Jōyō kanji as core building blocks of the Japanese writing system

Some observations from database construction

Terry Joyce, Hisashi Masuda & Taeko Ogawa
Tama University / Hiroshima Shudo University / Tokai Gakuin University, Japan

The architecture of writing systems metaphor has special relevance for understanding the structural nature of the Japanese writing system, and, more specifically, for appreciating how the 2,136 kanji of the 常用漢字表 /jō-yō-kan-ji-hyō/* 'List of characters for general use' function as the core building blocks in the orthographic representation of a considerable proportion of the Japanese lexicon. In seeking to illuminate the multiple layers of internal structure within Japanese kanji, the Japanese lexicon, and the Japanese writing system, the paper draws on insights and observations gained from an ongoing project to construct a large-scale Japanese lexical database system. Reflecting structural distinctions within the database, the paper consists of three main sections addressing the different structural levels of kanji components, jōyō kanji, and the lexicon.

Keywords: Japanese writing system; building blocks; jōyō kanji; components; orthographic structure; database

1. Introduction

The Japanese people number among the overwhelming majority of people who did not invent writing for themselves. In a process that has been repeated uncountable times throughout human history, the ancient Japanese people came to know about writing through contact with a neighboring culture – in their case China (via Korea) – and to borrow the Chinese writing system. However, because Chinese and Japanese are typologically very different languages, that initial borrowing – always a quirk of historical fate rather than a matter of option – set in motion a

* When providing phonological glosses, this paper uses hyphens to mark kanji-kanji boundaries and periods to indicate kanji-hiragana boundaries.

Written Language & Literacy 17:2 (2014), **173–194**. DOI 10.1075/wll.17.2.01joy
ISSN 1387–6732 / E-ISSN 1570–6001

long and gradual process of adapting Chinese characters to the Japanese language. Although the process entailed the independent development of two supplementary syllabographic scripts, 平仮名 /hira-ga-na/ and 片仮名 /kata-ka-na/, the creations of the kana scripts did not replace the use of morphographic 漢字 /kan-ji/ (literally, 'Chinese characters') (Joyce 2011), but led to the 漢字かな混じり文 /kan-ji.kana.ma.jiri.bun/ 'mixed kanji and kana writing' that uniquely characterizes the present Japanese writing system.

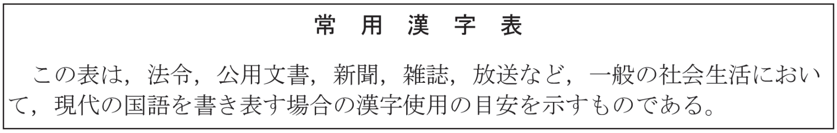
As Joyce (2011) details, the Japanese writing system has gained a distinctive reputation amongst writing system researchers because of its complexity. The general sentiment is succinctly conveyed – and more objectively than many – by Coulmas (1989) when he comments that "under the hands of the Japanese, Chinese characters were transformed to become what is often said to be the most intricate and complicated writing system ever used by a sizeable population" (Coulmas 1989: 122). Setting aside here questions of whether such portrayals are completely justified, still studies of the Japanese writing system are also likely to yield interesting insights into the architectural principles inherent within writing systems. Given the axiom of writing systems research that writing systems are related to language, such principles are also reflections of the combinatorial principles of language. To the extent that architecture is primarily about how various building blocks can be combined, these metaphors have special resonance for understanding the architectural principles of the Japanese writing system at multiple levels. At the syntactic level, the notion of building blocks is highly applicable to the functional demarcation realized by utilizing multiple scripts. As Taylor and Park (1995) note and as touched on briefly in Section 2, content words are generally represented by kanji, or by katakana in the case of foreign loanwords, and functional words are represented in hiragana. Accordingly, from the architectural analogy, content words could be likened to the building blocks of sentences and functional words to the bonding between them.

However, the building blocks metaphor has even deeper significance at the lexical level of content words. As Joyce (2002, 2011) argues and as explained briefly in Section 5, it is most informative to recognize that Japanese kanji are essentially morphographic in nature. And, while there is immerse diversity in the orthographic structures of the Japanese lexicon, as Section 5 also seeks to demonstrate, a substantial proportion of Japanese words are still represented orthographically by combinations of kanji or combinations of kanji and kana. Accordingly, at the lexical level, kanji can be regarded as the core building blocks for the orthographic representation of most Japanese polymorphemic words. Finally, the building blocks metaphor also has relevance for thinking about the internal structure of kanji at the sub-grapheme level. As briefly sketched out in Section 4, although kanji vary considerably in their complex, most complex kanji consist of reoccurring components that are organized according to a few basic configurations. Accordingly, at the sub-grapheme level, these components can also be regarded as the building blocks of kanji.

As potentially the more illustrative of the architectural principles of writing systems, this paper focuses on the lexical and sub-grapheme levels of building blocks. Thus, after a brief outline of the Japanese writing system in Section 2, Section 3 introduces the 2010 revision of the 常用漢字表 /jō-yō-kan-ji-hyō/ 'List of characters for general use', while Sections 4 and 5 address the sub-grapheme and the lexical levels, respectively. Sections 3–5 consist of an initial overview followed by an introduction of the relevant component within the larger database system under ongoing construction. Finally, Section 6 offers a few summary remarks to conclude the paper.

2. Brief outline of the Japanese writing system

This section offers a few basic remarks about the Japanese writing system for the benefit of the more general reader (for fuller outlines, see, for example, Joyce (2011), Joyce, Hodoščček & Nishina (2012), and Smith (1996)). In the interest of brevity, our strategy to briefly comment on a short piece of authentic Japanese text also attempts to partially dual-task in taking, as a natural example, the first sentence from the official jōyō kanji list document (Bunkachō 2010).



常用漢字表

この表は，法令，公用文書，新聞，雑誌，放送など，一般の社会生活において，現代の国語を書き表す場合の漢字使用の目安を示すものである。

Figure 1. Example of Japanese text from jōyō kanji list document [English translation: 'Jōyō Kanji List. This list is the standard for kanji usage when writing modern Japanese in general social activities, such as laws, public documents, newspapers, magazines, and broadcasting.']

As already noted, one of the most basic characteristics of the Japanese writing system is its employment of a mixture of scripts in essentially complementary ways (Joyce, Hodoščček & Nishina 2012). The example text is quite typical in its compositional balance of mainly morphographic kanji, which are generally more complex in form, and syllabographic hiragana, which primarily represent grammatical words, verb and adjective conjugating elements, and some basic words. However, this particular sentence has no katakana orthography, which tends to be used for foreign loanwords or as a form of italicization, nor rōmaji (Roman alphabet) orthography words, which are more common in advertising and glossier magazines. Again, as noted earlier, in the utilization of multi-scripts, Japanese orthographic conventions largely serve to signal a functional differentiation between content and grammatical building blocks.

In principle, Japanese kanji have two kinds of pronunciations; a 訓読み /kun-yo.mi/ 'Native-Japanese pronunciation' and an 音読み /on-yo.mi/ 'Sino-Japanese pronunciation'. For example, 用 – appearing three times within the short extract – has a basic meaning of 'to use' and has a Native-Japanese pronunciation of /mochi.iru/ and a Sino-Japanese pronunciation of /yō/. Sino-Japanese pronunciations tend to be used when a kanji is representing a morpheme within a compound word, which is the case for the three appearances of 用 in 常用 /jō-yō/ 'general use' (usual + use), 公用 /kō-yō/ 'public use' (public + use), and 使用 /shi-yō/ 'use, employ' (use + use).

The most frequent hiragana within the short text is の /no/, appearing six times. Four of the occurrences are in representing a single-mora word that indicates a possessive or modification relationship, while the other two occurrences are in representing the second elements of two bi-mora words, この /kono/ 'this' and もの /mono/ 'thing', respectively.

3. Jōyō kanji

3.1 Overview

The architecture of writing systems metaphor is particularly appealing in thinking about the Japanese writing system, given how naturally and intuitively analogies to building blocks emerge from the morphographic nature of kanji (Joyce 2011), while simultaneously recognizing that, as a naturally-evolved system, kanji also possess internal structure. However, before continuing to further develop the building block metaphor, this section briefly describes the jōyō kanji list and its 2010 revision.

In its listing of 49,963 kanji, Morohashi's (1955–1960) often-cited 大漢和辞典 /dai-kan-wa-ji-ten/ 'Comprehensive Chinese-Japanese dictionary' is strong testimony for the vast number of kanji that have historically been used within the Japanese writing system. To put the figure in clearer perspective, however, it should also be noted that there are separate entries for graphic variants of a kanji and for many obscure kanji only used a few times in some rare texts. Still, Twine (1991) has estimated that the number of kanji commonly used in writing Japanese was more than 10,000 at the end of the Tokugawa period (1603–1868).

Since the mid-20th century, the Japanese Ministry of Education has issued a series of guidelines concerning kanji usage. With the specific intention of greatly reducing the number of kanji in daily use and at simplifying some kanji forms, the first major guideline was the 当用漢字表 /tō-yō-kan-ji-hyō/ list of 1946 which prescribed 1,850 kanji. However, with subsequent guidelines there has been a slight trend towards expansion, where the jōyō kanji list, issued in October 1981, consisted of 1,945 kanji and its revision, issued in November 2010, removed five characters and added 196 new kanji to create a new official list of 2,136 kanji.[1] While some domains of great cultural and heritage significance, such as place, period, and family names, continue to mean that educated Japanese people are expected to know considerably more than 2,000 kanji, still, as a de facto standard for functional literacy within Japan, the jōyō kanji set can be regarded as a core component of the Japanese writing system.

Although it is beyond the scope of this paper to describe the jōyō kanji list in detail, a few cursory comments about the 2010 revision are in order. Rather than representing a fundamental reform in kanji policy, the revision may be regarded more as a periodic fine-tuning of the kanji set. For instance, two of the dropped kanji relate to traditional measurements of less relevance for modern Japanese society; namely, 勺 /shaku/ and /seki/ with a meaning of 'approximately 18 ml' and 匁 /monme/ and /me/ meaning 'approximately 3.75 grams'. Many of the additions remedy some gaps in basic Japanese vocabulary, such as 呂 of お風呂 /o.fu-ro/ 'bath', 鬱 of 鬱病 /utsu-byō/ 'depression', and both kanji for 挨拶 /hai-satsu/ 'greetings'. A number of other additions relate to Japanese proper nouns, such as 茨 of 茨城県 /ibara-ki-ken/ 'Ibaraki Prefecture', 那 of 那覇 /na-ha/ 'Naha' (main city in Okinawa), and 藤 /tō/ and /fuji/ 'wisteria (flower)', which is a component of many common family names, such as 佐藤 /sa-tō/ 'Satō' and 藤原 /fuji-wara/ 'Fujiwara'.

It is important to point out that the jōyō kanji list is only a guideline and, although generally conformed to by newspapers and official documents, it does not represent an absolute upper limit. In the modern age of electronically-mediated communication, a more practical limitation on kanji usage is the character encoding (JIS X-0208-1990) of the 日本工業規格 /ni-hon-kō-gyō-ki-kaku/ 'Japanese Industrial Standards' (JIS), which defines the character set used by Japanese computers and electronic devices such as smartphones. The JIS set consists of 6,355 kanji, further divided into 2,965 level 1 (JIS1) and 3,390 level 2 (JIS2) kanji, with the division largely reflecting usage frequencies. Most jōyō kanji are JIS1 kanji, but, of the recently added 196 kanji, 30 are JIS2 kanji. Table 1 presents kanji coverage data for the corpus word lists created by Joyce, Horoščok and Nishina (2012) from

[1] The list includes a further division between 1,006 教育漢字 /kyō-iku-kan-ji/ 'education kanji' taught during elementary school and the remaining 1,130 kanji taught at high-school.

the *National Institute for Japanese Language and Linguistics*' (NINJAL) (2011) 'Balanced Corpus of Contemporary Written Japanese' (BCCWJ).[2]

Table 1. Kanji coverage for BCCWJ-based corpus word lists (Joyce, Horoščєk & Nishina 2012)

	Counts		**Ratios**	
Kanji set	**Types**	**Tokens**	**Types**	**Tokens**
Jōyō	2,136	74,885,048	33.03	96.12
JIS1 + JIS2	4,093	2,805,816	63.30	3.60
Others	237	214,194	3.67	0.00
Totals	**6,466**	**77,905,058**	**100.00**	**100.00**

Table 1 clearly illustrates the central importance of the 2,136 jōyō kanji within contemporary written Japanese, for although they only represent 33.03% of the types (with 4,093 JIS kanji accounting for 63.30%), they account for the vast majority of tokens at 96.12%. For the 1981 jōyō kanji list, the type and token ratios were quite similar, at 30.08% and 95.47%, respectively. Given that the 2010 revision was made towards the end of the period sampled by the BCCWJ corpus, it is likely that counts and ratios for some of the additional kanji will increase as the revised list gradually exerts its influence over written Japanese, but it is also clear that many of the additional kanji were already in fairly common usage because of their importance for basic Japanese vocabulary, family names, and place names.

3.2 Jōyō kanji database

Following ten pages of explanation, the majority of the 164 pages of the official jōyō kanji list document are devoted to a fairly simple listing of the 2,136 kanji. The format is of four columns; one for the kanji (and variant forms), one for official jōyō pronunciations associated with the kanji, one for limited sets of word examples, and one for notes.

There is, however, a great deal more important information associated with jōyō kanji than the format of the official list implies. Accordingly, as a core component in developing a larger lexical database system to aid language researchers, we have constructed a database for the lexical properties of the revised jōyō kanji

[2] Developed under a five-year project (2006–2011) (NINJAL 2011, http://www.tokutei-corpus.jp/), BCCWJ is an approximately 100-million word corpus.

list. This section provides a very brief outline of the jōyō kanji database, primarily through a schematic image of its structure in Figure 2 and an example of summary frequency distributions for onyomi and kunyomi in Table 2. Given that a number of the data types at the single-kanji level interact with data types at both the sub-grapheme and compound-word levels within the larger lexical database system under construction and consistent with the notion of jōyō kanji as the core building blocks, Sections 4 and 5 deal with some of the data types that connect at these other levels.

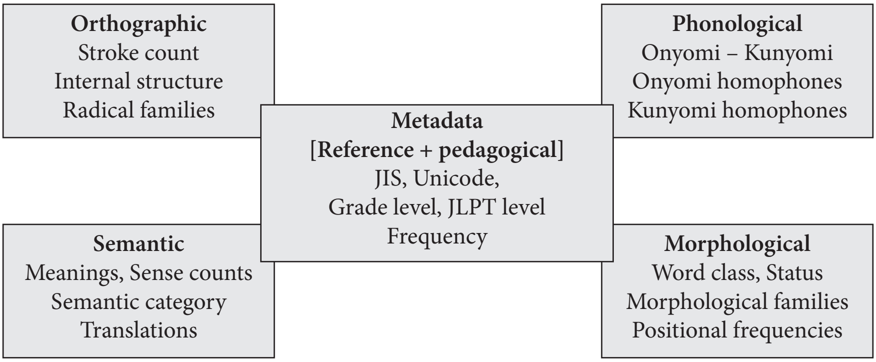


Figure 2. Schematic illustration of the jōyō kanji database structure

As Figure 2 shows, data within the jōyō kanji database is organized under five broad groupings. The first group of metadata includes reference identifications to various coding standards (such as JIS and Unicode) as well as data of pedagogical and usage relevance, such as kyōiku kanji grade levels, Japanese Language Proficiency Test (JLPT) levels, and frequency data from various sources. The second group of orthographic properties essentially interfaces to relevant parts of the kanji component database, as outlined in Section 4. The third group of phonological properties starts from the official onyomi and kunyomi associated with each kanji to listing homophones. The frequency distributions of onyomi and kunyomi presented in Table 2 are based in this data group. The fourth group is of semantic properties, which lists the range of meanings associated with each kanji and their semantic categories. The fifth group of morphological properties includes word class, status (free or bound morphemes), and information about the morphological families of a particular kanji and positional frequencies. This data group is of particular relevance to the larger lexical database under construction and to the orthographic structures of Japanese words to be discussed further in Section 5.

Table 2. Frequency distributions of onyomi and kunyomi for jōyō kanji

	Onyomi per kanji						
Kunyomi per kanji	**0**	**1**	**2**	**3**	**4**	**5**	**Total**
0	0	741	78	2	0	0	**821**
1	66	685	93	7	0	0	**851**
2	9	238	55	5	0	1	**308**
3	1	77	15	2	0	0	**95**
4	0	35	10	1	0	0	**46**
5	0	7	0	0	0	0	**7**
6	0	1	0	0	0	0	**1**
7	0	3	0	0	0	0	**3**
8	0	0	1	0	0	0	**1**
9	0	0	1	0	0	0	**1**
10	0	0	2	0	0	0	**2**
Total	**76**	**1,787**	**255**	**17**	**0**	**1**	**2,136**

[Note: Counts include special pronunciations]

The frequency distributions are essentially unaffected by the 2010 revision, which remain strongly skewed towards kanji associated with 0–2 kunyomi and 0–2 onyomi, representing 92% of jōyō kanji, such as the most frequent 741 kanji (35%) that have one onyomi and no kunyomi. There have, however, been slight increases in the numbers of kanji associated with the most frequent onyomi. For example, there are now 67 kanji associated with the onyomi of /kō/, compared to 64 prior to the revision, and 66 kanji linked with /shō/, compared to 63 before.

Having established the importance of the jōyō kanji set and briefly introduced the database of their lexical properties, the following two sections seek to further illustrate the role of kanji within the overall architectural framework of the Japanese writing system. More specifically, Section 4 discusses the internal structures of kanji based on newly conducted analyses of their components, while Section 5 moves to discuss the orthographic structures of Japanese words, based on coding the component scripts of the entries within both dictionary and corpus word lists.

4. Internal structures of kanji

4.1 Overview

Jōyō kanji vary greatly in terms of their visual complexity. At one extreme, the simplest is the one stroke kanji 一 /ichi/ meaning 'one', while, at the other extreme, the most complex jōyō kanji of 鬱 /utsu/ 'depression' has 29 strokes. The average number of strokes for jōyō kanji is 10.5.

Naturally, this scale of complexity is only possible because of the existence of internal structures, or patterns, which, in turn, relate to the principles of kanji formation that govern the combinations of kanji components (Joyce 2011). Traditionally, 214 components, mostly basic kanji and their variants, have been accorded special status of 部首 /bushu/ '(semantic) radical' for the purpose of organizing the entries within kanji dictionaries. For instance, Figure 3 shows the kanji 鯨 /kujira/ 'whale', which is a combination of 魚 /sakana/ 'fish', which is the radical component, and 京 /miyako/ 'capital'. 魚 is also the radical for one other jōyō kanji, 鮮 /sen/ and /aza-yaka/ 'fresh, vivid, clear', and 18 JIS1 kanji (鯵, 鮎, 鰯, 鰻, 鰍, 鰹, 鯉, 鮭, 鯖, 鮫, 鯛, 鱈, 鰭, 鮒, 鮪, 鱒, 鱗, and 鰐). It also appears in the right position of the jōyō kanji 漁 /asa.ru/ 'fishing; fishery' (but is not the radical) and the top position of JIS1 kanji 魯 /ro/ 'foolish; Russia'. Similarly, 京 occupies the left position of the jōyō kanji 就 /shū/ and /tsu.ku/ 'settle; take position' and appears in the right position of the jōyō kanji 涼 /ryō/ and /suzu.shii/ 'cool; refreshing' and of 3 JIS1 kanji (椋, 掠, and 諒) and in the bottom position of the jōyō kanji 景 /kei/ 'scenery; view'. Figure 3 also illustrates how both the two main components could conceivably also be further divided into smaller patterns that reoccur in a number of other kanji.

魚 = ク + 田 + 灬
鯨 = +
京 = 亠 + 口 + 小

Figure 3. Example of traditional component division and conceivable finer divisions into smaller reoccurring patterns

It should be noted, however, that the radical-based conventions for describing the internal structures are rather arbitrary in nature and provide no insights into the positional variability of the components within kanji. It is also debatable as to how informative it is to decompose complex characters beyond the radical level into smaller patterns of strokes. Accordingly, the following sub-section describes a new analysis of the revised jōyō kanji and JIS1 kanji according to three basic configurations.

4.2 Kanji component database

It is possible to adopt a number of different approaches towards the analysis of kanji into their components. At one extreme, one could essentially focus on the purely physical form in analyzing kanji according to various patterns from the whole kanji gestalt down to the smallest reoccurring patterns of strokes. At the other extreme, the analysis might seek to be as etymologically accurate as possible, drawing on specialist knowledge about structures and shape transitions, irrespective of any visual similarities that might exist. Alternatively, researchers might seek to strike some balance that incorporates the general knowledge of native speakers.

In spirit, two studies conducted by Saito, Kawakami and Masuda (1995, 1997) are probably more consistent with the first approach of emphasizing physical visual form. While both studies addressed the 2,965 JIS1 kanji, Saito et al. (1995) focused on the 1,668 kanji that they classified as having a left-right configuration of main components and Saito et al. (1997) reported on 807 kanji that they classified as conforming to a top-bottom configuration. However, the internal structures of the remaining approximately 17% of JIS1 kanji were not considered. In addition to concerns about the appropriateness of some of their classifications, as touched on more shortly, another minor shortcoming with their studies is that the 2010 jōyō kanji revision added 30 JIS2 kanji rendering their data somewhat obsolete.

In contrast, our analysis of the internal structures of kanji and the resultant database of kanji components and distributions seeks to embrace a meaningful middle approach. By both referring to scholarly analysis (Shirakawa 2012) and considering visual similarities between jōyō kanji, our analysis attempts to more faithfully reflect contemporary conventions concerning kanji instruction within Japanese schools. Accordingly, in addition to the two main left-right and top-bottom configuration utilized in Saito et al. (1995, 1997), our analysis also includes a third main category of enclosure-enclosed to capture a number of other traditional radical arrangements, as illustrated in Figure 4.

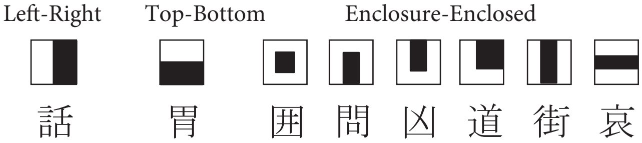


Figure 4. Three configurations employed by internal structure analysis with examples

By including the enclosure configuration and by adopting a more reserved attitude towards attempting to reduce every kanji down to one of just two basic configurations, it is possible to avoid some of the questionable classifications made by Saito et al. (1995). For example, although they treat 街 as a left-right configuration consisting of 彳 + rest, our analysis can more appropriately classify this as an enclosure configuration, consisting of the historically recognized 行構え /gyōgamae/ enclosure around the enclosed component of 圭. Moreover, from a reluctance to merely impose arbitrary configurations, such as on kanji that consist of three vertically-aligned elements but where there is no rationale for grouping the middle element with the top element over grouping it with the bottom element (for example, 魚 is classified as non-divisible rather than a particular top-bottom combination of three elements, as illustrated in Figure 3), our category of non-divisible kanji also includes a number of ambivalent kanji. Thus, while our analysis occasionally departs from traditional radical-based classifications, it seeks to achieve a more equitable treatment of kanji components, irrespective of their status as traditional radical and position.

Table 3. Results of analyzing the 2,136 jōyō and 2,965 JIS1 kanji according to the three configurations of left-right, top-bottom, enclosures, and a non-divisible category, together with total numbers of components

	2,136 jōyō kanji		**2,965 JIS1 kanji**	
Category	**Numbers**	**Percentage**	**Numbers**	**Percentage**
Non-divisible	185	8.7	218	7.4
Divisible	1,951	91.3	2,747	92.6
Left-right	1,166	54.6	1,664	56.1
Top-bottom	563	26.4	798	26.9
Enclosure-enclosed	222	10.4	285	9.6
Components	1,072		1,290	

Table 3 presents the analysis results for both jōyō and JIS1 kanji. In the case of jōyō kanji, 1,951 kanji (91.3%) are divisible according to the three configurations, while 2,747 JIS1 kanji (92.6%) are similarly divisible. Within both sets, over half of these divisible kanji conform to the left-right configuration, while more than one-quarter have a top-bottom configuration. Similarly, for both sets, enclosure-enclosed kanji is the smallest category representing about 10% of divisible kanji.

The divisible jōyō kanji consist of a total of 1,072 different components, while 1,290 components are combined in forming the JIS1 kanji. If it were necessary to remember and recognize all 2,136 jōyō kanji as mutually unrelated, visual objects, then the burden on memory would be quite considerable. However, reflecting general instructional practices that greatly emphasize internal structures and shared components, the actual load is arguably much lower, perhaps closer to 48%[3] based

[3] The reduction is even higher at 55.3% for the 2,747 divisible JIS1 kanji, which are formed from 1,290 kanji components (plus 35 of the 218 non-divisible JIS1 kanji).

on 1,111 separate forms (1,072 components plus 39 of 185 non-divisible kanji that are not used as components within any other kanji). Moreover, within the 1,072 jōyō kanji components, 316 (29.5%) are jōyō kanji themselves, with 77 appearing in left-right kanji, 192 in top-bottom kanji, and 47 in enclosure-enclosed kanji (and 146 are included within the non-divisible total). On average, the components of divisible jōyō kanji appear within 3.6 kanji, although the most frequent component is 氵, さんずいへん /san-zui-hen/ 'water radical', which appears on the left side of 112 left-right kanji.[4]

Table 4. Number of internal positions that components can occupy

Number of internal positions	Jōyō components		JIS1 components	
	Numbers	Percentage	Numbers	Percentage
1	761	71.0	867	67.2
2	207	19.3	263	20.4
3	63	5.9	95	7.4
4	28	2.6	43	3.3
5	12	1.1	21	1.6
6	1	0.1	1	0.1
Total	**1,072**	**100.0**	**1,290**	**100.0**

Table 5. Number of components that appear in the different configuration positions

	Left	Right	Top	Bottom	Enclosure	Enclosed
Jōyō kanji	215	598	244	257	55	173
JIS1 kanji	248	753	299	382	61	218

Tables 4 and 5 present summary data for the number of internal positions that components can occupy and the number of components that appear in each of the 6 possible internal positions (3 configurations x 2 components), respectively. Thus, Table 4 shows that for both jōyō and JIS1 components, approximately 70% only appear in one configuration position, while approximately 20% can appear in two positions. Only the 口 component can occupy all 6 configuration positions. Table 5 indicates that for both jōyō and JIS1 kanji, there are more right components than left components, more bottom components than top components,

[4] For JIS1 kanji components, the average is 4.3 kanji and the most frequent component is 木 'tree', which appears in 164 JIS1 kanji, either as the left, right, top, bottom, or enclosed element.

and more enclosed components than enclosure components. These findings are consistent with the fact that within traditional radical-based classifications, left components, top components, and enclosure components were generally used as the radical component.

5. Orthographic structures of Japanese words

5.1 Overview

As Joyce (2002, 2011) argues, the most meaningful orthographic classification of Japanese kanji is that they are essentially morphographic in nature. A number of kanji can stand alone as free morphemes (i.e. simplex words), such as 嵐 /arashi/ 'storm, tempest' recently added to the jōyō kanji list. However, in most cases, kanji are used in combination with other morpheme representations. In the case of kanji associated with Native-Japanese morphemes of verbal, adjective, and adverb meanings, the kanji represents the stem morpheme and is combined with 送り仮名 /oku.ri.ga-na/ referring to the hiragana script representation of the other conjugational morphemes (although, admittedly, the exact placement of the morpheme boundary can be more problematic). Figure 5 seeks to illustrate this by showing the orthographic representations for four conjugations of the Native-Japanese verb 書く /ka.ku/ 'write'.

Orthography	Phonological gloss	Meaning	Conjugation
書く	ka.ku	write	plain, present, affirmative
書きます	ka.kimasu	write	polite, present, affirmative
書かない	ka.kanai	not write	plain, present, negative
書きません	ka.kimasen	not write	polite, present, negative

Figure 5. Orthographic representation of four conjugations of the Japanese verb 書く

Similarly, the Native-Japanese of 暗い /kura.i/ 'dark' is a combination of a kanji representing the stem morpheme and a hiragana character for the conjugation ending of /i/ indicating the plain, present, affirmative (Japanese *i*-ending adjectives conjugate for tense and aspect).

Compounding is also a highly productive process of word formation involving both the Sino-Japanese and Native-Japanese stratums of the lexicon. In introducing the distinction between onyomi and kunyomi, Section 2 noted that 使用 /shiyō/ meaning 'use' is a combination of two Sino-Japanese morphemes both associated with similar meanings (使 /shi/ 'to use' + 用 /yō/ 'to use'). According to Nomura (1988), two-kanji compound words – referring, more accurately, to two-morpheme compound words orthographically represented by the two kanji that represent the respective morphemes – are the most common word structure in the Japanese language. Naturally, other combinations are also possible in forming longer compound words. Even two years after the Great Tohoku Earthquake in Japan, predictive input on entering 使用 into an internet search engine offers up the compound word of 使用済み核燃料 /shi-yō-zu.mi.kaku-nen-ryō/ 'spent nuclear fuel' as a candidate search term. Figure 6 attempts to illustrate the layers of compounding processes that underlie this compound word.

使 'use' + 用 'use' 済み 'finish, spent' 核 'nuclear' 燃 'burn' + 料 'material'
使用済み 'used, spent' 核 'nuclear' + 燃料 'fuel'
使用済み 'used, spent' + 核燃料 'nuclear fuel'

Figure 6. Analysis of compounding processes underlying 使用済み核燃料

While it is true that the component scripts of the Japanese writing system are generally employed in complementary ways, in reality, the situation is somewhat more complex, because it is also true that orthographic variation is a major characteristic of the Japanese writing system, at least, for more common Japanese words. In exploring the complex relationships between units of language and units of writing within the content of the Japanese writing system, Joyce et al. (2012) quantitatively analyzed the degree of orthographic variation within their corpus word lists created from the BCCWJ. However, in introducing one of their relevant findings, at this point, it is also expedient for subsequent discussions to briefly note the two lexical-units of short-unit words (SUWs) and long-unit words (LUWs) employed within the BCCWJ project. Although the labels themselves evoke notions of length, the distinction is really more about lexical status (such as between free and bound morphemes (i.e. affixes)), for SUWs are mainly basic words, or dictionary headwords, while LUWs are basically complex words and phrases (see also Joyce et al. (2012) for further discussion of issues related to these concepts). Now, returning to issue of orthographic variation, Joyce et al. (2012) found that, for the most frequent 100 lemmas in the four main word lists of nouns, verbs, *i*-adjectives, and adverbs, the average number of orthographic variations across the SUW word lists is 8.44 (min = 6.46; max = 10.19). For instance, there are five orthographic variants for 玉葱 /tama-negi/ 'onions' in the SUW noun list. As the second kanji is not a jōyō kanji, the five variants in descending order of frequency are 玉ねぎ (1,026 occurrences; 0.49), タマネギ (446 occurrences; 0.21), たまねぎ (345 occurrences; 0.17), 玉葱 (172 occurrences; 0.08) and 玉ネギ (94 occurrences; 0.05).

Even from this short outline of the orthographic structures of Japanese words, it should be clear from both the multi-script nature of the Japanese writing system and the high productivity of compounding processes within the Japanese language that there is a great deal of variation in the orthographic structures of the Japanese lexicon. The following sub-section reports on our efforts to analyze the orthographic structures of Japanese words.

5.2 Orthographic structure data

Within our construction of the large-scale lexical database system, we are utilizing a number of lexical resources, including various dictionaries and smaller existing databases. For instance, reflecting its encyclopedic nature, one of the more authoritative dictionaries of the Japanese language is Iwanami's (2008) 広辞苑 /kō-ji-en/ 'Kōjien' dictionary. Although the sixth edition has 232,795 headword lines, after excluding kanji that are not part of the union of the jōyō and JIS1 kanji sets, we have created a list of 215,597 headwords. Another core resource is the BCCWJ-based word lists created by Joyce et al. (2012). In the interests of brevity, summary data for the corpus word lists is combined with the orthographic coding results in Table 8 provided shortly.

In order to analyze the orthographic representation of Japanese words in terms of their orthographic structures, an orthographic code was attached to all headword entries within both word lists. Examples of some of these orthographic codes are shown in Table 6.

Table 6. Some orthographic code examples

Length	Examples
1	嵐 = C; は [/wa/ topic marker] = H
2	漢字 = 2C; 書く = CH; もの = 2H; ヒト = 2K
3	核燃料 = 3C; 食べる = C2H; 山登り = 2CH; しかし = 3H; イルカ = 3K
4	漢字使用 = 4C; 使用済み = 3CH; 書きます = C3H

Note: Basic codes are C = kanji, H = hiragana, K = katakana.

Analysis of the Kōjien list reveals a total of 1,152 separate orthographic codes, although 578 (50.2%) are unique with a token frequency of one. More specifically, Table 7 presents the 10 most frequent orthographic codes for the list, which indicates that two-kanji compound words account for 37.5% of the list, followed by 3C and 4C compounds that together account for 61.5%. Moreover, all of the nine most frequent orthographic codes are either all kanji orthographic representations or kanji plus hiragana representations; only the tenth most frequent orthographic code of 4K does not involve kanji.

Table 7. Ten most frequent orthographic codes observed within the Kōjien headword listing

Code	Token frequency	Percentage
2C	80,949	37.5
3C	32,614	15.1
4C	19,245	8.9
2CH	8,916	4.1
CHC	5,604	2.6
CHCH	4,688	2.2
C	4,625	2.1
5C	4,495	2.1
CH	4,394	2.0
4K	3,469	1.6

However, it is vital to immediately acknowledge a few caveats relating to the overall balance or relevance of the Kōjien list given its broader encyclopedic scope. The first is that it includes 19,328 head words (approximately 9%) that are foreign names or loan words where the main entry field is in rōmaji transcription. A second point is the Kōjien's convention to provide kanji orthography in the main entry field even if another orthographic representation is actually more common, such as 海豚 /iruka/ 'dolphin' (literally 'sea' + 'pig') rather than the more common katakana representation of イルカ. Consistent with its encyclopedic nature, a third concern is for balance given the considerable numbers of technical terms and historical words, such as from Japanese and Chinese classics, that the Kōjien covers and which predominately have kanji orthography main entries. While all these points undeniably raise certain concerns both for the relative value of utilizing the Kōjien dictionary list as a prime source within the larger database project and about how representative these results alone might be about the orthographic structures of the Japanese lexicon, still, the general pattern is consistent with the results obtained from the corpus word lists, which are introduced next.

Given the acknowledged issues with the list of Kōjien headwords and because the corpus-based word lists allow for comparisons of type and token frequencies, a little more attention is given to presenting summaries of the orthographic coding data for the corpus word lists, starting with Table 8 which presents a number

Table 8. Token and various type counts for both SUW and LUW word lists

Count	SUW	LUW
Tokens	104,344,054	83,290,629
Lemma types	173,010	2,359,295
Orthographic codes for lemma types	242	42,226
Unique orthographic codes for lemma types	37 (15.3%)	33,073 (78.3%)
Orthographic form types	234,821	2,480,161
Orthographic codes for orthographic form types	680	50,914
Unique orthographic codes for orthographic form types	236 (34.7%)	39,935 (78.4%)

Note: Status of unique orthographic codes based on tokens/types = 1. These counts do not include the symbols list from Joyce et al. (2012).

of count totals. The first important point to stress is that different notions of the word are being used in the Kōjien dictionary list (i.e. 215,597 headwords) and the corpus word lists. To a large extent, this reflects the BCCWJ's policy of differentiating between SUWs and LUWs and the serious underlying issues of defining the orthographic word for highly agglutinative languages like Japanese (Joyce et al. 2012; NINJAL 2011).

A second important comment to make concerns the distinction between lemma and orthographic form. Within the parsing dictionary developed as part of the BCCWJ corpus project for annotating the corpus, the term 'lemma' essentially refers to the headword entry, while orthographic forms cover the orthographic variants of the lemma (NINJAL 2011). Just as there are more orthographic forms compared to lemmas, for both the SUW and LUW lists, there are more different kinds of orthographic codes for the orthographic forms types compared to the lemma types. In the case of the SUW lists, although there are 242 orthographic codes for lemma types, there are 680 different orthographic codes for the orthographic form types. It is also imperative to note that the four different counts of different orthographic codes (for lemmas and orthographic forms for both SUW and LUW lists) all include considerable amounts of unique orthography codes (with token frequency of one), ranging from 15.3% for SUW lemma types to approximately 78% for both the LUW types. Still, the 205 different orthographic codes for SUW lemma types (total of 242 minus 37 unique codes) represent a level of variety in terms of the orthographic structures of Japanese words that is rather inconceivable from the perspectives of alphabetic writing systems. Tables 9 and 10 present the ten most frequent orthographic codes for lemma types as sorted by type and token counts for the SUW and LUW lists, respectively.

Table 9. Ten most frequent orthographic codes for lemma types observed within the SUW corpus word lists as a function of both token and type counts

By tokens			By types		
Code	Tokens	Percentage	Code	Types	Percentage
H	36,085,085	34.6	2C	61,299	35.4
2C	19,563,921	18.7	4K	26,380	15.2
C	15,944,686	15.3	5K	15,453	8.9
CH	14,262,240	13.7	3K	14,377	8.3
2H	6,137,554	5.9	6K	9,127	5.5
C2H	3,130,369	3.0	2CH	5,154	3.0
3K	1,780,241	1.7	7K	4,616	2.7
4K	1,701,708	1.6	CHCH	4,600	2.7
2CH	901,607	0.9	CHC	3,071	1.8
3H	900,752	0.9	CHC2H	2,599	1.5

Table 10. Ten most frequent orthographic codes for lemma types observed within the LUW corpus word lists as a function of both token and type counts

By tokens			By types		
Code	Tokens	Percentage	Code	Types	Percentage
H	30,336,066	36.4	4C	362,317	15.4
2C	10,017,730	12.0	3C	220,257	9.3
CH	8,280,353	9.9	5C	210,503	8.9
C	6,383,794	7.7	6C	139,538	5.9
2H	6,375,684	7.7	2C	102,876	4.4
C2H	3,551,506	4.3	7C	70,747	3.0
3H	2,820,616	3.4	8C	39,639	1.7
3C	2,619,383	3.1	4K	23,525	1.0
3K	1,765,535	2.1	7K	22,967	1.0
4K	1,668,133	2.0	6K	22,490	1.0

As Tables 9 and 10 provide a wealth of interesting insights into the orthographic structures of Japanese words, it is not possible to unravel every detail within this paper, but a few basic observations can serve to highlight the significance of the jōyō kanji set within the Japanese writing system. Focusing firstly on the token data (on left of both tables), the overall patterns are fairly similar, with most of the same orthographic structures appearing within the most frequent top ten and in similar rank orders. The most frequent orthographic structure in both cases is the single hiragana orthography word. This is consistent with expectations in naturally reflecting the functional demarcation between content words and grammatical words, as many grammatical functional morphemes are mono-mora morphemes represented orthographically by a single hiragana. For instance, these include many of the closed word class of 助詞 /jo-shi/ 'grammatical markers', such as の /no/ 'possessive or modification marker' and も /mo/ 'too' (inclusion marker), and conjugational elements, such as い /i/ 'adjectival ending'. The finding that the second most frequent orthographic structure by tokens is the two-kanji compound word is also exactly as one might expect, given that this orthographic structure is by far the most frequent among the Kōjien dictionary list of headwords.

Turning next to the type data (on right of both tables), the first point is to acknowledge that the most frequent orthographic codes are somewhat different for the SUW and LUW word lists, which is, once more, a reflection of the fundamental difference between the SUW and LUW units. Looking first at the most frequent orthographic structures for the SUW lists the most frequent is the 2C compound word, which is consistent with the dominance of this orthographic structure within the Kōjien word list and its second position within the token rankings. The high positions of katakana orthography words, from second to fifth most frequent (4K, 5K, 3K, and 6K, respectively), is testimony to the openness of the Japanese language to foreign loanwords, particularly loanwords from the English language. It should be also noted, however, that although these orthographic structures are very common by type, only the 3K and 4K orthographic structures appear within the top ten most frequent by tokens (at positions seven and eight, respectively).

The second key observation to be drawn from the type data relates to the productivity of compounding, noted earlier, which is clearly reflected in the type rankings for both the Kōjien and the LUW word list. In the case of the Kōjien list, the 2C compound word is the most frequent at 37.5%, but many of the 3C and 4C compound words – second and third most frequent orthographic structures at 15.1% and 8.9%, respectively – will be 2C compound words with either another one or two morphemes combined. The importance of compounding is even more obvious in the case of the LUW corpus list, for the seven most frequent structures are all kanji orthography words, ranging in descending order from 4C, 3C, 5C, 6C, 2C, 7C to 8C and together accounting for 48.6% of all the orthographic structures of LUWs.

6. Conclusion

The value of a metaphor lies in its power to capture and convey something of profound significance about the phenomenon in question. Much of the appeal of the building blocks metaphor for thinking about writing systems is its compatibility with linguistic axioms about the structured nature of language. In contemplating on the application of the metaphor to the Japanese writing system, our primary source of inspiration has been to conceive of the jōyō kanji list as the core building blocks for the orthographic representation of most Japanese words. The schematic illustration in Figure 7 attempts to encapsulate the intricate embedded structural relationships that this paper has sought to elucidate.

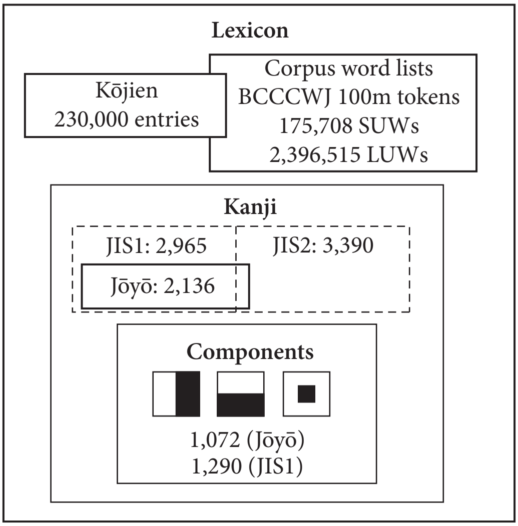


Figure 7. Schematic illustration of jōyō kanji as core building blocks of Japanese writing system

Our survey of these layers of structure has largely been informed by a number of observations and insights gained from our ongoing project to construct a large-scale lexical database system for the Japanese language. Mirroring both the architectural principles of the Japanese writing system and the organizational structure of the lexical database system, our discussions consisted of three main sections.

Section 3 was concerned with the jōyō kanji list and its 2010 revision. Rather than reflecting a drastic shift in kanji policy, however, the slight increase to 2,136 kanji for daily usage should be seen as more of a temporal readjustment. The central importance of the jōyō kanji list for general Japanese media and communication was underscored by coverage data derived from the corpus word lists created by Joyce et al. (2012), where even for a period mainly sampled before the 2010 revision, the 2,136 jōyō kanji account for 96.12% of all kanji tokens within the BCCWJ corpus. Accordingly, the jōyō kanji database and its interfaces is an integral part of our large-scale lexical database system. As Section 3.2 sought to outline, the jōyō kanji database includes various forms of metadata, orthographic, phonological, semantic, and morphological properties, such as the frequency distributions of jōyō kanji pronunciations summarized in Table 2.

Section 4 addressed the internal structures of kanji. After briefly explaining the traditional radical classification system, Section 4.2 introduced the kanji component database which is the product of a new analysis of jōyō and JIS1 kanji according to three basic configurations of left-right, top-bottom, and enclosure-enclosed patterns. The results clearly highlight the internally-structured nature of kanji, because approximately 90% of the kanji in both sets conform to the three basic configurations and the 1,951 divisible jōyō kanji are constituted from 1,072 components.

Finally, Section 5 focused on the orthographic structures of the Japanese lexicon. More specifically, Section 5.2 introduced the results of analyzing the orthographic structures observed within both a list of headwords extracted from the Kōjien dictionary and the corpus word lists created by Joyce et al. (2012) by applying orthographic codes. The ranking results for the orthographic codes of the corpus word lists by tokens are consistent with the functional demarcation between content words, the building blocks of sentences, and grammatical words, the cement, which is a basic feature of the Japanese writing system's utilization of multiple scripts. However, reflecting the morphographic nature of kanji (Joyce 2011), the orthographic structure data for the Kōjien dictionary and, especially, the type ranking of LUW lists provide unequivocal testimony for two fundamental phenomenons; namely that kanji orthography words dominate within the Japanese writing system and that compounding is a highly productive principle of word formation for the Japanese lexicon.

Applying the architecture of writing systems metaphor to thinking about the Japanese writing system has been particularly fruitful. With its inspiration to conceptualize the jōyō kanji list as the core building block, this paper has sought to illuminate the key layers of structure within the Japanese writing system, in general, and within Japanese kanji, in particular, with pertinent observations drawn from an ongoing project to construct a large-scale Japanese lexical database system.

References

Bunkachō [Japanese Agency for Cultural Affairs] (2010). Jōyōkanjihyō [Jōyō kanji list]. http://www.bunka.go.jp/kokugo_nihongo/pdf/jouyoukanjihyou_h22.pdf (accessed 10th March 2013).

Coulmas, Florian (1989). *The writing systems of the world*. Oxford: Basil Blackwell.

Iwanami (2008). *Kōjien* [Kōjien dictionary]. 6th edition. Tokyo: Iwanami Shoten.

Joyce, Terry (2002). The Japanese mental lexicon: the lexical retrieval and representation of two-kanji compound words from a morphological perspective. Unpublished doctoral thesis. University of Tsukuba, Japan.

Joyce, Terry (2011). The significance of the morphographic principle for the classification of writing-systems. In Susanne R. Borgwaldt & Terry Joyce (eds.), *Typology of writing systems*. Special issue of *Written Language and Literacy* 14.1: 58–81.

Joyce, Terry, Bor Hodošček & Kikuko Nishina (2012). Orthographic representation and variation within the Japanese writing system: some corpus-based observations. In Terry Joyce & David Roberts (eds.), *Units of language – units of writing*. Special issue of *Written Language and Literacy* 15.2: 254–278.

Nomura, Masaaki (1988). Niji kango no kozo [The structure of two-kanji Sino-Japanese words]. *Nihongogaku* [Study of Japanese Language] 7: 44–55.

Morohashi, Tetsuji (Chief-editor) (1955–1960). *Daikawajiten* [Comprehensive Chinese-Japanese dictionary], 13 vols. Tokyo: Taishukan.

National Institute for Japanese Language and Linguistics (NINJAL) (2011). *Gendai nihongo kakikotoba kinkō kōpasu* [Balanced corpus of contemporary written Japanese]. [Data DVD]. Tokyo: Center for Corpus Development, National Institute for Japanese Language and Linguistics.

Saito, Hirofumi, Masahiro Kawakami & Hisashi Masuda (1995). Kanji kōsei ni okeru buhin (bushu) no shutsugen hindō hyō [Frequency of semantic and phonetic components of radical types in complex left-right kanji]. *Studies in Informatics and Sciences* 1: 113–134.

Saito, Hirofumi, Masahiro Kawakami & Hisashi Masuda (1997). Jōge bunri kanji ni okeru buhin (bushu) no shutugen hindō hyō II [Frequency of components of radical types in complex (top-bottom) kanji]. *Studies in Informatics and Sciences* 6: 115–130.

Shirakawa, Shizuka (2012). Jōyō jikai [Analysis of Jōyō kanji]. 2nd edition. Tokyo: Heibonsha.

Smith, Janet S. (1996). Japanese writing. In Peter T. Daniels & William S. Bright (eds.), *The world's writing system*, 209–217. New York: Oxford University Press.

Taylor, Insup & Kwonsaeng Park (1995). Differential processing of content words and function words: Chinese characters vs. phonetic scripts. In Insup Taylor & David R. Olson (eds.), *Scripts and literacy: Reading and learning to read alphabets, syllabaries and characters* (Neuropsychology and cognition 7), 185–195. Dordrecht, Boston and London: Kluwer.

Twine, Nanette (1991). *Language and the modern state: The reform of written Japanese*. London: Routledge.

Corresponding author

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

terry@tama.ac.jp