Analysis (ComAComA 2014)*. [Workshop held during 25th International Conference on Computational Linguistics (COLING 2014)], 24 August. Dublin, Ireland.

Received on Nov. 30, 2015.