Orthographic representation and variation within the Japanese writing system

Some corpus-based observations

Terry Joyce, Bor Hodošček & Kikuko Nishina
Tama University / Tokyo Institute of Technology /
Tokyo Institute of Technology, Japan

Given its multi-scriptal nature, the Japanese writing system can potentially yield some important insights into the complex relationships that can exist between units of language and units of writing. This paper discusses some of the difficult issues surrounding the notions of orthographic representation and variation within the Japanese writing system, as seen from the perspective of creating word lists based on the Kokuritsu Kokugo Kenkyūjo's 'Balanced Corpus of Contemporary Written Japanese' (BCCWJ) Project. More specifically, the paper (i) reflects on the treatment of lemmas within UniDic, the morphological analyzer dictionary developed for the project, (ii) notes some concerns for extracting word lists that stem from the project's approach towards defining orthographic words which draws on its conceptualization of short and long unit words, and (iii) attempts to quantify the extent of orthographic variation within the Japanese writing system as represented by the BCCWJ.

Keywords: Japanese; Balanced Corpus of Contemporary Written Japanese (BCCWJ); kanji; hiragana; katakana; orthographic variation; UniDic

1. Introduction

In expressing their language in writing, the Japanese employ a mixture of script types. Thus, the Japanese writing system consists of morphographic kanji, the two syllabographic sets of hiragana and katakana, alphabetic rōmaji, Arabic numerals, and various other symbols (Joyce 2011; Smith 1996; Tranter 2008). Depending on one's perspective, this unique mixture of graphic elements can be seen either as the worst possible solution to written language (DeFrancis 1989) or as potentially supporting great orthographic flexibility (Backhouse 1984). Whichever position one favors, unquestionably, the Japanese writing system provides an exceptional

Written Language & Literacy 15:2 (2012), **254–278**. DOI 10.1075/wll.15.2.07joy
ISSN 1387–6732 / E-ISSN 1570–6001

case study for investigating the complex relationships between units of language and units of writing.

Following a description of the principal components of the Japanese writing system and a brief outline of the 現代日本語書き言葉均衡コーパス /gendai nihongo kakikotoba kinkō kōpasu/'Balanced Corpus of Contemporary Written Japanese' (BCCWJ) Project[1] at 国立国語研究所 /Kokuritsu Kokugo Kenkyūjo/ 'National Institute for Japanese Language and Literature' (NINJAL) (Maekawa 2007; NINJAL 2011a), this paper discusses some issues concerning orthographic representation within the Japanese writing system. More specifically, from the perspective of creating some word lists from the BCCWJ, the paper reflects on (i) the treatment of lemmas by UniDic, a morphological analyzer dictionary (Den et al. 2007, NINJAL 2011a) developed within the BCCWJ Project, (ii) some problems for extracting the word lists due to the definitions of short-unit words (SUWs; essentially morphemes) and long-unit words (LUWs; closer to phrases), and, drawing on the word lists, (iii) capturing the extent of orthographic variation within the contemporary Japanese writing system as represented by the BCCWJ. As these are all intricate issues and their full discussion is beyond the scope of this paper, the paper primarily seeks to shed some light on the problems and to note the BCCWJ as an essential resource for their further investigation.

2. The Japanese writing system

The Japanese use a mixture of script types to express their language in writing (Gottlieb 2008; Kess & Miyamoto 1999; Joyce 2011; Smith 1996; Taylor & Taylor 1995; Tranter 2008). The principal component scripts are morphographic 漢字 /kanji/ 'Chinese characters' and the two 'kana' syllabographic sets of 平仮名/ひらがな /hiragana/ and 片仮名/カタカナ /katakana/, which are supplemented with alphabetic ローマ字 /rōmaji/ 'Roman alphabet', Arabic numerals, and various other symbols. Commenting on the mixture of kanji, kana, and the alphabet, Kess and Miyamoto (1999:9) observe that the "Japanese may have the unique distinction of employing all three extant means" of expressing language in writing. This part of the paper offers a few remarks about each of the component scripts and some comments about the conventions that guide how the elements are employed in largely separate and complementary ways within an overall system.

2.1 Morphographic kanji

The process by which one culture learns of writing from a neighboring culture and borrows their writing system is one that has been repeated many times throughout

history, and the Japanese number among the vast majority of peoples who did not invent writing for themselves. For Japan, the neighboring culture was China, the writing system was Chinese characters, and the borrowing probably started as early as the first century CE[2] (Miller 1967; Seeley 1991; Shibatani 1990).

Historically, a large number of kanji have been used within the Japanese writing system,[3] but since the middle of the 20th century, the Japanese government, through the Ministry of Education, has provided a series of guidelines concerning kanji usage that have generally sought to reduce the number of kanji in daily use and to simplify some kanji forms. In 1946, the 当用漢字表 /tōyō kanjihyō/ list was issued which consisted of 1,850 kanji. In October 1981, the 常用漢字表 /jōyō kanjihyō/ 'List of characters for general use' was issued which consisted of 1,945 kanji. However, the jōyō kanji list was further modified in November 2010, by removing five characters and adding 196 new kanji, to create a new official list of 2,136 kanji. The jōyō kanji list consists of 1,006 教育漢字 /kyōiku kanji/ 'education kanji' which are taught during the six years of elementary school and 1,130 kanji which are taught at high-school.

To put the number of 2,136 jōyō kanji in clearer perspective, however, it must be noted that this is not an upper limit on the number of kanji used within the modern Japanese writing system because the jōyō kanji list is only a guideline and is less prescriptive in nature than the earlier tōyō kanji list. A stronger candidate for an upper limit figure would be the 6,355 kanji of the Japanese Industrial Standard (JIS) code (JIS X-0208-1990 includes 2,965 level 1 kanji and 3,390 level 2 kanji) (Lunde 1993), which defines the character sets for computers and electronic communication devices such as cell phones. On the other hand, newspapers and official documents generally follow the jōyō kanji guidelines and it has been estimated that the list covers over 98% of all newspaper kanji tokens, with 90% being covered by the 1,006 kyōiku kanji (Kess & Miyamoto 1999). The small percentage not covered is mainly due to kanji outside the list used in family and place names, which keep the total number of kanji used in newspapers at approximately 3,200 to 3,300 (Seeley 1991; Smith 1996).

Kanji vary considerably in terms of their visual complexity. Within the 2,136 jōyō kanji, there are 14 kanji that consist of just one or two strokes, as shown in (1), but there are also 11 kanji that consist of 21 or more strokes, as shown in (2). The average number of strokes for the jōyō kanji is 10.47 (SD = 3.79, min = 1, max = 29).

(1) 一 乙 九 七 十 人 丁 刀 二 入 八 又 了 力

(2) 艦 顧 魔 躍 露 鶴 驚 襲 籠 鑑 鬱

A vital key to understanding how the Japanese writing system functions lies in understanding how Japanese kanji, as units of writing, relate to units of

language. As Joyce and Borgwaldt (2011) observe, one of the most enduring concerns for typologies of writing systems has been to meaningfully characterize the elusive distinction between phonographic, or cenemic, writing systems from non-phonographic, or pleremic, writing systems. Consistent with Joyce's (2011) argument that it is more appropriate to refer to pleremic writing systems as morphographic than logographic, it is fundamental to appreciate the morphographic nature of Japanese kanji.

It is instructive to look at the two ways in which the Japanese people came to use Chinese characters for writing their own vastly different language when they borrowed writing. As Martin (1972) observes, they were to have far-reaching consequences, for they are essentially how kanji are still employed today, resulting in the dual system of 音読み /on'yomi/ 'Sino-Japanese pronunciations' and 訓読み /kun'yomi/ 'Native-Japanese pronunciations' associated with kanji. At first, 'writing' in Japan meant writing in Chinese following Chinese syntax.[4] However, that did not last long, for the Japanese set up conventions, known as 訓読 /kundoku/, that indicated reading order and pronunciations for reading classical Chinese according to Japanese syntax (Habein 1984; Miller 1967). Of course, once that happened, the language was no longer Chinese but Japanese and the Chinese words and morphemes represented by the characters entered the Japanese language as loan words and morphemes, with Japanese imitations of the Chinese readings. The first way in which the Japanese came to use Chinese characters was, therefore, to write these new loan words and morphemes.[5] For instance, the kanji 読 meaning 'read' has a basic Sino-Japanese pronunciation of ドク /doku/.

Although many Chinese words and morphemes for things entered the Japanese language as on'yomi for kanji, the Japanese people already had their own words for many of those things. So, the second method of using kanji was to associate the kanji with these native Japanese words referring to the same things. Thus, 読 became linked to the stem of the verb 読む /yo.mu/ meaning 'to read'.[6] Accordingly, a basic feature of the Japanese lexicon is a distinction between Sino-Japanese words, words borrowed from Chinese, and Native-Japanese words.[7] So, for instance, in addition to a Native-Japanese phrase of 本を読む /hon o yomu/ meaning 'to read a book', there is a Sino-Japanese phrase of 読書する /dokusho suru/ meaning 'to read (books)', which consists of 読 combined with 書 /sho/ 'writing' to form the noun of 読書 '(book) reading' and する /suru/ which means 'to do'.

2.2 Syllabographic katakana and hiragana

During the ninth century, the Japanese developed the two native syllabaries of katakana and hiragana. Although they were developed separately, both were based on the use of kanji in phonetic transcription, as employed in the 万葉集

/man'yōshū/ (759 CE), an anthology of Japanese verse, and consequently, characters used in this way are often referred to as 万葉仮名 /man'yōgana/ (Miller 1967; Shibatani 1990).[8] While the development of the kana scripts made it possible to write the sounds of the Japanese language without the use of kanji – as indeed women did during the Heian period (794–1185) – the kana scripts did not replace the use of kanji completely, but led to the mixed system of modern Japanese writing.

The symbols of katakana and hiragana represent syllables, or more precisely, morae as syllables of equal duration. Both kana sets consist of 46 basic characters, such as katakana ア /a/ and カ /ka/ and hiragana あ /a/ and か /ka/. The basic set is extended to 71 characters by adding diacritics: either ゛ called 濁音 /dakuon/ 'voiced', where カ + ゛ → ガ /ga/ and か + ゛ → が /ga/, or ゜ called 半濁音 /handakuon/ 'semi-voiced', where ハ /ha/ + ゜ → パ /pa/and は /ha/ + ゜ → ぱ /pa/. The sets are further extended to 107 characters by combining certain basic characters in combinations known as 拗音 /yōon/ 'contracted sounds', such as カ + a reduced ヤ /ya/ → キャ /kya/ and か + a reduced や /ya/ → きゃ /kya/.

Katakana is mainly used for foreign names, loan words, emphasis, onomatopoeia, and to indicate species within scientific taxonomies, as in (3) to (7), respectively.

(3) イグナス・ゲルブ /igunasu・gerubu/ 'Ignace Gelb' (grammatology scholar)

(4) テキスト /tekisuto/ 'text'

(5) イタイ・イタイ /itai-itai/ 'painful' (rather than 痛い・痛い for emphasis)

(6) サクサク /sakusaku/ 'crunchy, crisp, flaky'

(7) ヒト /hito/ 'homo sapiens' (as opposed to 人 /hito/ 'man, people')

In contrast, hiragana is mainly used for functional Japanese words, such as conjunctions and particles (8–9), and inflectional elements of verbs and i-adjectives,[9] known as 送り仮名 /okurigana/ (10–11).

(8) しかし /shikashi/ 'but, however'

(9) から /kara/ 'from'

(10) 飲みました /no.mimashita/ 'drank' (past polite form)

(11) 珍しかった /mezura.shikatta/ 'unusual' (past plain form)

2.3 Alphabetic rōmaji and Arabic numerals

The Japanese writing system also includes rōmaji. There are two main conventions for transliterating Japanese with rōmaji: ヘボン式 /hebonshiki/ 'Hepburn

system' (adopted in this paper), which was proposed by the American missionary James Curtis Hepburn (1815–1911), and 訓令式 /kunreishiki/ 'Cabinet Ordinance System', which was introduced by the Japanese government in 1954.[10] While the name of Japan's famous mountain is rendered as /fujisan/ in the Hepburn system, it is transliterated as /huzisan/ according to the Cabinet Ordinance System. Rōmaji is used as a supplementary gloss on public transport systems, such as railway and highway signs, and is very common in advertising and the mass media. Although Arabic numerals are frequently used within horizontally-arranged text, particularly scientific texts, kanji for numbers are also used, especially in vertically-arranged text.

2.4 A multi-script writing system

The standard orthographic convention of modern Japanese writing, known as 漢字かな混じり文 /kanji-kana-majiribun/ 'mixed kanji and kana writing', is to employ the various scripts in largely separate and complementary ways to represent the Japanese language in writing, as illustrated in (12).

(12) 日本語の表記システムには漢字、ひらがな、カタカナ、ローマ字が使い分けられている。
/Nihongo no hyōki shisutemu ni wa kanji, hiragana, katakana, rōmaji ga tsukaiwakerarete iru/
'Within the Japanese writing system, kanji, hiragana, katakana, and rōmaji are used in separate ways'

Kanji are used to write nouns, either singularly or in compounds (e.g. 日本語 /nihongo/ Japanese language), as well as the stems of verbs (e.g 使 /tsuka/ 'use' and 分 /wa/ 'divide' in 使い分けられている /tsukawakerareteiru/ 'use in separate ways', indicating the senses of 'use' and 'separately', respectively), and i-adjectives. Hiragana is used for inflectional elements (e.g the okurigana of 使い分けられている) and other grammatical elements (e.g には /ni wa/ 'within'). Katakana is used typically for foreign names and loan words (e.g システム /shisutemu/ 'system'), species names for animals and plants, onomatopoeic expressions, and for emphasis.[11]

The multi-scriptal nature of the Japanese writing system can be seen as either a very poor solution to the problem of written language or as allowing for great orthographic flexibility. Indeed, both positions have been championed. Taking the more negative view, DeFrancis (1989: 138) has referred to the Japanese writing system as "one of the worst overall systems of writing ever created", while, more upbeat in tone, Backhouse (1984: 220) has astutely observed that this mixture of scripts "makes for a potential flexibility of orthography on a scale that is inconceivable in the case of more familiar writing systems". While it is beyond the scope of this paper to debate which is the more valid appraisal, the present paper seeks to capture something of the extent of orthographic variation within the Japanese

writing system from word lists created from the BCCWJ and to reflect on some issues of orthographic representation, which can potentially highlight the complex relationships between units of language and units of writing.

3. Outline of the BCCWJ

Developed under a five-year project from 2006–2011 (Maekawa 2007; NINJAL 2011a),[12] the objective of the BCCWJ project was to compile a tagged corpus of contemporary written Japanese language that is comparable in terms of scale and coverage of various sub-corpora to the British National Corpus (BNC).[13] The BCCWJ project was structured through a number of special-concern groups, including data handling, tools and annotation, UniDic, Japanese linguistics, and language policy.[14]

The BCCWJ is a 100 million word corpus consisting of three sub-corpora that are fairly similar in size, although each was collected using different sampling techniques and for different objectives. The first sub-corpus is the publications sub-corpus of approximately 35 million word tokens, which includes books, magazines, and newspapers. Its primary purpose was to sample all written language produced in Japan between 2001 and 2005 to create a sub-corpus suitable for cross-sectional research. The second sub-corpus is the library sub-corpus of approximately 30 million word tokens, which includes a selection of books held at certain libraries within the Tokyo area. Its purpose was to create a representative sample of widely consumed written Japanese that has cultural significance, by sampling from a selection of literary books common to several library holdings within the Greater Tokyo Metropolitan area. The third sub-corpus is the special purpose sub-corpus of approximately 35 million word tokens, which includes white papers, internet material (blog and message board postings to the Yahoo! ブログ /Yahoo! Burogu/ and Yahoo!知恵袋 /Yahoo! Chiebukuro/ systems, respectively), Diet minutes, and best-selling books. The special purpose corpus consists of a number of smaller corpora from different sources, reflecting its aim to be a heterogeneous collection to facilitate research into more specialized and unconventional aspects of written language, such as written Japanese on the Internet and transcription practices used for Diet minutes.

The computational results and data yielded by the BCCWJ project will undoubtedly become a fundamental reference resource in shaping future language policies and standards in a number of areas. For instance, the corpus will be especially valuable for developing kanji instruction programs, particularly for educational kanji, in the context of policies for compulsory Japanese language education. The BCCWJ corpus will also provide vital data for establishing various national language

policies, such as permissible kanji for family registers, jōyō kanji guidelines, and promoting literacy for ordinary Japanese citizens to fully participate in the information society. Moreover, the corpus will have major industrial implications, from influencing the kanji used in newspapers, the orthographic representation of specialized vocabularies, to JIS codes for computer and communication technologies. Thus, through its three-part composition and its various sampling methods, the BCCWJ Project sought to construct a corpus that could be beneficial for a wide range of research purposes and various aspects of language policy formulation. Table 1 presents details of the various BCCWJ sub-corpora used in the present study.

Table 1. Details of the various BCCWJ sub-corpora utilized in the present study

Sub-corpus	Sampling period	Genre	Number of tokens	
			SUW	LUW
Publications	2001–2005	Books	28,552,283	22,857,932
		Magazines	4,444,492	3,480,831
		Newspapers	1,370,233	997,535
Library	1986–2005	Books	30,377,866	25,092,641
Special purpose	2005–2007	Textbooks	928,448	746,170
	1976–2005	Bestseller books	3,742,261	3,185,745
	1976–2005	White papers	4,882,812	3,100,617
	1976–2005	Diet minutes	5,102,469	4,007,842
	1976–2005	Legislation	1,079,146	706,313
	2008	Local government pamphlets	3,755,161	2,308,450
	1980–2005	Poetry	225,273	202,425
	2004–2005	Yahoo! Chiebukuro	10,256,877	8,613,610
	2008–2009	Yahoo! Burogu	10,194,143	8,285,554
		Total	**104,911,464**	**83,585,665**

Note: Based on a similar summary table in NINJAL (2011c: 86)

4. Creation of word lists utilizing UniDic

This part of the paper briefly describes the UniDic morphological analyzer dictionary (Den et al. 2007; NINJAL 2011b) and the creation of word lists based on the BCCWJ.

As illustrated in Table 2, in addition to POS information, UniDic entries consist of three basic levels: a lemma level, a word-forms level that distinguishes

Table 2. UniDic entry consisting of 3 basic levels

語彙素 **/goiso/** **'lemma'**	語形 **/gokei/** **'word forms'**	書字形**/shojikei/** **'orthographic forms'**
矢張り (adverb) 'also; as I thought; still; in spite of; absolutely; of course'	ヤハリ /yahari/	やはり
		矢張り
	ヤッパリ /yappari/	やっぱり
		矢っ張り
	ヤッパ /yappa/	やっぱ

Note: Based on a figure in Ogura et al. (2010: 86)

Table 3. Four main word lists and 'others' group with total counts for tokens, lemma types, and orthographic types, together with LOTR and SOFR and their means across the word lists, based on both SUWs and LUWs

Word list		**Tokens**	**Lemma types**	**Orthographic types**	**LOTR**	**SOFR**
Nouns	SUW	35,450,054	99,229	141,649	0.70	0.77
	LUW	23,993,514	1,935,336	2,037,164	0.95	0.96
Verbs	SUW	14,117,717	9,311	27,540	0.34	0.39
	LUW	10,556,859	95,640	127,960	0.75	0.82
i-Adjectives	SUW	1,584,323	779	3,061	0.25	0.22
	LUW	1,331,979	11,776	15,815	0.74	0.82
Adverbs	SUW	1,825,691	3,050	7,078	0.43	0.35
	LUW	2,266,356	31,556	37,462	0.84	0.88
Others	SUW	51,724,211	63,339	98,464	0.64	0.77
	LUW	45,436,956	322,207	329,962	0.98	0.98
Totals	SUW	**104,701,996**	**175,708**	**277,792**	**0.63**	**0.74**
	LUW	**83,585,664**	**2,396,515**	**2,548,363**	**0.94**	**0.96**

Note: LOTR = lemma : orthographic type ratio and SOFR = single orthographic form ratio

pronunciations, and an orthographic-forms level that distinguishes between orthographic variants. The currently-available version of UniDic (version 2.1.0; released February 2011) lists 195,782 lemmas, for which there are 198,937 word forms and 318,690 orthographic types.

The created word lists are essentially type and token frequency counts for different word classes (part-of-speech (POS)) (with sub-corpus information). Although 13 different word lists were created, based on the POS classification within UniDic, because of space limitations, Table 3 presents the details of the four

main word classes of nouns, verbs, i-adjectives, and adverbs (with the remaining lists grouped as 'other')[15] based on both SUWs and LUWs. More specifically, each list consists of total counts for tokens, lemma types (units of language), and orthographic forms (units of writing), together with two calculated ratios: a lemma-orthographic type ratio (LOTR) and a single orthographic form ratio (SOFR), which indicates the ratio of lemma types with only one orthographic form. In total, the SUW-based word lists contain 175,708 lemma types and 277,792 orthographic representations while the LUW-based word lists contain 2,396,515 lemma types and 2,548,363 orthographic representations. As discussed in more detail below, these total figures indicate that the relationship between lemma types and orthographic representations is far from straightforward in the case of the Japanese writing system.

5. Orthographic representation: Issues and data

This part of the paper addresses a few issues related to the notion of orthographic representation within the Japanese writing system that will highlight the complex relationships that exist between units of language and units of writing. The first issue is the treatment of lemmas within UniDic, which reflects the tension between the goals of developing an efficient morphological analyzer dictionary and maintaining traditional nuances of meaning between homophonous entries. The second issue centers on the distinction between SUWs and LUWs within the BCCWJ. Essentially, the issue is about defining units of language that are appropriate for parsing a highly agglutinative language that does not employ spaces to delimit words. Drawing directly on the data of the created word lists, the third issue is to try and quantitatively capture the extent of orthographic variation within the Japanese writing system using the calculated token, lemma type, and orthographic type counts. As a number of examples taken from both the SUW and LUW word lists will demonstrate, the relationships between units of language and units of writing are commonly one-to-many.

5.1 Treatment of lemmas within UniDic

As a core tool for analyzing the BCCWJ, UniDic has undergone continuous development. For instance, Ogura et al. (2010), building on their work on UniDic 1.3.12, outline a policy change regarding the treatment of homophones for the next version of UniDic (to be distributed in XML format). Because of the high incidence of homophones in Japanese, UniDic is seeking to adopt a consistent treatment policy for the accurate tagging of the BCCWJ. However, this is a

difficult issue and the objectives of a parsing dictionary may not be consistent with traditional notions of kanji usage. The examples used in Ogura et al. can also help to illustrate the issue. UniDic version 1.3.12 distinguishes between three homophonous lemmas associated with the Native-Japanese pronunciation of あう /a.u/, as listed in 12–14.

(12) 合う /a.u/ 'fit, match, agree with, be correct'

(13) 会う /a.u/ 'meet' with 逢う as an orthographic variant

(14) 遭う /a.u/ 'encounter (by chance)' with 遇う as orthographic variant

Ogura et al. present the token counts for all inflected forms of these respective orthographic forms within the BCCWJ, as shown in (15).

(15) あう 2,057 会う 10,057 遭う 89
合う 5,737 逢う 696 遇う 913

Based on these frequencies, which indicate that Japanese writers are generally able to distinguish between the two most frequent homophones of 合う and 会う, and on further analysis of the hiragana orthographic representations to determine the intended lemma meanings, future versions of UniDic will only distinguish between two lemmas, even though the orthographic variants of 会う have different nuances of meaning, as shown in Table 4.

Table 4. Treatment of あう /a.u/ homophones in future versions of UniDic

Lemma	Orthographic forms	Meanings
合う	あう	fit, match, agree with, be correct
	合う	fit, match, agree with, be correct
会う	あう	meet, encounter
	会う	meet, encounter
	逢う	meet, encounter (date or tryst nuance)
	遭う	meet, encounter (undesirable nuance)
	遇う	meet, encounter (unexpected nuance)

A second example discussed by Ogura et al. may seem even more radical in treating all the various orthographic representations of the homophone おさまる /osamaru/ as variants of the one lemma 収まる. UniDic version 1.3.12 distinguishes between four lemmas, as listed in 16–19.

(16) 収まる /osa.maru/ 'be obtained; end'

(17) 治まる /osa.maru/ 'be at peace, be quelled'

(18) 修まる /osa.maru/ 'order (one's life); study, cultivate, master'

(19) 納まる /osa.maru/ 'be paid (in), supplied; stay (in stomach); be contented'

Similarly, Ogura et al. present token frequencies for the respective orthographic forms within the BCCWJ, as shown in (20).[16]

(20) おさまる 844 収まる 649 治まる 243 納る 16
修まる 5 収る 9 納まる 181 蔵まる 1

In contrast to the previous example, the hiragana representation of おさまる is the most frequent. This indicates that Japanese writers are often less certain about which is the correct kanji representation when writing and so commonly adopt the hiragana representation. Based on these frequencies and considering the high frequency of the hiragana representation, future versions of UniDic will regard all the orthographic forms as variants listed under the one lemma 収まる, despite the considerable semantic variation represented by the different kanji, as shown in Table 5.

Table 5. Treatment of おさまる /osa.maru/ homophones in future versions of UniDic

Lemma	Orthographic forms	Meanings
収まる	おさまる	[covering all senses]
	修まる	'order (one's life); study, cultivate, master'
	収まる	'be obtained; end'
	収る	'be obtained; end'
	治まる	'be at peace, be quelled'
	納まる	'be paid (in), be supplied; stay (in the stomach); be contented'
	納る	'be paid (in), be supplied; stay (in the stomach); be contented'
	蔵まる	(= 納まる) [noun = storehouse; warehouse]

5.2 Language units: SUWs and LUWs

The BCCWJ project employs two language-unit concepts for analytical purposes. As the Japanese writing system does not use visible word delimiters (i.e spaces between words), the exact scope of the language unit has always been rather elusive in Japanese, with no clear candidate capable of satisfying varying research objectives. Based on a number of lexical and morphological studies of the Japanese language, NINJAL elected to employ the smallest possible unit in order to capture the maximum degree of variation and which would be relevant for both

morphological and syntactic boundaries. Accordingly, the SUW was established as the primary language unit, from which a second unit, the LUW, could be constructed. The potential benefit of adopting two language units is in allowing different levels of linguistic analysis. For lexical and morphological research involving concordance searches of the corpus, the SUW provides the greatest flexibility for identifying and selecting appropriate examples, while the LUW may be a more appropriate unit for investigating some issues such as lexical comparisons within genre-related studies.

The SUW generally corresponds to the Japanese morpheme, as the smallest meaningful unit. In contrast, the LUW is closer to the 文節 /bunsetsu/ 'phrase' that consists of an independent element (noun or verb) and a dependent element (particle or inflection). Although based on the bunsetsu, the process of identifying LUWs is rather complex because it involves both top-down and bottom-up processing. In the initial top-down parsing step, a sentence is chunked into bunsetsu that are analyzed in terms of SUWs and in the subsequent bottom-up parsing step, LUWs are constructed from component SUWs (independent and dependent elements). Accordingly, LUWs include combinations formed from SUWs plus common conjugations, inflections and affixes, and combinations formed by joining noun and verb elements into compound units.[17] The correspondences between SUWs and LUWs within a phrase are illustrated in Figure 1.

LUW	日本語 /nihon-go/ Japanese language		の	表記システム /hyōki-shisutemu/ 'writing system'		に	は
SUW	日本 /nihon/ 'Japan'	語 /go/ 'language'	の /no/ POSS	表記 /hyōki/ 'writing'	システム /shisutemu/ 'system'	に /ni/ LOC	は /wa/ TOP

Figure 1. Correspondences between SUWs and LUWs within a phrase
(POSS = possessive marker, LOC = locative marker, and TOP = topic marker)

The principal disadvantage of two language units from the perspective of creating Japanese word lists is that it is not possible with SUWs alone to extract all inflected forms of verbs and i-adjectives, all words formed by derivational morphology, and, of special importance for a language where compounding is extremely productive, all compound words. Therefore, in order to create more comprehensive word lists, it was necessary to utilize both SUWs and the LUWs in extracting words from BCCWJ. As shown in Table 3 above, the total number of SUW lemma types is 175,708, while the number of LUW lemma types is 2,396,515. The considerable gap between these numbers of lemma units identifiable according to

the two different language units clearly underscores the different ways of thinking about language units that are encapsulated within these two unit concepts. The approximately thirteen-fold increase in the number of LUW lemma types over the number of SUW lemma types clearly testifies to the highly agglutinative nature of the Japanese language.

5.3 Capturing the extent of orthographic variation within the Japanese writing system

As noted earlier, the various orthographic components of the Japanese writing system are generally employed in separate ways, but it is also true that any Japanese word that has a kanji orthography representation can also be represented in both hiragana and katakana.[18] This is the source of the asymmetrical relationship between units of language and units of writing where one lemma can be represented by many orthographic variations, such as shown in Table 2 above, for example, where a single lemma is associated with five orthographic forms within UniDic. The results of the created word lists, as summarized in Table 3 above, also provide some indication of the extent of orthographic variation where within the SUW-based word lists 175,708 lemma types are represented by 277,792 orthographic forms. This part seeks to illustrate the extent of this orthographic variation in a little more detail, by using the calculated ratios presented in Table 3 and by presenting some examples of orthographic variation observed in the created word lists.

The first ratio is the LOTR, which is the ratio of lemmas to orthographic types. For example, a lemma represented by two orthographic forms would have a ratio of 0.5, such that lower LOTRs indicate more orthographic variation than higher LOTRs. The results indicate considerable levels of orthographic variation across the various word lists. For instance, in the case of SUW-based word lists, the LOTRs are 0.70, 0.34, 0.25, and 0.43 for nouns, verbs, i-adjectives, and adverbs, respectively. Thus, although the low LOTR of 0.25 for i-adjectives suggests that, on average, each of these lemmas is associated with four orthographic representations, the mean LOTR of 0.70 for nouns suggests that, on average, noun lemmas are associated with less orthographic variation. In specific contrast to the verb and i-adjective word classes, where the generally permissive attitudes towards okurigana usage is likely to be a major factor, the lower mean LOTR for the noun word list may reflect a number of factors. Possible influences could include the distributional properties of this large open class, the fact that Japanese nouns do not inflect and, apart from nominalized verbs, they do not have okurigana, as well as the characteristics of SUWs used in generating the list.

The second calculated ratio is the SOFR that indicates the ratio of lemmas within a word list that have a single orthographic form. Supplementing the LOTR, the SOFR is useful in better understanding the distributions of LOTRs for the open word classes, such as the noun class that has an extremely long distributional tail of low frequency lemmas and which, due to their low occurrences within the BCCWJ, invariably only have one orthographic form. The SOFR for the SUW-based noun list is 0.77, which is consistent with the suggestion that the majority of lemmas in the long tail have only one orthographic representation. This high SOFR also reflects the fact that noun lemmas account for approximately 56% of all lemma types. In marked contrast, the SOFRs for SUW verbs, i-adjectives, and adverbs are far lower at 0.39, 0.22, and 0.35, respectively. The SOFRs for the LUW-based word lists reveal considerably longer distributional tails compared to the ratios for the SUW forms. This probably reflects both the nature of LUWs as compounded SUWs, where some SUW lemmas are completely subsumed within LUWs that consist of more than one SUW, and the size of the corpus, where, because the LUWs are on average longer than the SUWs, an even larger corpus is probably necessary to fully uncover the variation within LUWs.

In order to obtain a clearer sense of the orthographic variation, the lemmas within each of the four main word-class lists for both SUWs and LUWs were ranked by frequency and the resultant distributions are plotted in Figures 2 and 3,

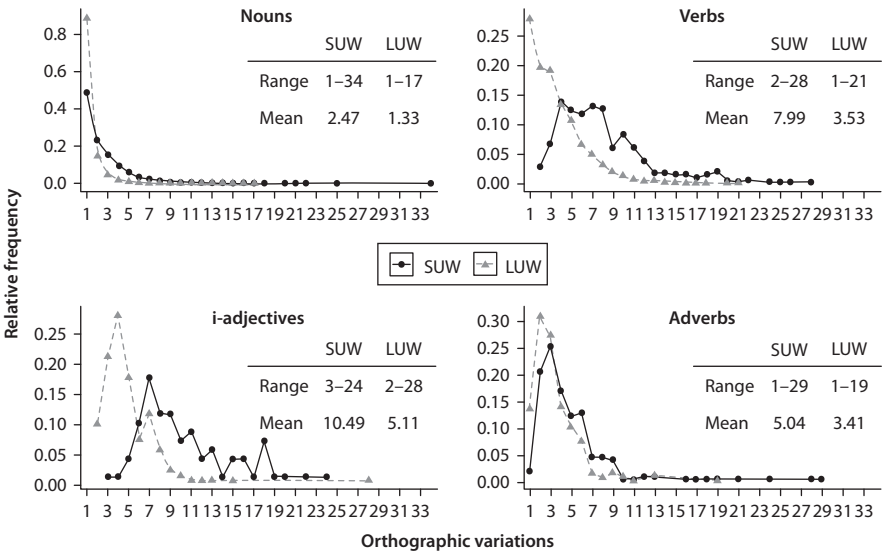

Figure 2. Distributions of orthographic variation as a function of number of variants (x-axis) and relative frequency (y-axis) among the most frequent 90% of lemmas in terms of token counts for each word list, together with ranges and means of orthographic variation

together with the respective ranges and means of orthographic variation. Figure 2 illustrates the orthographic variation among the lemmas that account for 90% of all occurrences in each word list; this level of coverage is adopted to strike a reasonable balance between minimizing the effects of many low frequency lemmas with no variation and gaining a realistic sense of the overall patterns of variation for each word class. In contrast, Figure 3 presents more focused comparisons, by illustrating the orthographic variation among the most frequent 100 lemmas within each word class list.[19]

Because the distributions in Figure 2 represent the orthographic variation within the lemmas that account for a cumulative frequency of 90% of each word class total, the scales vary considerably in terms of the numbers of lemmas with different levels of orthographic variation. Focusing initially on the SUW-based word lists, although nouns have by far the greatest number of lemmas with only one orthographic representation (3,749) and have the lowest variation mean (2.47), it should also be noted that there are more noun lemmas with seven variations (163), for instance, compared to the other word classes combined (73), and nouns have the widest range of orthographic variation (1–34) compared to the other word classes. In contrast, the highest means for orthographic variation within the SUWs are observed for the verbs and i-adjectives, which are roughly three to four times higher than the mean for nouns (2.47 compared to 7.99 and 10.49, respectively). It is also instructive to look at the modes in Figure 2; compared to the modes of 4 for verbs and 7 for i-adjectives, the nouns and adverbs have modes of 1 and 3, respectively. Turning to the LUW lemmas, the distributions for nouns, i-adjectives, and adverbs are quite similar to those for SUW lemmas in their general shapes, although reflecting their higher LOTRs, the distributions for LUWs are naturally more skewed to the right. In addition to stronger right skewing, while the LUW i-adjectives have the greatest range of orthographic variation (2–28), the mean is only 5.11. The long distributional tail for LUW verbs is likely to result from the inclusion of many LUW verbs consisting of a SUW noun plus the dummy 'do' verb, with the overall distribution more closely mirroring the distributional properties of SUW nouns.

Concerning the graphs in Figure 3, it is noteworthy that the ranges for orthographic variation are essentially unchanged for the SUWs and only slightly narrower for the LUWs. Thus, the overall impression is of very similar distributions of orthographic variation across the four word classes for both SUWs and LUWs. Moreover, comparing the upper range limits in Figure 2 to those in Figure 3, it is clear that a great deal of orthographic variation occurs within the extremely small sets of the 100 most frequent lemmas in these main word classes. Overall, these distributions indicate that orthographic variation is a major characteristic of the

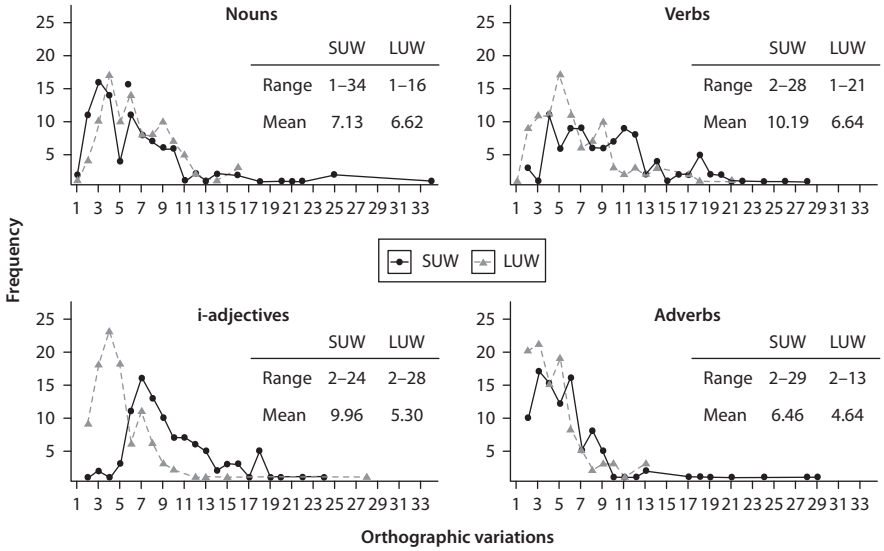

Figure 3. Distributions of orthographic variation as a function of number of variants (x-axis) and frequency (y-axis) among the most frequent 100 lemmas for each word list, together with ranges and means of orthographic variation

Japanese writing system, at least in the orthographic representation of the most common lemmas of the Japanese language.

Tables 6 and 7 present some orthographic variation examples for SUW and LUW lemmas, respectively. For the sake of brevity, however, only a selection is singled out for more general comment. For instance, much of the variation within the 9 orthographic variants of 聞き取る /ki.ki.to.ru/ 'catch (words)' relates to kanji selection (either 聞 as the standard kanji for 'to listen' or 聴 with nuance of 'to hear; to listen') and whether the /to/ of /to.ru/ (with sense of 'to take; catch') should be represented in kanji or hiragana. In contrast, the 11 variants for 面白い /omoshiro.i/ 'interesting; amusing' appear to reflect more stylistic factors; particularly, the non-standard usage of katakana in both the 面白 stem and the okurigana elements and the non-standard Japanese dialect form オモロイ /omoro.i/. Moreover, while the kanji orthography representation of 全然 (4,869) is approximately five times more frequent than the hiragana orthography representation ぜんぜん (906), many of the other variations are attempting to more accurately represent the pronunciation in a given situation, such as the lengthened initial /e/ vowel in ぜーんぜん /zēnzen/ for emphasis or mock imitation of a child-like pronunciation in じぇんじぇん /jenjen/.

Table 6. Examples of orthographic variation in SUWs across the four word classes of nouns, verbs, adjectives, and adverbs

	Tokens	Ratio		Tokens	Ratio
Noun			**Verb**		
玉葱 /tamanegi/ 'onions'			聞き取る /ki.ki.to.ru/ 'catch (words)'		
玉ねぎ	1,026	0.49	聞き取る	397	0.68
タマネギ	446	0.21	聞きとる	91	0.16
たまねぎ	345	0.17	聴き取る	52	0.09
玉葱	176	0.08	ききとる	18	0.03
玉ネギ	94	0.05	聴きとる	17	0.03
	2,087		きき取る	3	0.01
			聴取る	2	0.00
			尋きとる	1	0.00
			聞取る	1	0.00
				571	
Adjective			**Adverb**		
面白い /omoshiro.i/ 'interesting; amusing'			全然 /zenzen/ 'not at all (-); entirely (+)'		
面白い	8,006	0.67	全然	4,869	0.83
おもしろい	3,770	0.31	ぜんぜん	906	0.15
おもろい	137	0.01	ゼンゼン	34	0.01
オモシロイ	38	0.00	ぜーんぜん	27	0.00
オモシロい	32	0.00	ぜ〜んぜん	9	0.00
オモロい	16	0.00	ぜんっぜん	7	0.00
オモロイ	13	0.00	じぇんじぇん	3	0.00
おもしろーい	7	0.00	全々	3	0.00
おもしろし	5	0.00	全っ然	2	0.00
面白し	3	0.00	ぜっんぜん	2	0.00
面白ーい	2	0.00	ぜんぜ〜ん	1	0.00
	12,029		じぇんっじぇん	1	0.00
			ぜん〜ぜん	1	0.00
				5,865	

Although not listed in Table 7, in the interest of brevity, the 28 variants of the i-adjective 格好良い /kakkō-ī/ 'cool; smart-looking' observed in the LUW list are also highly illustrative of the orthographic variation phenomenon. The variations

result from various combinations of kanji (格好 and nonstandard 恰好) or kana (both hiragana and katakana) representations for the first SUW of /kakkō/ (as well as hiragana and katakana representations for its phonological variants of /kakko/ and /katchō/) together with kanji or kana representations of the second SUW of /ī/ (and its phonological variant as /yoi/). Although this particular example is admittedly rather extreme in having so many attested orthographic variations, given their compositional nature, LUWs will generally have more orthographic variation than their constituent SUWs (such as, in the case of 格好良い, 28 variations compared to the 8 and 22 orthographic variations for 格好 and 良い, respectively).

Table 7. Examples of orthographic variation in LUWs across the four word classes of nouns, verbs, adjectives, and adverbs

	Tokens	**Ratio**		**Tokens**	**Ratio**
Noun			**Verb**		
洗濯|物 /sentaku-mono/ 'laundry'			御|願い|する /o-negai-suru/ 'to ask (politely)'		
洗濯|物	622	0.90	お|願い|する	11,886	0.96
せんたく|もの	22	0.03	お|ねがい|する	405	0.03
洗濯|もの	22	0.03	御|願い|する	70	0.01
洗たく|物	15	0.02	お|願|する	27	0.00
洗濯|モノ	4	0.01	お|願ひ|する	10	0.00
せんたく|物	2	0.00		**12,398**	
洗濯|もん	1	0.00			
	688				
Adjective			**Adverb**		
申し訳|無い /mōshiwakenai/ 'no excuse'			必ず|しも /kanarazu-shimo/ 'necessarily (-)'		
申し訳|ない	1,157	0.82	必ず|しも	3,372	0.92
申しわけ|ない	217	0.15	かならず|しも	266	0.07
申訳|ない	16	0.01	必らず|しも	3	0.00
申し訳|無い	10	0.01	必|しも	3	0.00
もうしわけ|ない	2	0.00		**3,649**	
申しわけ|無い	1	0.00			
申訳|無い	1	0.00			
	1,404				

Note: In these LUW examples, SUW boundaries are marked by |

6. Conclusion

This paper has sought to highlight a number of issues relevant to orthographic representation within the Japanese writing system, which beneficially illuminate the complex nature of the relationships that can exist between units of language and units of writing. As briefly documented in Joyce (2011), the Japanese writing system has frequently been described by writing systems researchers as one of the most complicated systems ever developed, and one factor that has undoubtedly motivated some of those portrayals is the multi-scriptal nature of the Japanese writing system. The principle mixture of morphographic kanji and the two syllabographic sets of hiragana and katakana, supplemented with alphabetic rōmaji and symbols, means that within the Japanese writing system the fundamental relationship between units of language and units of writing is a one-to-many relationship, which, on the one hand, entails a higher degree of complexity, but, on the other hand, offers a great level of orthographic flexibility.

Indicative that the relationship between lemma types and orthographic representations is not straightforward in the case of the Japanese writing system, a summary of the generated word lists showed that there are 277,792 orthographic representations for the 175,708 lemma types in the SUW word lists and that the mean LOTR for SUWs is 0.63, which indicates that, on average, SUW lemmas have approximately two orthographic representations.

Although presented primarily from the perspective of creating the word lists, the issues discussed in Part 5 have wider and more profound implications for thinking about the complex relationships between units of language and units of writing within the context of the Japanese writing system. Given the intricate nature of these issues, that touch on sociolinguistic aspects of writing, such as language policy, their full discussion is beyond the scope of this paper. Instead, this paper has sought to draw attention to the problems as warranting further examination and to highlight the significant contribution that the BCCWJ can make for such investigations.

At one level, the first issue of what to treat as lemmas within UniDic is about developing an efficient morphological parsing dictionary by establishing a consistent policy for the accurate tagging of the BCCWJ, but, at another level, the issue highlights the difficulties of meeting the diverse research objectives of various researchers, such as lexicographers, corpus linguists, educators and policy makers, as well as Japanese language instructors. The issue is also closely connected with the multi-scriptal nature of the Japanese writing system, because, as illustrated in Table 2, UniDic entries need to specify the range of orthographic forms associated with a lemma, which will often include both kanji orthography

and kana orthography representations. However, as noted in Part 5.1, UniDic's lemma treatment policy is at the expense of maintaining traditional distinctions between different kanji representations of homophone words that sometimes signal nuances between senses.

The second issue is about defining units of language that are appropriate for parsing a highly agglutinative language, such as Japanese, that does not employ spaces to delimit words, and the distinction between SUWs and LUWs within the BCCWJ. While the recognition of two linguistic units may facilitate different levels of linguistic analysis, it seems fair to remark that neither linguistic unit corresponds very well to a general notion of the word; SUWs often necessitate a deeper awareness of morphological structures while LUWs tend to emphasize phrasal structures. As the summary of the word lists in Table 3 indicates, it is clearly necessary to employ both linguistic units in order to extract a wider range of word types from a Japanese language corpus, such as the BCCWJ.

Given the complex relationships between units of language and units of writing, due to the multi-scriptal nature of the Japanese writing system, problems with identifying separate lemmas, and the distinction between SUWs and LUWs, the third issue addressed in Part 5 was an attempt to capture the extent of orthographic variation within the Japanese writing system. Although simple comparisons of the lemma type counts to orthographic forms counts in Table 3 indicate that, on average, there are approximately two orthographic representations for every SUW-based lemma in the BCCWJ, this study also calculated LOTRs and plotted their distributions in order to more accurately quantify the extent of orthographic variation. These distributions of orthographic variation reveal considerable levels of orthographic variation, with up to 34 orthographic variants for one SUW noun lemma and mean variations of 8.44 for SUWs and 5.80 for LUWs (based on the most frequent 100 lemmas). Moreover, from comparisons of the graphs presented in Figures 2 and 3, it is clear that the SUW and LUW lemmas with high degrees of orthographic variation are predominately within the most frequent 100 lemmas for each of the four word classes focused on in this paper.

Admittedly, the present paper has not touched on the complex sociolinguistic factors that might influence the selection of one orthographic representation over another when writing in Japanese, such as the association of kanji orthography with education and erudition or the popularity of katakana and rōmaji in the mass media (see Tranter 2008 for some discussion of nonconventional script choices in terms of language function). It has, however, presented quantitative data from word lists generated from the recently-compiled BCCWJ – particularly, the highly-similar distributions of orthographic variation across the four main word

classes of nouns, verbs, i-adjectives, and adverbs – that clearly demonstrate that orthographic variation is a major characteristic of the Japanese writing system, at least, in terms of the orthographic representation of the most common words of the Japanese language.

Notes

1. http://www.tokuteicorpus.jp

2. It is also widely recognized that Korea played an important role in the transmission process (Miller 1967; Seeley 1991).

3. Indeed, Morohashi's (1955–1960) 大漢和辞典 /daikanwajiten/ 'Comprehensive Chinese-Japanese dictionary' has entries for 49,963 kanji; although it is not a reliable index of usage due to the inclusion of various graphic versions of the same kanji as well as highly obscure kanji used only a few times within a rare text. Still, it has been estimated that the number of characters used for writing Japanese was more than 10,000 at the end of the Tokugawa period (1603–1868) (Twine 1991).

4. Shibatani (1990) observes that the preface to the 古事記 /kojiki/ 'Record of Ancient Matters', the earliest extant text (712 CE), was written like this.

5. Actually, the situation is somewhat more complex, because some kanji have multiple on'yomi associated with different meanings, reflecting the fact that some Chinese characters were borrowed repeatedly at different periods and from different regions of Chinese (Miller 1967; Shibatani 1990).

6. When providing the pronunciation of a Japanese word, this paper follows the convention of using a period to indicate the boundary between different scripts in the case of multi-script orthographic representations. In a couple of subsequent examples, ・ (a Japanese middle-dot) is used to mark word boundaries.

7. There is also a further distinction of 外来語 /gairaigo/ 'borrowed words', which refers to words borrowed from languages other than Chinese and not written with Chinese characters. It may also be noted that the differentiation of certain words as either Sino-Japanese or Native-Japanese are not always etymologically accurate.

8. Katakana was developed by Buddhist priests as part of the kundoku reading conventions for annotating texts written in Chinese, but because of limited space between lines and the need to write quickly while taking notes, the priests would write man'yōgana in a simplified form, usually by isolating a distinctive feature of the character. Hiragana also developed from man'yōgana, but the process was somewhat different. From the early Heian period (794–1185), man'yōgana were written in a cursive hand called 草仮名 /sōgana/ 'grass style', and hiragana represent a simplified and more cursive form of sōgana (Habein 1984).

9. Japanese adjectives are traditionally divided into two categories: i-adjectives that end in *-i* and inflect and na-adjectives formed from nouns by adding な /na/.

10. Although some transcription conventions include a macron to indicate the lengthening of some long vowel sounds (Smith 1996), as one of the anonymous reviewers suggests, transcriptions aimed at Japanese readers tend not to use them.

11. As one anonymous reviewer points out, the kana scripts can also be used as 振り仮名 /furigana/ referring to the convention of indicating the pronunciation of a kanji orthography word or word element in a kana script usually placed next to or above the word. The convention is mainly used when it is expected that a reader may not know the appropriate pronunciation, and so it is most commonly employed in reading materials for children. However, this reading-aid convention is not regarded as a form of orthographic variation within this paper.

12. The BCCWJ project is actually part of the NINJAL's larger Kotonoha Project (kotonoha is a term of classical Japanese meaning 'word of language') that also includes the Corpus of Spontaneous Japanese (CSJ: a corpus of speeches) and the Taiyō Corpus (a corpus of the Taiyō periodical published between 1895–1925).

13. http://www.natcorp.ox.ac.uk.

14. The third author was the leader of the 作文支援システム班 /sakubun shien shisutemu han/ 'Composition Support Group', of which the first author was also a research member.

15. 'Others' includes proper nouns, na-adjectives, affixes, interjections, particles, auxiliary verbs, prenominals, and conjunctions. The data in Table 3 is also adjusted by excluding the spacing and symbol lists that include punctuation and so have extremely high token counts.

16. In addition to distinctions between the four lemmas, the variations in orthographic word forms in this example also reflect different okurigana practices for representing the inflectional elements.

17. For example, one particularly common kind of LUW is created by combining SUW Sino-Japanese nouns with the SUW dummy verb 'to do', as in the example given earlier of 読書 '(book) reading' and する /suru/.

18. One anonymous reviewer suggested that one might expect minimal levels of orthographic variation for Sino-Japanese words, because they are generally perceived to be Chinese in origin and because Chinese does not allow for orthographic variation. While we hesitate to make claims about the orthographic practices of one language based on the orthographic conventions of another, evidence from the BCCWJ seems to clearly refute the suggestion. While it is true that, in comparison to the mean LOFR of 0.52 for Native-Japanese nouns, the mean ratio of 0.84 for Sino-Japanese nouns is much higher, still it does indicate the presence of orthographic variation among Sino-Japanese nouns. Moreover, while lexical stratum comparisons are less meaningful for verbs and i-adjectives, the LOFRs for Sino-Japanese and Native-Japanese adverbs are highly similar, at 0.41 and 0.43, respectively. Taking the Sino-Japanese adverb of 多分 /tabun/ 'perhaps' as an example, 44% of its occurrences within the BCCWJ are written in kanji orthography and the other 56% are written in hiragana, which testifies to a considerable degree of latitude concerning its orthographic representation.

19. It should be noted that in the case of i-adjectives, fewer lemmas are included within the Figure 2 plots than in the Figure 3 plots, because 75 types alone account for 90% of the occurrences (tokens).

References

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

DeFrancis, John (1989). *Visible speech: The diverse oneness of writing systems*. Honolulu, HI: University of Hawaii Press.

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

Gottlieb, Nanette (2008). Japan: language policy and planning in transition. In Robert B. Kaplan & Richard B. Baldauf Jr. (eds.), *Language planning and policy in Asia, Vol. 1: Japan, Nepal, Taiwan and Chinese characters* (Language planning and policy), 102–169. Bristol / Buffalo/ Toronto: Multilingual Matters.

Habein, Yaeko Sato (1984). *The history of the Japanese written language*. Tokyo: University of Tokyo Press.

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 & Susanne R. Borgwaldt (2011). Typology of writing systems: Special issue introduction. In Susanne R. Borgwaldt & Terry Joyce (eds.), *Typology of writing systems.* Special issue of *Written Language and Literacy* 14(1): 1–11.

Kess, Joseph F. & Tadao Miyamoto (1999). *The Japanese mental lexicon: Psycholinguistic studies of kana and kanji processing*. Philadelphia/Amsterdam: John Benjamins.

Lunde, Ken R. (1993). *Understanding Japanese information processing*. Sebastopol, CA: O'Reilly & Associates.

Maekawa, Kikuo (2007). Design of a balanced corpus of contemporary written Japanese, *Proceedings of the Symposium on Large-Scale Knowledge Resources (LKR2007)* (1–3 March, Tokyo Institute of Technology, Tokyo, Japan), 55–58.

Martin, Samuel E. (1972). Nonalphabetic writing systems: Some observations. In James F. Kavanagh & Ignatius G. Mattingly (eds.), *Language by ear and by eye: The relationship between speech and reading*, 81–102. Cambridge, MA: MIT Press.

Miller, Roy Andrew (1967). *The Japanese language*. Chicago: University of Chicago Press.

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

National Institute for Japanese Language and Literature (NINJAL) [Kokuritsu Kokugo Kenkyūjo] (2011a). *Tokuteiryōiki kenkyū nihongo kōpasu kenkyū seika hōkoku* [Priority-area research 'Japanese corpus': Research report] [DVD format of data and research reports]. Tokyo: General Headquarters, Priority-Area Research 'Japanese Corpus'.

National Institute for Japanese Language and Literature (NINJAL) [Kokuritsu Kokugo Kenkyūjo] (2011b). *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.

National Institute for Japanese Language and Literature (NINJAL) [Kokuritsu Kokugo Kenkyūjo] (2011c). *Gendai nihongo kakikotoba kinkō kōpasu: Riyō no tebiki*. Dai 1.0 han

[BCCWJ user's manual: version 1.0]. Tokyo: Center for Corpus Development, National Institute for Japanese Language and Linguistics.
Ogura, Hideki, Toshinobu Ogiso, Hanae Koiso, Yutaka Hara & Sayaka Miyauchi (2010). Keitaiso kaiseki jisho UniDic ni okeru goiso midashi no rikkō hōshin [Criteria for the lemmatization of UniDic], *Tokuteiryōiki kenkyū 'nihongo kōpasu' heisei 21 nendo kōkai waakushoppu (Kenkyū seika hōkokukai) yokōshū* [Priority-area research 'Japanese corpus': proceedings of the 2010 public workshop]. Tokyo: General Headquarters, Priority-Area Research "Japanese Corpus".
Seeley, Christopher (1991). *A history of writing in Japan*. Leiden: E.J. Brill.
Shibatani, Masayoshi (1990). *The languages of Japan*. Cambridge, England: Cambridge University Press.
Smith, Janet Shibamoto (1996). Japanese writing. In Peter T. Daniels & William Bright (eds.), *The world's writing systems*, 209–217. New York: Oxford University Press.
Taylor, Insup & M. Martin Taylor (1995). *Writing and literacy in Chinese, Korean and Japanese*. Amsterdam: John Benjamins.
Tranter, Nicolas (2008). Nonconventional script choice in Japan. *International Journal of the Sociology of Language* 192: 133–151.
Twine, Nanette (1991). *Language and the modern state: The reform of written Japanese*. London: Routledge.

Author's addresses:

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

terry@tama.ac.jp

Bor Hodošček
Tokyo Institute of Technology
Department of Human System Science
W9-622, 2-12-1 O-okayama, Meguro-ku
Tokyo, 152-8552
Japan

hodoscek.b.aa@m.titech.ac.jp

Kikuko Nishina
Tokyo Institute of Technology
3-2-3-302, Shin-machi, Setagaya-ku
Tokyo, 154-0014
Japan

knishina@ryu.titech.ac.jp