Constructing Databases of Japanese Three- and Four-Kanji Compound Words

Some Observations Concerning Their Morphological Structures

Terry Joyce · Hisashi Masuda

Abstract. As the principal component of the multiple script Japanese writing system (JWS), morphographic kanji function as the core units of graphematic representation for a considerable proportion of the Japanese lexicon (Joyce and Masuda, 2018; 2019; Joyce, Masuda, and Ogawa, 2014; Kobayashi, Yamashita, and Kageyama, 2016; Nomura, 1975; 1988). Deeply entwined with the morphographic nature of Japanese kanji (Joyce, 2011), the Japanese language offers especially fascinating opportunities for both linguistic and psycholinguistic investigations of compound words (Joyce, 2002; 2004; Masuda and Joyce, 2018). As contributions to the ongoing construction of a larger database project of Japanese lexical properties (Joyce, Hodoščeк, and Masuda, 2017; Joyce, Masuda, and Ogawa, 2014), which aims to facilitate such investigations in terms of experimental designs and stimuli preparation, this paper reports on two new database components for three-kanji (3KCWs) and four-kanji compound words (4KCWs) respectively. More specifically, the paper focuses on the results of analyzing their morphological structures. In contrast to 3KCWs, where the dominant morphological structure is attaching suffixes to existing two-kanji compound words (2KCWs), such as 可能性 /ka-nō-sei/ *potentiality; possibility* [[can + ability = possible; potential] + nature; *-ity* ending], for 4KCWs, the dominant structure is compounding with two 2KCWs combined, such as 自分自身 /ji-bun-ji-shin/ *oneself* [[oneself + one's lot = oneself] + [oneself + someone = oneself]].

1. Introduction

One of the most fundamental characteristics of contemporary written Japanese is its simultaneous employment of multiple scripts, which is

Terry Joyce 0000-0001-9625-1979
School of Global Studies, Tama University, 802 Engyo, Fujisawa, Kanagawa, 252-0805, Japan
E-mail: terry@tama.ac.jp

Hisashi Masuda 0000-0001-8619-6275
Faculty of Health Sciences, Hiroshima Shudo University, 1-1-1 Ozukahigashi, Asaminami-ku, Hiroshima, 731-3195, Japan
E-mail: hmasuda@shudo-u.ac.jp

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2020. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 5.
Fluxus Editions, Brest, 2021, pp. 579–619. https://doi.org/10.36824/2020-graf-joyc
ISBN: 978-2-9570549-7-8, e-ISBN: 978-2-9570549-9-2

referred to as 漢字仮名交じり文 /kan-ji-ka-na.ma.jiri.bun/[1] *mixed kanji and kana writing* [kanji + kana + mixed + writing] in Japanese (for fuller accounts of the Japanese writing system (JWS), see Joyce and Masuda, 2018; 2019, as well as Kess and Miyamoto, 1999; Konno, 2013; Smith, 1996; Smith and Schmidt, 1996; Taylor and Taylor, 2014). The four component scripts are morphographic 漢字 /kan-ji/ *kanji* [Han + character], the two separate syllabographic 仮名 /ka-na/ *kana* [provisional + name] scripts of 平仮名 /hira-ga-na/ *hiragana* [smooth + provisional + name] and 片仮名 /kata-ka-na/ *katakana*, [part + provisional + name] and the phonemic alphabet of ローマ字 /rōma.ji/ *Roman letters* [Roman + character], which are supplemented by the small set of Arabic numerals 数字 /sū-ji/ *numbers* [number + character] (Joyce and Masuda, 2018; 2019). Undoubtedly, this unique aspect of the JWS contributes greatly to the highly fungible nature of Japanese written representations (Backhouse, 1984; Joyce, Hodošček, and Nishina, 2012; Joyce and Masuda, 2018; 2019; Miller, 2011; Robertson, 2015; 2017; Smith, 1996; Tranter, 2008). Although Joyce and Masuda (2019) have recently advocated an inclusive notion of intentionality as a promising approach to capturing the diverse motivational factors that influence Japanese graphematic representations, as they equally emphasize, instances of graphematic variation can only be appropriately interpreted with reference to Japanese orthographic conventions. Moreover, as such conventions are closely tied to the historical development of the JWS—from the initial adaption of Chinese characters, the early emergence of the kana scripts, and the relatively recent supplement with rōmaji (Joyce and Masuda, 2018; Lurie, 2012)—there are particularly strong affinities between the scripts and the different lexical strata of the Japanese language (Joyce and Masuda, 2018; 2019; Kageyama and Saito, 2016).

Citing Tamamura (1984) in illustration, Kageyama and Saito (2016) claim that studies of the Japanese language have traditionally distinguished between four 語種 /go-shu/ *word types* [word + type], or lexical strata. They are (1) indigenous 和語 /wa-go/ *Native-Japanese words* (NJ) [Japan + word], (2) 漢語 /kan-go/ *Sino-Japanese words* (SJ) [Han + word], entering from Chinese, (3) 外来語 /gai-rai-go/ *Foreign-Japanese words* (FJ) [outside + come + word], entering from foreign languages since the 16th century, and (4) 混成語 /kon-sei-go/ *hybrid words* [mix + create + word].[2]

[1] Unless redundant by context, such as within Table listings, Japanese words are represented conventionally and are usually followed by a phonological gloss between slash symbols, / /, English translation in italics, and morpheme meanings and their concatenation within square brackets [ ]. Within the phonological glosses, word boundaries are indicated by spaces, kanji-kanji boundaries by hyphens, and other script boundaries by periods, with macrons, such as ō, indicating long vowels.

[2] It should, however, be noted that more recent classifications also cover four types, but the categories differ. Although Shibatani (1990) and Kageyama and Saito

The close affinities between the component scripts and the different lexical strata are manifest in a set of general tendencies.[3] Broadly, these are for kanji to represent both SJ and NJ content words as well as NJ verb and adjective stems, for hiragana to represent NJ functional elements such as grammatical markers and inflections, for katakana to represent both FJ and mimetic words, and for rōmaji to represent FJ words and names (Joyce and Masuda, 2019; Kageyama and Saito, 2016).

TABLE 1. Affinities between Japanese lexical strata and JWS component scripts

Stratum	Script	Examples
NJ	Kanji	山 /yama/ *mountain*; 筆 /fude/ *calligraphy brush*
	Kanji-Hiragana	高い /taka.i/ *tall*; 書く /ka.ku/ *to write*
	Hiragana	これ /kore/ *this*; の /no/ *possessive marker*
	Katakana	ワンワン /wanwan/ *doggy*; チカチカ /chikachika/ *flickering, twinkling*
SJ	Kanji	愛 /ai/ *love*; 大学 /dai-gaku/ *university* [big + study]; 正書法 /sei-sho-hō/ *orthography* [correct + write + way]
FJ	Katakana	ミルク /miruku/ *milk*; クラス /kurasu/ *class*; スマートフォン /sumātofon/ *smart phone*
	Rōmaji	PC /pīshī/ *personal computer*; CM /shīemu/ *TV commercial*
Hybrid	Kanji-Kanji	表玄関 /omote-gen-kan/ *front entrance* [NJ+SJ]
	Kanji-Katakana	野菜ジュース /ya-sai.jūsu/ *vegetable juice* [SJ+FJ]
	Hiragana-Katakana	あんパン /an.pan/ *bean-jam bun* [NJ+FJ]

Notes: NJ = native-Japanese; SJ = Sino-Japanese; FJ = foreign-Japanese

(2016) continue to recognize the same first three categories (i.e., NJ, SJ and FJ), their fourth category is *mimetic words* that "express non-linguistic sounds or cries or vividly express states or action or physical sensations" (Kageyama and Saito, 2016, p. 12). Usually, they are referred to as 擬音語・擬声語・擬態語 /gi-on-go・gi-sei-go・gi-tai-go/ in Japanese.

[3] As one source of deviation from these tendencies, Kageyama and Saito (2016) note that, because the different scripts have distinct perceptual characteristics, such as the stiff and formal impressions of kanji, writers may employ graphematic variants to convey certain nuances. However, as Joyce and Masuda (2019) describe in some detail, there is a wider range of intentionality factors underlying Japanese graphematic variation. Accordingly, they treat such script associations as one subcategory of script sensibilities, which is one of their three main factor categories, together with message context and creative representations.

Table 1 presents some examples of these script-lexical strata affinities. Structurally, Table 1 is closely based on Kageyama and Saito's (2016, p. 13) Table 1, entitled 'Classification of word types in traditional Japanese grammar' (as adapted, in turn, from Tamamura, 1984, p. 110), which is primarily from the perspective of the lexical strata. It has, however, been supplemented with a few additional examples from Joyce and Masuda (2019, p. 253) Table 1, entitled 'Examples of standard JWS orthographic conventions', which underscores the same script-strata associations, albeit primarily from the perspective of the JWS's component scripts. While granting that the range of examples in Table 1 may potentially obscure matters, a couple of deeply intertwined points, which are particularly germane to this paper, warrant highlighting. The first point is that, although exact script proportions vary across different genres of written Japanese, kanji are unquestionably the principal component script of the JWS. Indeed, in an interesting study of average script proportions, Igarashi (2007) reports that kanji represented approximately 72%, hiragana 18%, katakana 6% and alphabetic symbols and numbers 4% of the word lists that she extracted from three major newspapers, which, in targeting general adult readerships, closely conform to standard Japanese orthographic conventions.

The second significant point is that, because kanji have deep affinities with the two dominant Japanese lexical strata, both NJ and SJ words, they function as the core units of graphematic representation for a considerable proportion of the Japanese lexicon. Admittedly, this may superficially appear to be merely stating the reason why kanji are the dominant component script within the JWS, but the pluralistic links between kanji and both the NJ and SJ lexical strata are key to understanding the complex nature of Japanese morphographic kanji (Joyce, 2011; Kobayashi, Yamashita, and Kageyama, 2016). Although Kobayashi, Yamashita, and Kageyama (2016, p. 93) tender their remark with specific reference to SJ words, it is essentially impossible to discuss the graphematic representation of the Japanese lexicon as a whole "without some explanation of the *kanji* themselves" (italics in original). It is, therefore, expedient at this point to briefly draw on their succinct account and examples of how kanji became associated with both SJ and NJ words. By their definition, SJ words have entered the Japanese language due to lexical borrowing from the Chinese language; a process that essentially dates back to around the third and fourth centuries to when Chinese characters were initially borrowed and subsequently adapted for written Japanese. Consistent with their morphographic nature in Chinese, kanji represent either a single word or a morpheme, such as 木 meaning *tree*. Also reflecting different historical Chinese pronunciations, this particular kanji is associated with two SJ morphemes or, from the perspective of their phonological values, the two 音読み /on-yo.mi/ *SJ readings* [sound + reading] of /moku/ and /boku/, in different SJ compound words, as in (1).

(1) /moku/ 木馬 /moku-ba/ *wooden horse* [wood + horse]
材木 /zai-moku/ *timber* [material + tree]
/boku/ 木刀 /boku-tō/ *wooden sword* [tree + sword]
巨木 /kyo-boku/ *large tree* [giant + tree]

Moreover, as the Japanese language already had NJ words for many of the SJ morphemes represented by kanji, such as the NJ /ki/ for *tree*, it does not require a great leap of imagination to understand how kanji also came to be associated with those NJ morphemes and their phonological values; 訓読み /kun-yo.mi/ *NJ readings* [semantic + reading].[4] As Kobayashi, Yamashita, and Kageyama (ibid.) stress, although some SJ words are monomorphemic and, thus, graphematically represented by a single kanji, such as 茶 /cha/ *tea* and 損 /son/ *loss*, most SJ morphemes are bound morphemes in nature, such that they combine with other SJ morphemes to form compound words.

The Japanese language is particularly interesting from the perspectives of word formation processes and its morphological structures (Kageyama and Saito, 2016; Shibatani, 1990; Tamamura, 1984; 1985). However, as Kageyama and Saito (2016) observe, the application of various word formation processes varies markedly across the different lexical strata. Consistently, although compounding, which Shibatani (1990, p. 237) singles out as being the most productive process by far, is attested with both NJ and SJ elements, it is particularly prominent for SJ words (Kageyama and Saito, 2016; Kobayashi, Yamashita, and Kageyama, 2016). Some sense of the striking differences can be discerned from Joyce, Masuda and Ogawa's (2014) analyses of the graphematic representation codes that they applied to the headwords of the sixth edition of the 広辞苑 /kō-ji-en/ *Kōjien* dictionary (Shinmura, 2008). For example, with C standing for kanji, H for hiragana and K for katakana, 山 was coded as C, 高い as CH, 大学 as 2C, and 山登り /yama-nobo.ri/ *mountain climbing* [mountain + climb] as 2CH. Table 2 shows the ten most frequent graphematic representations codes for the list of Kōjien headwords.[5]

What is particularly striking about these results is that the first three graphematic representation codes of 2C, 3C and 4C (i.e., 2KCWs,

[4] The official list of characters for general use, known as the 常用漢字表 /jō-yō-kan-ji-hyō/ *Jōyō kanji list* (Agency for Cultural Affairs, 2010), also includes /ko/ as an NJ morpheme in some NJ compound words, such 木陰 /ko-kage/ *shade of tree* [tree + shade]. Kobayashi, Yamashita, and Kageyama (2016, p. 93) refer to it as an "allomorph (apophonic variant)" and acknowledge that "the same character 木 is used in such cases as well".

[5] The sixth edition of Kōjien has 232,795 headword entries, but the analyzed list consisted of 215,597 headwords after excluding all kanji that are not on the official jōyō or the Japanese Industrial Standard (JIS) level 1 lists. Of the 1,152 separate graphematic representations codes applied to the list, 578 (50.2%) were unique (i.e., frequency = 1).

Table 2. Ten most frequent graphematic representation codes observed for a list of Kōjien (Shinmura, 2008) headwords (based on Joyce, Masuda, and Ogawa, 2014, p. 188)

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

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

3KCWs and 4KCWs, respectively) together account for 61.5% of the graphematic representations for the Kōjien headword list. However, although such concatenations of kanji are prototypically characteristic of SJ compounds, it should be stressed immediately that, because their analysis was purely from the perspective of graphematic representation, Joyce, Masuda, and Ogawa (2014) did not seek to explicitly control for lexical strata. Thus, while it is reasonable to assume that the majority of those compound words are SJ compounds, it should also be acknowledged that the frequency counts, particular the 2KCW count, also include some proportion of NJ compound words.

Even though many combinations of two NJ morphemes are graphematically represented with two kanji, pronounced according to their NJ readings, such as 大雨 /ō-ame/ *heavy rain* [big + rain] (Masuda and Joyce, 2018), such combinations more frequently yield graphematic representations that are mixtures of kanji and hiragana.[6] Thus, in contrast to the three most frequent graphematic representations being predominately SJ compounds, the fourth to sixth most frequent codes of 2CH, CHC, and CHCH are likely to be predominately NJ compound words, as illustrated in (2).

(2) 2CH 南向き /minimi-mu.ki/ *facing south* [south + face toward]
底堅い /soko-gata.i/ *stable (market) after bottoming out* [bottom + firm][7]
CHC 食べ物 /ta.be.mono/ *food* [eat + thing]
泣き声 /na.ki.goe/ *cry, crying voice* [cry + voice]
CHCH 立ち読み /ta.chi.yo.mi/ *reading while standing (in store)* [stand + read]
売り買い /u.ri.ka.i/ *trade; buying and selling* [sell + buy]

[6] This is because the 連用形 /ren-yō-kei/ *infinitive form* [connect + use + form] of many NJ verbs and adjectives consists of a stem and inflection, which are graphematically represented by a kanji and a hiragana, respectively.

[7] Kageyama and Saito (2016, p. 20) cite this, together with 高止まり /takado.mari/ *remaining high* [high + stop] (2C2H), as evidence of newly coined NJ compounds being common in specialized fields, like the stock market.

In concluding their survey of the word-formation processes and productivity of SJ words, Kobayashi, Yamashita, and Kageyama (2016) single out two reasons why SJ words are so productive (as evidenced in the considerable gaps between the frequencies and percentages of the three most frequent graphematic representation codes compared to the subsequent three codes in Table 2). The first is what Kobayashi, Yamashita, and Kageyama (ibid., p. 129) refer to as a visual factor; namely, "that the meanings of the component morphemes are easily comprehended through the *kanji*" (italics in original). The second reason, which they regard as being the more important, is what they refer to as relaxed restrictions on compound lengths when the compound head is a SJ morpheme. In that context, Kobayashi, Yamashita, and Kageyama (ibid., p. 129) particularly emphasize "the iterative application of compounding rules to produce compounds four or more characters in length and the vigor of affixes that can attach to bases of three or more characters".[8]

Unquestionably, the morphology of Japanese compound words is an especially interesting topic from the perspectives of both writing systems research and the related areas of psycholinguistic research into visual word recognition and the mental lexicon. In light of growing research interest into the representation and retrieval of morphological information within the mental lexicon, Kobayashi et al.'s (2016, p. 129) claim that "kanji play an important role in providing the readers of written Japanese with a visual aid for capturing the meaning of a word at a glance" undoubtedly warrants further empirical investigation. One potentially fertile approach in that respect could be to conduct visual word recognition experiments that utilize the constituent priming paradigm with Japanese compound words of various lengths (Joyce, 2002; Masuda and Joyce, 2018). Moreover, given that kanji are associated with both NJ and SJ morphemes, analyses of the morphological structures of Japanese compound words can potentially further illuminate the intricate nature of morphography in the case of the JWS; a topic of potentially profound significance for writing systems research.

Against such background considerations, this paper reports on the construction of two new databases of 3KCWs and 4KCWs, which have been compiled as components of a larger database project concerned

[8] However, in order to more appropriately contextualize this comment, it should also be noted that Kobayashi et al.'s (2016) chapter outline only includes sections up to four-character SJ words. As they explain, although it is theoretically possible to construct SJ words of unlimited lengths, such words are inevitably combinations of compound word elements. In illustration, Kobayashi, Yamashita, and Kageyama (2016, pp. 114–115) analyze 新社屋建設案発表会 /shin-sha-oku ken-setsu-an hap-pyō-kai/ *presentation of plan for construction of new company building* according to its component structure, working from its head of 発表会 [[disclose + diagram = presentation] + gathering] for the announcement of the 建設案 [[build + establish = construction] + plan] for the 新社屋 [new + [company + building]].

with Japanese lexical properties (Joyce, Hodošček, and Masuda, 2017; Joyce, Masuda, and Ogawa, 2014). The overarching objective of the larger database project is to compile a database of scale, which can support linguistic and psycholinguistic research on the Japanese lexicon, such as facilitating the selection of stimuli for psycholinguistic surveys and priming experiments (Masuda and Joyce, 2018). Consistent with common practice (Kobayashi, Yamashita, and Kageyama, 2016), the component databases for the larger database project focus on different aspects of the Japanese lexicon, as such compound words according to their overall lengths and targeted lexical properties. For instance, Masuda and Joyce (2005) supplemented a list of 2KCW headwords extracted from the fifth edition of Kōjien (Shinmura, 1995) with various data relating to morphological family sizes, morphological structures and semantic categories, while Masuda, Joyce, et al. (2014) focused on semantic transparency ratings for 2KCWs. The present paper focuses primarily on the analyses of the new database components in terms of the morphological structures of the 3KCWs and 4KCWs, respectively. As the target compound words were extracted according to their graphematic representations, without explicitly controlling for lexical strata, while the majorities of the 3KCWs and 4KCWs will be SJ, inevitably, some proportion of both databases will be either NJ or hybrid compound words. After briefly outlining the extracting and cleaning of the two database lists in Section 2, Sections 3 and 4 present the results of analyzing the morphological structures of the 3KCW and 4KCWs, respectively. The paper ends with a short section of concluding remarks.

2. List Extraction and Cleaning

Although the analyzed lists of 3KCWs and 4KCWs were extracted on separate occasions, the two-stage extraction procedures were identical in both cases. During the respective first stages, all the relevant compound words were extracted from the set of corpus word lists (CWLs) that Joyce, Hodošček, and Nishina (2012) compiled from the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Joyce, Hodošček, and Masuda, 2017; Maekawa et al., 2013). Joyce et al.'s (2012) CWLs are grouped according to both the two word-units definitions[9] and the word class divisions used within the BCCWJ project, and all CWL files,

[9] The main lexical demarcation employed with the BCCWJ is a somewhat elusive one in distinguishing between short-unit words (SUWs) and long-unit words (LUWs). Although the short-long labels evoke a length-based contrast, as Joyce, Masuda, and Ogawa (2014) explain, the distinction is essentially of lexical status, such that SUWs include both bound morphemes and simple words (dictionary headwords) and LUWs are complex words and phrases.

apart from the proper noun files, were examined to check for the possible presence of target compound words. In addition to recording the CWL source file, all of the CWL's lexical information was retained for reference in analyzing the compound words. This information includes columns for the underlying lemma entry, lemma length (used to extract target compounds), number of graphematic variants, etymology code, BCCWJ frequency of the lemma, orthographic base (graphematic variants of a lemma), orthographic base pronunciations, lengths of orthographic bases, BCCWJ frequency of the orthographic base, and ratio of total lemma frequency covered by a particular orthographic base form. Stage 1 processing resulted in spreadsheets of 171,123 rows of 3KCWs and 298,944 rows of 4KCWs.

The substantial disparity in the numbers of spreadsheet rows for the 3KCWs and 4KCWs extracted from the CWLs is consistent with the analyses of graphematic representation codes that Joyce, Hodoščček, and Masuda (2017) also conducted for Joyce et al.'s (2012) CWLs. Focusing only on the relevant long-unit word (LUW) data, even though the first and second most frequent graphematic representation codes by types counts were 4C (15.4%) and 3C (9.3%), respectively, by token counts, the 3C code was only the eighth most frequent (3.1%) and the 4C code did not feature amongst the top ten codes at all. Those findings indicate that, although there are far fewer 3KCWs than 4KCWs within the Japanese lexicon overall, 3KCWs generally tend to occur more frequently than 4KCWs.

In order to derive lists of more practical lengths for analyses, the respective second stages commenced by first applying the criterion that the BCCWJ lemma frequencies (token counts) should be either equal to or greater than 10. Moreover, reflecting the automatic nature of the methods used in extracting the CWL source corpus, additional cleaning work was required to remove some non-words, some proper nouns and to merge for cases of lemma replications. Accordingly, Stage 2 processing resulted in database lists of 23,046 3KCW-lemmas and 23,159 4KCW-lemmas. Although the application of the frequency criterion yielded highly comparable lists in terms of the overall numbers of compound word lemmas that each database component contains, naturally, the impact of eliminating compound words with frequencies of less than 10 was far greater in the case of the 4KCWs. That is, although Stage 2 processing for the 3KCWs yielded a list that was 13.5% of the Stage 1 extracted list, Stage 2 processing for the 4KCWs yielded a list that contained only 7.75% of the Stage 1 extracted list. It should also be noted that while the distributions of lemma frequencies are generally consistent for both database lists, with both being typical of corpus frequencies, the 3KCWs are generally of higher token frequency counts compared to the 4KCWs, as the plots of log-transformed frequencies in Figure 1 indicate. More specifically, for the 3KCWs, the frequency range is from 10 to 18,395 with a mean of 88.5 and median of 25, while for the 4KCWs, the frequency range is from 10 to 4,127 with a mean of 43.1 and a median of 20.

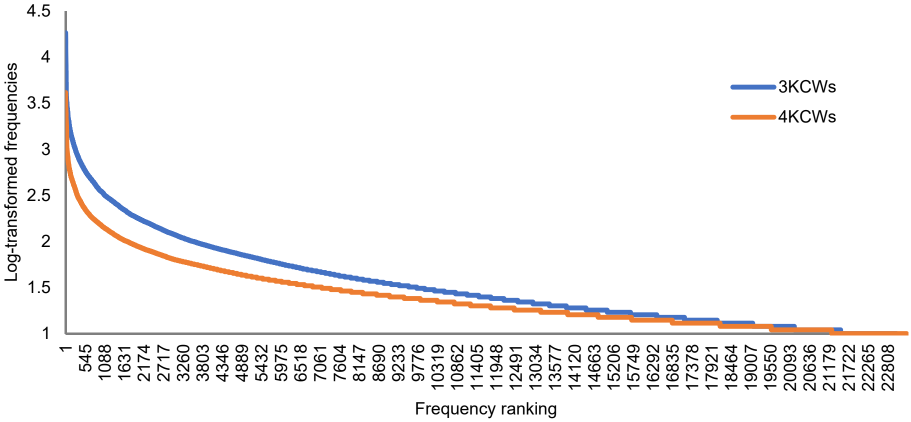

FIGURE 1. Log-transformed lemma frequencies of the 3KCWs and 4KCWs

Even though the BCCWJ's original lexical strata codes (i.e., NJ, SJ or hybrids) are retained within the respective databases compiled from the CWLs, the primary criterion for inclusion has been the appropriate lemma length of 3KCWs and 4KCWs, respectively. Thus, while acknowledging that both database lists contain some proportions of NJ and hybrid compound words and that awareness concerning the lexical stratum of the components often greatly informs the appropriate classifications of the compound words, the analyses of morphological structures reported in the subsequent sections do not explicitly consider the lexical stratum of the component elements. The conducted analyses of both database lists adopted similar conventions for denoting the constituent component kanji, which were designated as A, B, and C (3KCWs), as well as D (4KCWs), respectively, with square-brackets used to indicate internal structures, such as [AB]+C to indicate a 2KWC with a C addition and [AB]+[CD] to indicate a combination of two 2KCWs. Moreover, as Kobayashi, Yamashita, and Kageyama (2016, p. 108) emphasize, with SJ morphemes, in particular, it can often be quite difficult to discern both a morpheme's status, as either a free word or bound element, and the word-formation process that underlies a particular compound word, as either involving compounding or affix-derivation. Accordingly, in considering the appropriate classification of all compound words, we have also checked for alternative structures. To that end, all compound words were initially segmented and the component kanji were subsequently recombined to consider for all possible structures. For example, as 農業 /nō-gyō/ *agriculture* [agriculture + business] and 業者 /gyō-sha/ *trader, business person* [business + person] both exist as 2KCWs, it is necessary to consider all component meanings and usage patterns to determine that [AB]+C is the more coherent interpretation of 農業者 /nō-gyō-sha/ *agricultural worker* [agriculture + business + person].[10]

3. The Morphological Structures of the 3KCW Database

Although Joyce and Masuda (2019) tendered an initial report about compiling this database list of 23,046 3KCWs, this paper describes the results of analyzing their morphological structures in a little more detail. In addition to presenting a summary table of the morphological structures, Section 3.1 includes a table of the top 20 most frequent 3KCWs, as well as some general analyses of the A and C additions to 2KCWs. Three further sub-sections focus on the morphological structures, with Section 3.2 on the primary structure of [AB]+C, Section 3.3 on the secondary structure of A+[BC], and Section 3.4 on the remaining 3KCW structures.

3.1. Morphological Structures of the 3KCW Database: Summary and A + C Additions

Table 3 presents the breakdown of the 3KCW database list according to their morphological structures, with both type counts and their corresponding percentages. As the morphological structures of 3KCWs are generally transparent, it has been possible to confidently classify the database list according to eight morphological structures.[11] As Table 3 clearly indicates, the primary morphological structure of [AB]+C is highly dominant in accounting for 77.1% of the database list. In contrast, the secondary structure of A+[BC] only accounts for 21.3% overall, which is about one-third of the primary structure's percentage. However, taking the primary and secondary structures together, they account for the vast majority of 3KCWs, at 98.4% for the database list, with six other structures underlying the remaining 1.6%. Firmly underscoring the profound significance of 2KCWs within the Japanese lexicon

[10] Although we regard the [AB]+C classification as being the more plausible interpretation, we are also planning to conduct psycholinguistic surveys to investigate the extent to which alternative structures might be activated in the processing of such compound words.

[11] Table 3 also includes an adjustment category of multiple types for a few 3KCWs that are open to alternative analyses.

(Joyce, 2011; Joyce, Hodoščeck, and Masuda, 2017; Kobayashi, Yamashita, and Kageyama, 2016; Nomura, 1975; 1988), the majority of 3KCWs are 2KCWs combined with an additional morpheme; either predominately attached to the end or, in considerable cases, inserted at the beginning.

TABLE 3. Breakdown of the morphological structures in the 3KCW database

Morphological structure	Type counts	Percentage
[AB]+C	17,761	77.1
A+[BC]	4,904	21.3
[A(C*)]+[BC] (with (C*) omitted)	154	0.7
Non-divisible	93	0.4
Phonological transcription (当て字)	64	0.3
Monomorphemic (熟字訓)	45	0.2
A+B+C	25	0.1
[AB]+[(A*)C] (with (A*) omitted)	15	0.1
Multiple types (count adjustment)	-15	−0.1
Total	**23,046**	**100**

Table 4 presents the 20 most frequent 3KCWs based on token frequency counts, which indicates that frequency is independent of morphological structure. Although the primary morphological structure of [AB]+C is the most frequent among these most frequent 3KCWs, which is consistent with the overall analysis results, other morphological structures are also associated with highly frequent 3KCWs, such as 雰囲気 /fun-i-ki/ *mood; ambience* [atmosphere + surround + spirit], which is classified as non-divisible. Although each of the SJ morphemes contributes semantically to some degree to the overall meaning of this 3KCW, its original etymology is no longer obvious.

Understandably, a sizeable proportion, at 12.0% of the 3KCWs are combinations of number kanji with various numerical units and classifiers and, as may be also discerned from Table 4, some of these are of high frequencies, such as 三十分 / san-jip-pun/ *thirty minutes* [[three + ten = thirty] + minutes] and 十二月 /jū-ni-gatsu/ *December; 12 months* [[ten + two = twelve] + month].

Before turning to the dominant primary and secondary morphological structures, as the vast majority of 3KCWs involve either a single SJ or NJ morpheme being added to an existing 2KCW, it is beneficial to also note Masuda and Joyce's (2019) separate analyses of the A and C additions. Notwithstanding the challenges, with most kanji being associated with both multiple SJ and multiple NJ morphemes and that the status of

Table 4. 20 most frequent 3KCWs by token frequency counts for the orthographic base

3KCW	Structure	Gloss	Translation and explanation	Frequency
大丈夫	A+[BC]	dai-jō-bu	problem-free [big + [stature + man = healthy]]	16,861
可能性	[AB]+C	ka-nō-sei	possibility [[can + able = possible] + *-ity* ending]	13,555
不思議	A+[BC]	fu-shi-gi	mysterious [negative + [think + debate = conjecture; guess]]	13,044
雰囲気	Non-divisible	fun-i-ki	mood; ambience [atmosphere + surround + spirit]	11,427
三十分	[AB]+C	san-jip-pun	thirty minutes [[three + ten = thirty] + minutes]	11,042
十二月	[AB]+C	jū-ni-gatsu	December; 12 months [[ten + two = twelve] + month]	10,325
十一月	[AB]+C	jū-ichi-gatsu	November; 11 months [[ten + one = eleven] + month]	9,739
具体的	[AB]+C	gu-tai-teki	concrete, specific [[means + substance = tangible] + *-ic* adjectival noun (AN) ending]	9,334
基本的	[AB]+C	ki-hon-teki	fundamental, basic [[foundation + base = basics] + *-ic* AN ending]	8,251
第一項	[AB]+C	dai-ik-kō	first item [[ordinal number + one = first] + item; clause]	7,261
積極的	[AB]+C	sek-kyoku-teki	positive; active [[amass + poles = active, positive] + *-ic* AN ending]	7,060
大統領	A+[BC]	dai-tō-ryō	president [big + [govern + territory = ruler; leader; consul]]	6,992
出来事	[AB]+C	de-ki-goto	incident; event [[go out + come = occurrence; happening] + thing]	6,189
一般的	[AB]+C	ip-pan-teki	general, typical [[one + general = general; ordinary] + *-ic* AN ending]	5,932
不可欠	Non-divisible	fu-ka-ketsu	indispensable; essential [negative + can + lack; fail]	4,209
十五日	[AB]+C	jū-go-nichi	15th; 15 days [[ten + five] + day]	3,967
高齢者	[AB]+C	kō-rei-sha	elderly person/people [[high + age = old] + person]	3,902
青少年	[A(C*)]+[BC]	sei-shō-nen	youths, young people [[green + years* = youths] + [few + years = youth]]	3,751
二十日	[AB]+C	hatsu-ka	20th; 20 days [[two + ten] + day]	3,608
小学校	A+[BC]	shō-gak-kō	elementary school [small + [study + school = school]]	3,540

any given kanji can vary across different 3KCWs,[12] Masuda and Joyce analyzed the additional A and C components according to their morpheme status, as either free, bound or affix morphemes. The analysis results are presented in Table 5.

TABLE 5. Results of analysing A and C additional components in terms of their morpheme status

	A additions		C additions	
Morpheme status	Type count	Percentage	Type count	Percentage
Free	360	55.0	369	44.0
Bound	225	34.4	401	47.9
Affix	70	10.7	68	8.1
Total	**655**	**100.0**	**838**	**100.0**

3.2. Primary Morphological Structure of [AB]+C 3KCWs

As Table 3 vividly attests, 3KCWs overwhelmingly conform to the morphological structure of [AB]+C, where a single morpheme is appended to an existing 2KCW. Accordingly, Table 6 first presents the ten most frequent C-additions in terms of their type counts, which indicates their productivity in combining with multiple 2KCWs, then Table 7 presents the most frequent 3KCWs by token counts, for each of the 10 most frequent C-additions.

Being wholly consistent with Kobayashi et al.'s (2016, p. 127) comment that 的 /teki/ *AN ending*[13] is "a representative, highly productive Sino-Japanese affix that combines with a variety of bases," it is not in the least surprising to find that it is the most productive of the C-additions observed within the 3KCW database list. Indicative of its wide applicability, 的 is a C-component of 3KCWs across the database's entire frequency range. In addition to being a C-addition to four of the 20 most

[12] Kobayashi, Yamashita, and Kageyama (2016, pp. 95–96) classify one-character SJ morphemes as free (会 /kai/ *meeting*) or bound—either connectives (運転中 /unten-chū *while driving*), or the bases of verbs (信じる /shin.jiru/ *believe*), of ANs (急な /kyū.na/ *abrupt*), of adverbs (実に /jitsu.ni/ *actually*), of adnominal/adverbial modifiers (単なる /tan.naru/ *mere*). Kobayashi et al. also regard some bound morphemes as affixes due to their positional constraints, such as 最 /sai/ *most* as a prefix of 最先端 /sai-sen-tan/ *cutting edge* [most + front + edge] (p. 108).

[13] As Kageyama and Saito (2016, p. 18) note, the lexical category of adjectival noun does not exist in English or other European languages. While morphologically a noun, it can function syntactically as an adjective with -だ /-da/ in predicates and -な /-na/ adnominally.

TABLE 6. Top ten most frequent C-additions to [AB]+C 3KCWs by type counts

C-addition	Meaning	Type count
的	*-ic* AN ending	873
者	*-er* person-indicating ending	685
等	etc.; and so forth	577
性	*-ity* ending; nature	498
中	in/during [place or time]	352
化	*-ization* verbal noun (VN) ending	294
後	after	253
達	pluralizing ending	244
上	above; in terms of	239
人	*-er* person-indicating ending	227

TABLE 7. Most frequent [AB]+C 3KCWs by token counts, for each of the most frequent C-additions

3KCW	Gloss	Translation and explanation	Frequency
具体的	gu-tai-teki	concrete [[means + substance] + AN ending]	9,334
高齢者	kō-rei-sha	elderly person/people [[high + age] + person]	3,902
整備等	sei-bi-tō	maintenance etc. [[organize + equip] + etc]	608
可能性	ka-nō-sei	possibility [[can + able] + *-ity* ending]	13,555
世界中	se-kai-jū	around the world [[world + world] + thoughout]	2,034
生活化	sei-katsu-ka	living [[life + active] + VN ending]	1,195
十年後	jū-nen-go	after 10 years; 10 years later [[ten + year] + after]	476
子供達	ko-domo-tachi	children [[child + accompany] + pluralizer]	886
事実上	ji-jitsu-jō	as a matter of fact [[thing + real] + in terms of]	1,396
外国人	gai-koku-jin	foreigner [[outside + country] + *-er* person]	2,361

frequent 3KCWs (Table 4), some other 3KCW examples that vary in terms of their token frequencies are listed in (3).

(3) 比較的 /hi-kaku-teki/ *comparatively* 3,515
[[compare + contrast] + *-ic* AN]
国際的 /koku-sai-teki/ *international* 2,106
[[country + occasion; side] + *-ic* AN]

本質的 /hon-shitsu-teki/ *intrinsic; substantial* 1,026
[[true + quality] + *-ic* AN]
潜在的 /sen-zai-teki/ *implicit, latent* 506
[[conceal + exist] + *-ic* AN]
挑戦的 /chō-sen-teki/ *challenging; provocative* 99
[[contend + battle] + *-ic* AN]

As Kobayashi, Yamashita, and Kageyama (2016) point out, 的 combines with various bases, as their examples in (4) illustrate, and, consistently, it is one of the most frequent D-additions to 4KCWs.

(4) 私的 /shi-teki/ or /watashi-teki/ private; personal [I + -ic AN]
活動的 /katsu-dō-teki/ *active; dynamic* [[active + move] + *-ic* AN]
政治家的 /sei-ji-ka-teki/ *politician-like*
[[[politics + rule] + person] + *-ic* AN]
共産主義的 /kyō-san-shu-gi-teki/ *communistic*
[[[together + produce] + [principle + meaning]] + *-ic* AN]
草分け的 /kusa-wa.ke.teki/ *pioneering* [[grass + divide] + *-ic* AN]
カリスマ的 /karisuma.teki/ *charismatic* [charisma + *-ic* AN]

As Table 6 shows, the tenth most productive C-addition is 人 person, reflecting its generic sense, but as Kobayashi, Yamashita, and Kageyama (ibid., p. 127) point out, it is associated with two SJ morphemes. The first is /jin/, which attaches to both nouns and stems of ANs, such as the examples in (5).

(5) 外国人 gai-koku-jin *foreigner* [[outside + country] + -er person]
芸能人 gei-nō-jin *performer* [[perform + talent] + *-er* person]
有名人 yū-mei-jin *famous person* [[possess + name] + *-er* person

The second SJ morpheme is /nin/, which only attaches to verbal nouns (VN),[14] such as the examples in (6). A further restriction is that while /nin/ attaches to NJ bases, such as 受け取り人 /u.ke.to.ri.nin/ *receipent* [[receive + take] + person], /jin/ does not, apart from the single exception of 暇人 /hima-jin/ *person of leisure* [leisure + person].

(6) 通行人 /tsū-kō-nin/ *passerby* [[pass through + go] + person]
弁護人 /ben-go-nin/ *advocate; defender*
[[speech + safeguard] + person]
管理人 /kan-ri-nin/ *manager; administrator*
[[control + arrange] + person]

[14] As Kageyama and Saito (2016, p. 18) also stress, the verbal noun (VN) is another lexical categories that does not exist in European languages. Kagayama and Saito describe VNs as "hybrid category" of a noun that can function as a verb when combined with the dummy verb する /suru/.

3.3. Secondary Morphological Structure of A+[BC] 3KCWs

Although not as common as the primary structure of [AB]+C 3KCWs, the secondary structure of A+[BC] accounts for approximately one-fifth (21.3%) of the 3KCW database list. Table 8 presents the ten most frequent A-additions in terms of their type counts, while Table 9 presents, for each of the ten most frequent A-additions, the most frequent A+[BC] 3KCWs with the respective A-additions.

TABLE 8. Top ten most frequent A-additions to A+[BC] 3KCWs by type counts

A-addition	Meaning	Type count
御	honorific prefix	430
大	large; big	313
各	each; every	152
不	negative prefix *non-*	143
新	new	127
一	one	126
無	negative prefix *un-*, *non-*	95
同	same	93
諸	various; several	90
全	all; whole	86

TABLE 9. Most frequent A+[BC] 3KCWs by token counts, for each of the most frequent A-additions

3KCW	Gloss	Translation and explanation	Frequency
御指摘	go-shi-teki	as you indicate [honorific + [point + pinch]]	2,171
大丈夫	dai-jō-bu	problem-free [big + [stature + man = healthy]]	16,861
各地域	kaku-chi-iki	each region [each + [ground + region]]	388
不思議	fu-shi-gi	mysterious [negative + [think + debate]]	13,044
新幹線	shin-kan-sen	bullet train [new + [trunk + line]]	1,118
一時間	ichi-ji-kan	one hour [one + [time + interval]]	6,515
無意識	mu-i-shiki	unconsciousness [un- + [mind + know]]	1,263
同級生	dō-kyū-sei	classmate [same + [rank; class + student]	840
諸外国	sho-gai-koku	various foreign countries [various + [out + country]	744
全世界	zen-se-kai	whole world [all + [[world + world]	619

While Kobayashi, Yamashita, and Kageyama (2016, p. 123) acknowledge that there are considerable cases of SJ words “where the distinction between affix and compound constituent is not clear,” they also stress that A-additions often represent substantive semantic concepts. Indeed, they provide a number of examples, which they organize according to five semantic functions, including (a) limiting or modifying the base meaning, (b) verbal meaning that corresponds to the base noun’s argument, (c) limiting the base noun’s reference, (d) adverbially modifying a predicate-like base, and (e) indicating negation. Examples for each of these five semantic functions are given in (7).

(7)	好成績	/kō-sei-seki/ *good results* [pleasing + [become + achievements]]
	反体制	/han-tai-sei/ *anti-establishment* [opposite + [body + system]]
	本製品	/hon-sei-hin/ *this product* [this; main + [manufacture + goods]]
	急成長	/kyū-sei-chō/ *rapid growth* [rapid + [become + long]]
	未経験	/mi-kei-ken/ *inexperienced* [not yet + [pass thru + effect]]

Instructively, Kobayashi, Yamashita, and Kageyama (ibid., pp. 126–127) differentiate between the four A-additions that signify negative senses in terms of their nuances and the categories of bases to which they attach. Consistent with its fourth place ranking amongst the most productive A-additions, Kobayashi, Yamashita, and Kageyama (ibid.) comment that 不 /fu/ and /bu/ *negative* is the most productive and attaches to nouns, adjectival nouns and verbal nouns, as in (8), respectively.

(8)	不景気	/fu-kei-ki/ *recession* [negation + [view + spirit; atmosphere]]
	不確実	/fu-kaku-jitsu/ *uncertain* [negation + [confirm + reality]]
	不承知	/fu-shō-chi/ *disapproval* [negation + [acquiesce + know]]

The next most productive A-addition with negative connotations is 無 /mu/ *lacking, non-existent*, which attaches to nouns and verbal nouns, but not adjectival nouns, as in (9).

(9)	無関心	/mu-kan-shin/ *unconcerned* [lacking + [connection + heart]]
	無関係	/mu-kan-kei/ *unrelated* [lacking + [connection + connection]]

As an A-addition of 45 3KCWs within the database list, the third most productive of the A-additions with negative connotations is 未 /mi/ *not yet*, which attaches to nouns and verbal nouns, but not adjectival nouns, as in (10).

(10)	未成年	/mi-sei-nen/ *not of age* [not yet [become + age]]
	未解決	/mi-kai-ketsu/ *unresolved* [not yet + [unravel; solve + decide; fix]]

While only attested as an A-addition to 38 A+[BC] 3KCWs within the database list, 非 /hi/ *negation* also attaches to nouns, adjectival nouns and verbal nouns, as in (11) respectively.

(11) 非人情 /hi-nin-jō/ *inhuman* [negation + [person + feelings]]
非合法 /hi-gō-hō/ *unlawful* [negation + [fit; suit + law, rule]
非公認 /hi-kō-nin/ *unauthorized*
[negation + [public; official + acknowledge]]

3.4. Other 3KCW Morphological Structures

Although our analysis of the morphological structures of the 3KCW database reveals that the two structures of [AB]+C and A+[BC] account for the vast majority (98.4%) of 3KCWs, as Table 3 also indicates, six other morphological structures underlie a small percentage of 3KCWs. Accordingly, this section turns to present examples of 3KCWs that conform to those other morphological structures.

Albeit on a distinctly smaller scale (0.7%), the third most frequent morphological structure is [A(C*)]+[BC], where the C-component of an [AC] 2KCW is omitted and the resultant A is attached to a related [BC] 2KCW. The practice of omitting the C-component of an [AC] 2KCWs is undoubtedly a form of clipping that is common with SJ words (Kobayashi, Yamashita, and Kageyama, 2016, p. 128).[15] As such, superficially, this structure may appear to resemble the A+[BC] structure, in the sense, that it effectively involves an A-component being inserted before a [BC] 2KCW. It is, however, appropriate to differentiate them, because the [A(C*)]+[BC] structure crucially hinges on the semantic relationship between the [AC] and [BC] 2KCWs, due to their shared C-component, as the examples in (12) illustrate.

(12) [A(C*)]+[BC] (with (C*) omitted)
視聴覚 /shi-chō-kaku/ *audiovisual* [視覚 vision + 聴覚 hearing]
入出国 /nyū-shutsu-koku/ *immigration*
[入国 enter country + 出国 depart county]

[15] Clipping with SJ words most typically involves 4KCWs being shortened to 2KCWs, such as 模擬試験 /mo-gi-shi-ken/ *practice test* → 模試 /mo-shi/ (A + C) or 高等学校 /kō-tō-gak-kō/ *high school* → 高校 /kō-kō/ (A + D). One important consequence of such clipping processes is that the resultant 2KCWs tend to have far higher frequencies than the corresponding 4KCW, such as 就活 /shū-katsu/ *job hunting* [take position + activity] which is derived by clipping from 就職活動 /shū-shoku-katsu-dō/ *job hunting* [position + post + lively + move].

The fourth category of non-divisible is necessary to handle the small set of exceptions (0.4%). Some 3KCWs are classified as non-divisible, because the compound word's etymology and morphological structure are not clear, even though the meanings of the component morphemes are usually related to the overall meaning. Other 3KCWs classified as non-divisible are the results of clipping processes applied to longer compound words. One example of each is presented in (13).

(13) Non-divisible

方程式 /hō-tei-shiki/ *equation; formula*
[direction + formula + expression]
食洗機 /shoku-sen-ki/ *dishwasher* ← 食器洗浄機
[[eat + ware = dishes] + [[wash + clean = washing] + machine]]

The fifth category is phonological transcriptions (0.3%), known as 当て字 /a.te.ji/ *phonological transcription* [apply + character] in Japanese, which refers to the convention of phonologically representing a word's syllables with kanji. Although phonological transcriptions are essentially a form of the rebus principle, the individual kanji used for such graphematic representations often have some degree of semantic relevance to the word's meaning, such as in the first example in (14), but sometimes less so, as in the second example.

(14) Phonological transcriptions (当て字)

歌舞伎 /kabuki/ *kabuki; Japanese classical drama* [sing + dance + art]
目論見 /mokuromi/ *plan; scheme; plot* [eye + argument + see]

The sixth category is monomorphemic words (0.2%), known as 熟字訓 /juku-ji-kun/ *monomorphemic word* [compound + character + semantic translation] in Japanese, which refers to the convention of representing the meaning of an NJ word with kanji that are semantically related. In contrast to phonological transcriptions where the kanji are representing the syllables of the word, there is usually no phonological correspondence between the elements of the graphematic representation, but the meanings of the component kanji are related to the word's meaning. The first example in (15) may be regarded as the prototypical example that is frequently cited in illustration.

(15) Monomorphemic words (熟字訓)

五月雨 /samidare/ *early summer rain* [five + month + rain]
波止場 /hatoba/ *wharf; quay* [wave + stop + place]

The seventh morphological structure is A+B+C (0.1%), as the concatenation of three morphemes that together constitute some form of set or may be regarded as exemplars of the compound word's meaning, as both the examples in (16) indicate.

(16) A+B+C

衣食住	/i-shoku-jū/ *necessities of life* [clothing + food + shelter]
産官学	/san-kan-gaku/ *industry, government and academia* [industry + government + academia]

The eighth and final morphological structure is [AB]+[(A*)C] (0.1%), where the A-component of an [AC] 2KCW is omitted and the resultant C is attached to an [AB] 2KCW. Like the [A(C*)]+[BC] structure, the omitting of the A-component of an [AC] 2KCW is also a form of clipping. Also similar to the [A(C*)]+[BC] structure, it is appropriate to differentiate this from the primary morphological structure of [AB]+C 2KCWs, because the [AB]+[(A*)C] structure also hinges on the semantic relationships between the [AB] and [AC] 2KCWs, due to their shared A-component, as the examples in (17) illustrate.

(17) [AB]+[(A*)C] (with (A*) omitted)

国内外	/koku-nai-gai/ *domestic + foreign* [国内 domestic + 国外 foreign]
十五六	/jū-go-roku/ *15 or 16* [十五 15 + 十六 16]

4. Morphological Structure Results for the 4KCW Database

Having presented the results of analyzing the morphological structures of the 3KCW database component in some detail, this paper now turns to present the results for the 4KCW database. Adopting a similar organization to the previous section, Section 4.1 starts with a summary table of the morphological structures and a table of the top 20 most frequent 4KCWs. Four further sub-sections focus on the various morphological structures, with Section 4.2 on the primary structure of [AB]+[CD], Section 4.3 on the second structure of [ABC]+D, Section 4.4 on the tertiary structure of A+[BCD], and Section 4.5 on the remaining 4KCW structures.

4.1. Morphological Structures of the 4KCW Database: Summary

Table 10 presents the breakdown of the database list of 23,159 4KCWs according to their morphological structures, with both type counts and their corresponding percentages.

As the morphological structures of 4KCWs are also generally highly transparent, it has been possible to confidently classify the database

TABLE 10. Breakdown of the morphological structures in the 4KCW database

Morphological structure	Type counts	Percentage
[AB]+[CD]	19,805	85.3
[ABC]+D	2,809	12.1
A+[BCD]	449	1.9
Non-divisible	23	0.1
[A(CD*)]+[BCD] (with (*CD) omitted)	18	0.1
[A(D*)]+[B(D*)]+[CD] (with both (*D) omitted)	16	0.1
A+B+C+D	16	0.1
Phonological transcriptions (当て字)	14	0.1
[AB]+C+D	6	0.0
Monomorphemic (熟字訓)	2	0.0
[A(D*)]+[BCD] (with (*D) omitted)	1	0.0
Total	**23,159**	**100**

list according to 11 morphological structures.[16] Similar to the results for the 3KCWs, the analyses of the morphological structures within the 4KCWs reveals that one structure dominates in accounting for 85.3% of the 4KCW types. However, in the case of 4KCWs, the primary morphological structure is [AB]+[CD], which is consistent with Kobayashi et al.'s (2016, p. 113) comment that "four-character S-J words are words composed of four S-J morphemes, which are typically divided into two words, each consisting of two morphemes". This primary structure also attests to the immense significance of 2KCWs within the Japanese lexicon (Joyce, 2011; Joyce, Hodoščěk, and Masuda, 2017; Kobayashi, Yamashita, and Kageyama, 2016; Nomura, 1975; 1988).

Reflecting the even greater dominance of the [AB]+[CD] structure, in contrast, the secondary structure of [ABC]+D and the tertiary structure of A+[BCD] account for 12.1% and 1.9%, respectively, of all 4KCW structures. Naturally, there are parallels between these morphological structures and the primary and secondary structures of 3KCWs, as they also involve combining an additional morpheme with an existing compound word and the marked preference is for attaching that additional morpheme to the end rather than inserting at the beginning. However, reflecting the even greater dominance of the primary structure, the secondary and tertiary structures are relatively less common for 4KCWs.

[16] Claiming that the structures of 4KCWs "can be categorized by the patterns of binary branching structures," Kobayashi, Yamashita, and Kageyama (2016, pp. 114–115) list nine patterns under four types, following Nomura (1975). Reflecting its importance, the first type is [AB]+[CB], the second is of 3KCWs plus additions (i.e., [ABC]+D and A+[BCD]), the third involves combinations (i.e., [ACD*]+[BCD] and [AD*]+[BD*]+[CD]), and the fourth is A+B+C+D. However, their list does not include either phonological transcriptions or monomorphemic words.

Table 11. 20 most frequent 4KCWs by token frequency counts

4KCW	Structure	Gloss	Translation and explanation	Frequency
自分自身	[AB]+[CD]	ji-bun-ji-shin	oneself [[oneself + one’s lot = oneself] + [oneself + someone = oneself]]	3,288
三十一日	[ABC]+D	san-jū-ichi-nichi	31st; 31 days [[three + ten + one] + day]	3,238
二十五日	[ABC]+D	ni-jū-go-nichi	25th; 25 days [[two + ten + five] + day]	3,058
二十一日	[ABC]+D	ni-jū-ichi-nichi	21st; 21 days [[two + ten + one] + day]	2,718
二十四日	[ABC]+D	ni-jū-yok-ka	24st; 24 days [[two + ten + four] + day]	2,716
二十三日	[ABC]+D	ni-jū-san-nichi	23st; 23 days [[two + ten + three] + day]	2,660
二十二日	[ABC]+D	ni-jū-ni-nichi	22rd; 22 days [[two + ten + two] + day]	2,659
二十八日	[ABC]+D	ni-jū-hachi-nichi	28th; 28 days [[two + ten + eight] + day]	2,642
都道府県	A+B+C+D	to-dō-fu-ken	administrative divisions [Tokyo + Hokkaido + Osaka/Kyoto + prefectures]	2,621
二十六日	[ABC]+D	ni-jū-roku-nichi	26th; 26 days [[two + ten + six] + day]	2,587
二十七日	[ABC]+D	ni-jū-shichi-nichi	27th; 27 days [[two + ten + seven] + day]	2,495
二十九日	[ABC]+D	ni-jū-ku-nichi	29th; 29 days [[two + ten + nine] + day]	2,492
政府委員	[AB]+[CD]	sei-fu-i-in	ministerial aide [[politics + government] + [committee + member]]	2,336
中小企業	[AB]+[CD]	chū-shō-ki-gyō	small-medium companies [[middle + small] + [plan + business = company]]	2,176
二千一年	[ABC]+D	ni-sen-ichi-nen	2001 [[two + thousand + one] + year]	2,053
金融機関	[AB]+[CD]	kin-yū-ki-kan	financial institutions [[money + melt] + [mechanism + concern]]	2,051
携帯電話	[AB]+[CD]	kei-tai-den-wa	mobile phones [[carry + belt = mobile] + [electric + talk = phone]]	2,025
人間関係	[AB]+[CD]	nin-gen-kan-kei	human relations [[person + interval = human] + [connect + connect]]	1,907
二千二年	[ABC]+D	ni-sen-ni-nen	2002 [[two + thousand + two] + year]	1,861
一生懸命	[AB]+[CD]	is-shō-ken-mei	with utmost effort [[one + life = lifetime] + [depend + fate = effort]]	1,861

Table 11 presents the 20 most frequent 4KCWs by token frequency counts. Comparing Table 11 with Table 3, which present the 20 most frequent 3KCWs, might initially appear to somewhat undermine the claim advanced earlier that morphological structures are independent of word frequencies. Even though Table 10 clearly indicates that the [AB]+[CD] structure is the highly dominant one for 4KCWs, at 85.3% of all types, 12 of the top 20 4KCWs have [ABC]+D structures and only seven have [AB]+[CD] structures. However, it should be noted that ten of those [ABC]+D 4KCWs are referring to dates of the month, such as 三十一日 /san-jū-ichi-nichi/ *the thirty-first; 31 days*, where the ABC kanji represents 31 and the D-addition represents day, with the other two [ABC]+D 4KCWs being year designations (e.g., 二千一年 2001). As such, their occurrences within the top 20 4KCWs should be attributed to the relative frequency levels of these compound word lemmas within the BCCWJ-based CWLs that are the basis for the 3KWC and 4KCW database lists. That is, while these particular 4KCWs are of high frequencies among the 4KCW database list, as noted earlier, the 4KCWs are generally of lower frequencies compared to the 3KCWs. It is also germane in this context to note that, although 12.0% of the 3KCWs are combinations of number kanji with various numerical units and classifiers, only 1,134 (4.9%) of the 4KCW list are of such combinations, with another 332 (1.4%) 4KCW that are only numbers. Of the 1,134 4KCWs that are combinations of a number and a unit or classifier, understandably, 991 (87.4%) of those are [ABC]+D structures.[17]

4.2. Primary Morphological Structure of [AB]+[CD] 4KCWs

As the summary results in Table 10 incontestably indicate, the primary morphological structure of 4KCWs is [AB]+[CD], where two 2KCWs are combined into a larger compound unit. Notwithstanding Kobayashi et al.'s (2016, p. 117) observation that the semantic head of most [AB]+[CD] 4KCWs is on the right-side (i.e., the CD-component), with some possessing dual heads, Table 12 presents the top 13 most frequent AB-components in terms of their type counts and Table 13 presents the most

[17] It bears repeating that the analyzed list of 4KCW represents only 7.75% of all 4KCWs within the CWLs, while the analyzed list of 3KCWs represents 13.5% of all 3KCWs. Thus, it is highly probably that many more 4KCWs exist that are combinations of numbers and numerical units, but which are of lower frequencies (lemma frequencies > 10). It also bears noting that compound words that consist of three number kanji and a numerical unit/classifier (e.g., 三十一 + 日) are likely to far less frequent in occurrence compared to both a single number kanji and classifier (i.e., 2KCWs such as 一回 /ik-kai/ *one-time* [one + time]) and two number kanji and classifier (i.e., 3KCWs such as 三十分 30 minutes).

frequent 4KCWs, in terms of their token counts, for each of the most frequent AB-components.

TABLE 12. Top 13 most frequent AB-components of [AB]+[CD] 4KCWs according to type counts

AB	Gloss	Translation and explanation	Type count
当該	tō-gai	appropriate; relevant [appropriate + above-stated]	112
経済	kei-zai	economic; finance [pass thru; expire + settle (debt, etc.)]	88
自己	ji-ko	self; oneself [oneself + self]	82
生活	sei-katsu	living; life [life + lively]	79
国際	koku-sai	international [country + occasion; side]	78
社会	sha-kai	society, community [association + meeting; association]	76
一般	ip-pan	general, typical [one + general]	67
経営	kei-ei	business, management [pass thru; expire + occupation]	66
基本	ki-hon	fundamental, basic [foundation + base]	61
教育	kyō-iku	education; instruction [teach + raise]	58
政治	sei-ji	politics; government [politics + reign; rule]	58
生産	sei-san	production; manufacture [life; birth + product; yield]	58
地域	chi-iki	region [earth + region]	58

Highly consistent with the large-scale BCCWJ corpus from which the 4KCW database list has been derived, the most frequent AB-components are related to general areas of human activity, such as 経済 /kei-zai/ *economics*, 社会 /sha-kai/ *society*, 経営 /kei-ei/ *business* and 政治 /sei-ji/ *politics*. The most productive AB-component is the adjective 当該, which appears as the AB-component of 112 4KCWs within the database list, such as in 当該各号 /tō-gai-kaku-gō/ *relevant items* [[relevant + above-stated] + [each + item]] and 当該年度 /tō-gai-nen-do/ *relevant year(s)* [[relevant + above-stated] + [year + time]]. Apart from 当該, the other 12 most frequent AB-compounds are either nouns or VNs. For example, the second most frequent AB-component is the noun 経済, which appears as the AB-component of 88 4KCWs, such as 経済成長 /kei-zai-sei-chō/ *economic growth* [[expire + settle] + [become + long]] and 経済発展 /kei-zai-hat-ten/ *economic development* [[expire + settle] + [start from + unfold]]. The fourth ranked AB-component of 生活 is a VN, which appears as the AB-component of 79 4KCWs, such as 生活環境 /sei-katsu-kan-kyō/ *living environment* [[life + lively] + [ring + boundary]] and 生活習慣 /sei-katsu-shū-kan/ *lifestyle; living habits* [[life + lively] + [learn + accustomed to]].

Table 13. Most frequent [AB]+[CD] 4KCWs by token counts, for each of the most frequent AB-components

[AB]+[CD]	Gloss	Translation and explanation	Type count
当該各号	tō-gai-kaku-gō	relevant items [[relevant + above-stated] + [each + item]]	214
経済成長	kei-zai-sei-chō	economic growth [[expire + settle] + [become + long]]	689
自己責任	ji-ko-seki-nin	self-responsibility [[oneself + self] + [condemn + duty]]	356
生活環境	sei-katsu-kan-kyō	living environment [[life + lively] + [ring + boundary]]	822
国際社会	koku-sai-sha-kai	international society [[country + side] + [company + meet]]	786
社会主義	sha-kai-shu-gi	socialism [[association + meeting] + [main + meaning]]	563
一般会計	ip-pan-kai-kei	general accounting [[one + general] + [meeting + measure]]	473
経営戦略	kei-ei-sen-ryaku	management strategy [[expire + work] + [battle + outline]]	199
基本方針	ki-hon-hō-shin	basic policy [[foundation + base] + [direction + needle]]	839
教育訓練	kyō-iku-kun-ren	education + training [[teach + raise] + [instruct + practice]]	380
政治活動	sei-ji-katsu-dō	political activity [[politics + rule] + [lively + move]]	198
生産活動	sei-san-katsu-dō	production activity [[life + product] + [lively + move]]	281
地域社会	chi-iki-sha-kai	regional community [[earth + region] + [company + meet]]	1,007

Kobayashi, Yamashita, and Kageyama (2016, pp. 116–117) comment that nearly all [AB]+[CD] 4KCWs function as either nouns, VNs or ANs, as some of their examples in (18) illustrate.

(18) [AB]+[CD] nouns

財務大臣 /zai-mu-dai-jin/ *Finance Minister*
[[money + duties] + [big + retainer]]
土地家屋 /to-chi-ka-oku/ *land and buildings*
[[soil + earth] + [house + roof]]

[AB]+[CD] VN

大学改革 /dai-gaku-kai-kaku/ *university reform*
[[big + learn] + [modify + reform]]
意気消沈 /i-ki-shō-chin/ *depressed in spirits*
[[mind + spirit] + [extinguish + sink]]

[AB]+[CD] AN

利用可能	/ri-yō-ka-nō/ *usable* [[benefit + use] + [can + ability]]
単純明快	/tan-jun-mei-kai/ *simple and clear* [[simple + pure] + [bright + pleasant]]

However, of the seven 4KCWs with [AB]+[CD] structures among the 20 most frequent by token frequency counts (Table 11), all are nouns apart from the one AN of 一生懸命 /is-shō-ken-mei/ *with utmost effort* [[one + life] + [depend + fate]]. Moreover, of the 4KCWs for each of the most frequent AB-components (Table 13), most are nouns, with just the three VNs of 教育訓練, 政治活動, and 生産活動.

Turning next to the CD-components of [AB]+[CD] 4KCWs, Table 14 presents the top ten most frequent CD-components in terms of their type counts and Table 15 presents the most frequent 4KCWs, in terms of their token counts, for each of the most frequent CD-components. Also highly consistent with the nature of corpora lexicons, the most frequent CD-components are also closely related to human activities. However, in contrast to the domain connotations of the AB-components, the most frequency CD-components by type counts primarily pertain to the notions of 関係 /kan-kei/ *relations*, 活動 /katsu-dō/ *activities*, 時間 /ji-kan/ *time* and 期間 /ki-kan/ *periods*, and 方法 /hō-hō/ *methods*, as well as 問題 /mon-dai/ *problems* and their 状況 /jō-kyō/ *situations* and 状態 /jō-tai/ *states*.

TABLE 14. Top 10 most frequent CD-components of [AB]+[CD] 4KCWs according to type counts

CD	Gloss	Translation and explanation	Type count
関係	kan-kei	relation; connection [connection + connection]	164
活動	katsu-dō	activity; action [lively + move]	156
以上	i-jō	… and upwards; beyond … [by means of + up]	154
時間	ji-kan	time; period [time + interval]	143
方法	hō-hō	method; process [way + method]	133
期間	ki-kan	period; term [period + interval]	124
主義	shu-gi	doctrine; -ism [main + meaning]	118
問題	mon-dai	problem; issue [ask + topic]	118
状況	jō-kyō	situation; circumstances [state + situation]	112
状態	jō-tai	state; condition [state + condition]	103

The most productive CD-component is the noun 関係, which appears as the CD-component of 164 [AB]+[CD] 4KCWs within the database list, such as in 人間関係 /nin-gen-kan-kei/ *human relations* [[human + space] + [connect + connect]] and 信頼関係 /shin-rai-kan-kei/ *relationship of mutual trust* [[faith + trust] + [connect + connect]]. The second most frequent

TABLE 15. Most frequent [AB]+[CD] 4KCWs by token counts, for each of the most frequent CD-components

[AB]+[CD]	Gloss	Translation and explanation	Type count
人間関係	nin-gen-kan-kei	human relations [[human + space] + [connect + connect]]	1,861
経済活動	kei-zai-katsu-dō	economic activity [[expire + settle] + [lively + move]]	519
必要以上	hitsu-yō-i-jō	more than necessary [[certain + need] + [by means of + up]]	504
労働時間	rō-dō-ji-kan	working hours [[labor + work] + [time + interval]]	790
応募方法	ō-bo-hō-hō	application method [[apply + recruit] + [way + method]]	292
一定期間	it-tei-ki-kan	fixed interval [[one + determine] + [period + interval]]	314
民主主義	min-shu-shu-gi	democracy [[people + main] + [main + meaning]]	1,102
環境問題	kan-kyō-mon-dai	environmental problem [[ring + boundary] + [ask + topic]]	900
実施状況	jis-shi-jō-kyō	implementation status [[real + perform] + [state + situation]]	326
健康状態	ken-kō-jō-tai	health condition [[healthy + ease] + [state + condition]]	326

CD-component is the VN of 活動, which is the only VN amongst the top ten CD-components. It is the CD-component of 156 4KCWs, such as 経済活動 /kei-zai-katsu-dō/ *economic activity* [[expire + settle] + [lively + move]] and 事業活動 /ji-gyō-katsu-dō/ *business activities* [[matter + business] + [lively + move]].

4.3. Secondary Morphological Structure of [ABC]+D 4KCWs

Reflecting the greater dominance of the primary [AB]+[CD] morphological structure for 4KCWs, the secondary structure of [ABC]+D only accounts for 12.1% of the 4KCW database list. Moreover, although this secondary structure closely parallels the primary [AB]+C morphological structure of 3KCWs, as noted earlier, where an additional morpheme is being attached to the end of an existing compound word, its coverage of only 12.1% stands in sharp contrast to the 77.1% prevalence of [AB]+C 3KCWs as the primary structure of 3KCWs. Moreover, further analyses of the D-additions of 4KCWs reveals that 26% are suffixes, which account for account for 61% of the [ABC]+D structures.

Table 16 presents the top ten most frequent D-additions to [ABC]+D 4KCWs by type counts and Table 17 presents the most frequent [ABC]+D 4KCWs, by token counts, for each of the most frequent D-additions.

TABLE 16. Top ten most frequent D-additions to [ABC]+D 4KCWs by type counts

D-addition	Meaning	Type count
等	etc.; and so forth	156
円	Japanese yen	152
人	*-er* person-indicating ending	147
条	article, clause, counter for articles	116
年	year	109
的	*-ic* AN ending	109
者	*-er* person-indicating ending	95
歳	age counter	76
間	between; interval	71
達	pluralizing ending	70

TABLE 17. Most frequent [ABC]+D 4KCWs by token counts, for each of the most frequent D-additions

[ABC]+D]	Gloss	Translation and explanation	Type count
高齢者等	kō-rei-sha-tō	such as the elderly [[high + age + person] + pluralizer]	93
千五百円	sen-go-hyaku-en	1,500 yen [[thousand + five + hundred] + yen]	691
被相続人	hi-sō-zoku-nin	decedent [[cover + together + continue] + person]	297
第十二条	dai-jū-ni-jō	article 12 [[number + ten + two] + article]	636
二千一年	ni-sen-ichi-nen	2001 [[two + thousand + one] + year]	2,053
中長期的	chū-chō-ki-teki	mid-long term-ish [[middle + long + period] + -ic]	229
被保険者	hi-ho-ken-sha	insured person [[cover + protect + precipitous] + person]	1,013
二十四歳	ni-jū-yon-sai	24 years old [[two + ten + four] + years of age]	597
二十年間	ni-jū-nen-kan	20 year period [[two + ten + year] + interval]	381
主人公達	shu-jin-kō-tachi	protagonists [[main + person + public] + pluralizer]	15

In light of the clear parallels in terms of word-formation processes, it is most expedient to first compare the most frequent C-additions of [AB]+C 3KCWs (Table 6) with the most frequent D-additions of [ABC]+D 4KCWs (Table 16). While such comparisons reveal that five morphemes are common to both lists (i.e., 的, 者, 等, 達, and 人), clearly, there are also differences in terms of their respective rankings. For instance, 的 is the most frequent C-addition, occurring in 873 [AB]+C 3KCWs, but it is only the sixth most frequent as a D-addition, occurring in 109 [ABC]+D 4KCWs, such as 中長期的 /chū-chō-ki-teki/ *mid-long term-ish* [[middle + long + period] + -ic]. However, in demonstrating that this morpheme attaches to both many 3KCWs and many 4KCWs, these results are highly consistent with Kobayashi et al.'s (2016, p. 127) observation, noted earlier, that 的 is a highly productive SJ affix that attaches to various bases. The most frequent D-addition is 等 which occurs in 156 4CKWs, such as 高齢者等 /kō-rei-sha-tō/ *such as the elderly* [[high + age + person] + pluralizer], while it is the third most frequent C-addition, occurring in 577 3KCWs. The largest shift in the respective frequency rankings for 3KCWs and 4KCWS is for 人, which is the third most frequent D-addition, occurring in 147 4KCWs, such as 被相続人 /hi-sō-zoku-nin/ *decedent* [[cover + together + continue] + person], as opposed to being the tenth most frequent C-addition, occurring in 227 3KCWs.

Comparing Tables 6 and 16 also reveals that five SJ morphemes are not common to both lists, but, highly congruent with earlier remarks about the likely frequency distributions of number kanji, these D-additions attach either solely or commonly to number kanji. Accordingly, it is not surprising to discover that the most frequent D-addition is 円 /en/ *Japanese yen currency*, which is a D-addition to 152 4KCWs, such as 千五百円 /sen-go-hyaku-en/ *1,500 yen* [[thousand + five + hundred] + yen] and of even larger sums, such as 五十万円 /go-jū-man-en/ *500,000 yen* [[five + ten + ten-thousand] + yen]. The fourth most frequent D-addition is 条 /jō/ *article, clause, counter for articles*, which occurs in 116 4KCWs, such as 第十二条 /dai-jū-ni-jō/ *article 12* [[number + ten + two] + article]. The fifth most frequent D-addition is 年 /nen/ *year*, which occurs in 109 4KCWs, such as 二千一年 /ni-sen-ichi-nen/ *2001* [[two + thousand + one] + year], while the eighth most frequent is 歳 /sai/ *age counter*, which occurs in 76 4KCWs, such as 二十四歳 /ni-jū-yon-sai/ *24 years old* [[two + ten + four] + years of age]. Although the ninth most frequent D-addition of 間 /kan/ *between; interval* also often combines with 3KCWs that involve numbers, such as 二十年間 /ni-jū-nen-kan/ *20 year period* [[two + ten + year] + interval], in such cases the C of the 3KCW invariably represents some time unit (such as minutes, days, months, and years). It can also attach to other kinds of 3KCWs, where the notion of between is spatial, such as 加盟国間 /ka-mei-koku-kan/ *between member states* [[add + alliance + country] + between].

4.4. Tertiary Morphological Structure of A+[BCD] 4KCWs

As with the secondary structure of 4KCWs, the tertiary structure of A+[BCD] has also been considerably marginalized to just 1.9% of all 4KCWs structures, due to the marked prevalence of the 4KCW primary structure. However, again, the parallels to the morphological structures of 3KCWs are present to the extent that the tertiary structure of A+[BCD] 4KCWs is similar to the secondary A+[BC] structure of 3KCWs, where an additional morpheme is being inserted at the beginning. Moreover, the tendency seen with 3KCWs to derive longer compounds by appending a final morpheme as opposed to inserting an initial morpheme is also observed for the 4KCWs. As with the secondary structure of [ABC]+D 4KCWs, further analysis of the A-additions reveals that 32% are prefixes, which account for 75% of the A+[BCD] structures.

Table 18 presents the top ten most frequent A-additions to A+[BCD] 4KCWs by type counts and Table 19 presents the most frequent A+[BCD] 4KCWs, by token counts, for each of the most frequent A-additions.

Table 18. Top ten most frequent A-additions to A+[BCD] 4KCWs by type counts

A-addition	Meaning	Type count
約	approximately	84
各	each; every	46
総	gross, whole, general	24
同	same	22
新	new	16
全	all, whole	16
非	negation prefix	16
大	large; big	14
翌	the following; next	12
副	vice-; assistant	11

Also reflecting the close parallels in terms of word formation, there is again merit in comparing the most frequent A-additions for 3KCWs (Table 8) with the most frequent A-additions of 4KCWs (Table 18). Five morphemes are common to both lists (i.e., 大, 各, 新, 同, and 全), but the shifts in their respective ranking orders are generally not as pronounced as the shifts between the C-additions and D-additions to 3KCWs and 4KCWs, respectively. However, in sharp contrast to 御 /o/ and /go/ *honorific prefix* being the most frequent A-addition for 3KCWs in terms of type counts, in the case of A+[BCD] 4KCWs, the most frequent A-addition is 約 /yaku/ *approximately,* which is an A-addition to 84 4KCWs, even though it is not amongst the top ten as an A-addition to 3KCWs.

TABLE 19. Most frequent A+[BCD] 4KCWs by token counts, for each of the most frequent A-additions

A+[BCD]	Gloss	Translation and explanation	Type count
約三十分	yaku-san-jip-pun	about 30 minutes [about + three + ten + minutes]	123
各市町村	kaku-shi-chō-son	each municipality [each + city + town + village]	113
総司令部	sō-shi-rei-bu	headquarters [general + official + orders + section]	134
同委員会	dō-i-in-kai	same committee [same + committee + member + meet]	116
新事業者	shin-ji-gyō-sha	new business person [new + thing + business + person]	94
全十二回	zen-jū-ni-kai	twelve times in total [all + ten + two + times]	37
非製造業	hi-sei-zō-gyō	nonmanufacturing sector [un + make + create + business]	115
大真面目	ō-majime	deadly serious [big + true + face + eye]	187
翌営業日	yoku-ei-gyō-bi	next working day [next + conduct + business + day]	23
副大統領	fuku-dai-tō-ryō	vice-president [vice + big + govern + territory]	67

As in both 約三十分 /yaku-san-jip-pun/ *about 30 minutes* [about + [three + ten + minutes]] and 約二百人 /yaku-ni-hyaku-nin/ *about 200 people* [about + [two + hundred + people]], 約 is typically inserted at the beginning of 3KCWs with [AB]+C structures, where the AB morphemes are numbers and the C-component is a numerical unit or classifier, such as minutes, people, and Japanese yen.

The second most frequent A-addition for 4KCWs is 各 /kaku/ *each; every*, which occurs in 46 4KCWs, such as in 各市町村 /kaku-shi-chō-son/ *each municipality* [each + [city + town + village]] and 各自治体 /kaku-ji-chi-tai/ *each municipality* [each + [[self + rule + body]]. Its ranking as the second most frequent A-addition is comparable to its ranking as the third most frequent A-addition for 3KCWs, which underscores the general productivity of this SJ morpheme as a prefix of both 3KCW and 4KCWs. Although not appearing within the top ten A-additions for 3KCWs, the third most frequent for 4KCWs is 総 /sō/ *gross, whole, general*, which occurs in 24 4KCWs, such as 総司令部 /sō-shi-rei-bu/ *headquarters* [general + [official + orders + section]] and 総事業費 /sō-ji-gyō-hi/ *total operating expenses* [gross + [matter + business + expenses]].

Of the four A-additions that function as negative prefixes (Kobayashi, Yamashita, and Kageyama, 2016), as noted earlier, only 非 /hi/ *negation* features within the top ten most frequent A-additions for 4KCWs, even though it was not amongst the top ten for 3KCWs. It occurs in 16 4KCWs, such as 非製造業 /hi-sei-zō-gyō/ *nonmanufacturing sector* [un + [make + create + business]] and 非喫煙者 /hi-kitsu-en-sha/ *non-smoker* [non + [consume + smoke + person]].

4.5. Other 4KCW Morphological Structures

Our analysis of the morphological structures of the 4KCW database reveals that a large majority (85.3%) have [AB]+[CD] structures, being the combination of two 2KCWs. The secondary and tertiary structures of 4KCWs involve one morpheme being added to an existing 3KCW, which together account for 14.0% of 4KCws However, as Table 10 also indicates, eight other morphological structures underlie a small percentage of 4KCWs. Accordingly, this section turns to present examples of those 4KCW structures.

For the 4KCWs, the first of these more marginal morphological structures is non-divisible (0.1%), which, as with the 3KCWs, is necessary to handle a small set of exceptions. The examples provided in (19) also illustrates that although the compound word's etymology and morphological structure are not clear, the meanings of the component morphemes are often related to the overall meaning.

(19) Non-divisible

炭水化物	/tan-sui-ka-butsu/ *carbohydrate* [coal + water + change + matter]
不可思議	/fu-ka-shi-gi/ *mystery; unfathomable* [negative + can + think + debate]

The second of the marginal morphological structures is [A(CD)*]+[BCD] (0.1%), where the CD-component of an [ACD] 2KCW is omitted and the resultant A is attached to a related [BCD] 3KCW. This is also a form of clipping, as noted earlier, and, once again, this structure may appear to resemble superficially the A+[BCD] structure outlined above, to the extent that an A-component is being inserted before a [BCD] 3KCW. However, as with the [A(C*)]+[BC] structure of 3KCWs, the [A(CD)*]+[BCD] structure crucially hinges on the semantic relationship between the [ACD] and [BCD] 3KCWs, due to their shared CD-components, as the examples in (20) illustrates.

(20) [A(CD*)]+[BCD] (with (*CD) omitted)

小中学生	/shō-chū-gaku-sei/ *elementary and junior-high school students* [小 of 小学生 elementary school student + [中学生 junior-high school student]]
土日曜日	/do-nichi-yō-bi/ *Saturday and Sunday* [土 of 土曜日 Saturday + [日曜日 Sunday]]s

The third of the marginal morphological structures is [A(D*)]+[B(D*)]+[CD] (0.1%), where (D*) is omitted from both an [A(D*)] and a [B(D*)] 2KCW and the resultant A and B morphemes are inserted at the beginning of a CD 2KCW. As yet another example of a compound word formation that involves clipping, this structure also attests to the fact that the clipping process is a commonplace phenomenon. It is also essential to carefully differentiate this [A(D*)]+[B(D*)]+[CD] structure from the primary 4KCW structure of [AB]+[CD]. Although it may again potentially appear as if an [AB] 2KCW is being combined with a [CD] 2KCW, crucially, the A and B morphemes that are being inserted before the [CD] 2KCW here do not occur together as a 2KCW. This morphological structure also depends on the semantic connections between three 2KCWs due to the D component that is shared by all, as the examples in (21) highlight.

(21) [A(D*)]+[B(D*)]+[CD] (with both (*D) omitted)
陸海空軍 /riku-kai-kū-gun/ *land, sea and air forces*
[land + sea + air + troops] [陸 of 陸軍 land forces + 海 of 海軍 navy + [空軍 air force]]
農林漁業 /nō-rin-gyo-gyō/ *agriculture, forestry and fishing*
[farm + forest + fish + industry] [農 of 農業 agriculture + 林 of 林業 forestry + [漁業 fishing industry]]

The fourth of the marginal morphological structures is A+B+C+D (0.1%), as the concatenation of four morphemes that together constitute a set of things, with the examples in (22) being prototypical.

(22) A+B+C+D
春夏秋冬 /shun-ka-shū-tō/ *four seasons*
[spring + summer + autumn + winter]
喜怒哀楽 /ki-do-ai-raku/ *human emotions*
[joy + anger + grief + pleasure]

The fifth of the marginal morphological structures is phonological transcription (当て字) (0.1%). As explained earlier for the 3KCWs structures, there are also 4KCWs where the kanji are being used conventionally to represent the word's syllables, as in (23).

(23) Phonological transcriptions (当て字)
滅茶滅茶 /me-cha-me-cha/ *disorderly, absurd; excessive*
[destroy + tea + destroy + tea]
無理矢理 /mu-ri-ya-ri/ *forcibly; against one's will*
[nothing + reason + arrow + reason]

The sixth of the marginal morphological structures for 4KCWs is [AB]+C+D (0.0%), where C and D morphemes are being attached to an [AB] 2KCW. This structure should also be distinguished from the primary morphological structure of [AB]+[CD], because, as with the A and

B morphemes inserted before a CD 2KCW in the [A(D*)]+[B(D*)]+[CD] structure, the C and D morphemes that are attached to form [AB]+C+D structures do not occur together as an independent 2KCW. The AB components of the 4KCWs that conform to this structure are number kanji and the C morpheme is 箇 /ka/ *counter for articles*, as the examples in (24).

(24) [AB]+C+D

十二箇月 /jū-ni-ka-getsu/ *12 month (period)*
[[ten + two] + counter + month]
十一箇国 /jū-ichi-ka-koku/ *11 countries*
[[ten + one] + counter + country]

The seventh of the marginal morphological structures is monomorphemic words (熟字訓) (0.0%). Again, in contrast to phonological transcriptions, although there is usually no phonological correspondence between the elements of the graphematic representation and the compound word pronunciation, the meanings of the component kanji usually relate to the word's overall meaning, which is the case with the example in (25).

(25) Monomorphemic words (熟字訓)

再従兄弟 /haitoko/ *second cousin*
[again + accompany + elder brother + younger brother]

The eighth and final of the marginal morphological structures for 4KCWs is [A(D*)]+[BCD] (0.0%), where the D-component of an AD 2KCW is omitted and inserted at the beginning of an BCD 3KCW. It is also important to distinguish this structure from both the secondary structure of A+[BCD] and from the second of the more marginal structures of [A(CD)*]+[BCD] with the (CD) element of the A(CD) 3KCW omitted. As with the second of the marginal structures, the distinction is well motivated based on the semantic connection between the D-component of the [A(D)*] 2KCW and the D component of the [BCD] 3KCW, as the example in (26) illustrate.

(26) [A(D*)]+[BCD] (with (*D) omitted)

産婦人科 /san-fu-jin-ka/ *maternity and gynaecology*
[産 of 産科 obstetrics + [婦人科 gynaecology]]

5. Concluding Remarks

This paper has outlined the construction of two new databases of 3KCWs and 4KCWs, as key components for a larger database project concerned with Japanese lexical properties (Joyce, Hodošček, and Masuda, 2017; Joyce, Masuda, and Ogawa, 2014). More specifically, this paper has focused on describing the results of analysing the extracted database lists according to the morphological structures that underlie the 3KCWs and 4KCWs, respectively. The results provide tangible quantitative indications of the degrees to which the dominant morphological structures differ for 3KCWs and 4KCWs (Kageyama and Saito, 2016; Kobayashi, Yamashita, and Kageyama, 2016; Shibatani, 1990; Tamamura, 1984; 1985).

In the case of the 3KCW database, although eight structures were identified in total, the analysis results clearly show that just two morphological structures underlie the vast majority (98.4%) of the 23,046 3KCWs. Moreover, although both structures involve adding a morpheme to an existing 2KCW, the results also reveal a striking preference for attaching an morpheme to the end of an existing 2KCW, such that the primary morphological structure of [AB]+C accounts for 77.1% of the 3KCWs. In comparison, the secondary structure of A+[BC], where the additional morpheme is added to the beginning of a 2KCW, only accounts for 21.3% of the 3KCWs. In sharp contrast to the results for the 3KCWs, in the case of the 4KCW database, although 11 structures were identified in total, the analysis results indicate that the dominant morphological structure of 4KCWs is overwhelmingly [AB]+[CD], where two 2KCWs are combined by compounding processes, and which account for 85.3% of the 23,159 4KCWs. Notwithstanding the pervasive nature of the primary structure, still, a considerable number of 4KCWs are formed by adding a morpheme to an existing 3KCW, where a marked preference for attachment to the end of compound words is also observed. Thus, at much reduced proportions compared to 3KCWs, for 4KCWs, the secondary structure is [ABC]+D, which accounts for 12.1%, and the tertiary structure is A+[BCD], which accounts for 1.9%.

Taken together, the results of analyzing the morphological structures of both database lists unquestionably underscore the immense significance of 2KCWs within the Japanese lexicon, not only as words in their own right, but as the component elements of longer compound words (Joyce, 2011; Joyce, Hodošček, and Masuda, 2017; Joyce, Masuda, and Ogawa, 2014; Kobayashi, Yamashita, and Kageyama, 2016; Nomura, 1975; 1988). Overall, these findings are entirely consistent with the morphographic nature of kanji (Joyce, 2011; Kobayashi, Yamashita, and Kageyama, 2016) because they vividly highlight how the concatenation of kanji in graphematically representing the vast majority of Japanese compound words is primarily the province of the morphological processes that underlie the formation of Japanese compound words. That is, while there are undeniably a limited number of exceptions, such as the non-divisible, phonological transcription and monomorphemic structures, the surface graphematic forms of most Japanese compound words conform to the morphographic principle (Joyce, 2011). However, kanji are associated with both NJ and SJ morphemes, many with multiple NJ and SJ allomorphs, and the status of those morphemes—as either free, bound or affixes—is often context-dependent. Accordingly, the present analyses of the rich morphological structures of Japanese compound words can potentially further elucidate the intricate nature of the morphographic principle in the case of the JWS; a topic that undoubtedly warrants greater attention from the perspective of writing systems research.

As indicated earlier, the task of analyzing the morphological structures of 3KCWs and 4KCWs has been greatly facilitated by the fact that, in fundamentally conforming to the morphographic principle, their structures are generally highly transparent. However, as also acknowledged, reflecting the productive nature of SJ morphemes, some compound words are conceivably open to alternative interpretations, such as 農業者 *agricultural worker*. Accordingly, even though the most plausible interpretation of 農業者 is to regard it as an example of the [AB]+C structure, we are also planning to conduct studies to obtain native-speaker rankings related to the psychological validity and credibility of the morphological structures. Such studies will also investigate the extent to which semantic shifts in the meanings of compound words might influence native-speaker interpretations of their morphological structures. For instance, although the meanings of the constituent morphemes are clear and the morphological structure of 新幹線 /shin-kan-sen/ [new + [trunk + line]] is unquestionably A+[BC], the compound word's contemporary meaning of *bullet train* represents a substantial semantic shift.

Moreover, the conducted analyses of the morphological structures of both 3KCWs and 4KCWs are essential for preparing to conduct various visual word recognition experiments to further investigate Kobayashi et al.'s (2016, p. 129) claims that kanji facilitate meaning comprehension, which have significant implications for the organization of morphological information within the mental lexicon. Joyce (2002; 2004) and Masuda and Joyce (2018) have already conducted a series of psycholinguistic experiments that have employed the constituent-priming paradigm to examine lexical-decision task responses to 2KCWs. As those studies have generally observed robust patterns of facilitated reaction times due to the prior presentation of the constituent kanji, across a variety of conditions, including very brief stimulus onset asynchrony intervals, the natural next steps are to extend this experimental approach to investigate the recognition processes of 3KCWs and 4KCWs. To that aim, the analyses of their morphological structures will be invaluable in terms of designing various experimental conditions and selecting suitable stimuli.

As already noted, these 3KCW and 4KCW databases have been compiled as new components of a larger database project concerned with various Japanese lexical properties (Joyce, Hodošček, and Masuda, 2017; Joyce, Masuda, and Ogawa, 2014; Masuda and Joyce, 2005; Masuda, Joyce, et al., 2014). In being extracted from Joyce, Hodošček, and Nishina (2012) CWLs, which were, in turn, extracted from the BCCJW, both database lists have automatically inherited a number of valuable data-fields, such as word class, lexical strata, token frequencies of lemma and orthographic base forms and pronunciation(s). In addition to those inherited data-fields, naturally, the work of analysing the morphological structures has itself generated a number of additional fields beyond just assigning a structure category, such as identifying all possible graphematic overlaps and counts related to morphological family sizes. These will all be checked as the work of integrating these new database components within the larger database progresses and as further database components are developed in the future, such as the planned analyses of the morphological structures of five-kanji compound words.[18] Thus, the present analyses of the morphological structures of both 3KCWs and 4KCWs represent a significance contribution to the larger database project of mapping out various Japanese lexical properties.

References

Agency for Cultural Affairs [文化庁] (2010). "常用漢字表 [Jōyō kanji list]." In: URL: http://kokugo.bunka.go.jp/kokugo_nihongo/joho/kijun/naikaku/pdf/joyokanjihyo_20101130.pdf.

Backhouse, A.E. (1984). "Aspects of the graphological structure of Japanese." In: *Visible Language* 18.3, pp. 219–228.

Igarashi, Yuko (2007). "The changing role of katakana in the Japanese writing system: Processing and pedagogical dimensions for native speakers and foreign learners." PhD thesis. University of Victoria, British Columbia, Canada.

Joyce, Terry (2002). "Constituent-morpheme priming: Implications from the morphology of two-kanji compound words." In: *Japanese Psychological Research* 44, pp. 79–90.

——— (2004). "Modeling the Japanese mental lexicon: Morphological, orthographic and phonological considerations." In: *Advances in Psychological Research: Volume*. Ed. by S.P. Shohov. Vol. 31. Hauppauge, NY: Nova Science, pp. 27–61.

——— (2011). "The significance of the morphographic principle for the classification of writing-systems." In: *Written Language & Literacy* 14.1, pp. 58–81.

Joyce, Terry, Bor Hodošček, and Hisashi Masuda (2017). "Constructing an ontology and database of Japanese lexical properties: Handling the orthographic complexity." In: *Written Language & Literacy* 20.1, pp. 27–51.

[18] Although longer Japanese compound words are frequently attested, as Kobayashi, Yamashita, and Kageyama (2016, pp. 114–115) observe, as longer compound words invariably involve recursive combinations of existing shorter compound words. Thus, the returns from analyzing beyond five-kanji compound words are likely to be rather limited.

Joyce, Terry, Bor Hodošček, and Kikuko Nishina (2012). "Orthographic representation and variation within the Japanese writing system: Some corpus-based observations." In: *Written Language & Literacy* 15.2, pp. 254–278.

Joyce, Terry and Hisashi Masuda (2018). "Introduction to the multi-script Japanese writing system and word processing." In: *Writing systems, reading processes, and cross-linguistic influences: Reflections from the Chinese, Japanese and Korean languages*. Ed. by Hye Pae. Vol. 7. Bilingual Processing and Acquisition. Amsterdam: John Benjamins, pp. 179–199.

——— (2019). "On the notions of graphematic representation and orthography from the perspective of the Japanese writing system." In: *Written Language & Literacy* 22.2, pp. 248–280.

Joyce, Terry, Hisashi Masuda, and Taeko Ogawa (2014). "Jōyō kanji as core building blocks of the Japanese writing system: Some observations from database construction." In: *Written Language & Literacy* 17.2, pp. 173–194.

Kageyama, Taro and Michiaki Saito (2016). "Vocabulary strata and word formation processes." In: *Handbook of Japanese lexicon and word formation*. Ed. by Taro Kageyama and Hideki Kishimoto. Vol. 3. Handbooks of Japanese Language and Linguistics. Boston, Berlin: Walter de Gruyter, pp. 11–50.

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

Kobayashi, Hideki, Kiyo Yamashita, and Taro Kageyama (2016). "Sino-Japanese words." In: *Handbook of Japanese lexicon and word formation*. Ed. by Taro Kageyama and Hideki Kishimoto. Vol. 3. Handbooks of Japanese Language and Linguistics. Boston, Berlin: Walter de Gruyter, pp. 93–131.

Konno, Shinji [今野真二] (2013). 正書法のない日本語 *[The Japanese language lacks orthography]*. 東京 [Tokyo]: 岩波書店 [Iwanami Shoten].

Lurie, David B. (2012). "The development of writing in Japan." In: *The shape of script: How and why writing systems change*. Ed. by S.D. Houston. Santa Fe, NM: School for Advanced Research Press, pp. 159–185.

Maekawa, Kikuo et al. (2013). "Balanced corpus of contemporary written Japanese." In: *Language Resources and Evaluation*, pp. 1–27.

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

——— (2018). "Constituent-priming investigations of the morphological activation of Japanese compound words." In: *Writing systems, reading processes, and cross-linguistic influences: Reflections from the Chinese,* *Japanese and Korean languages*. Ed. by Hye Pae. Amsterdam: John Benjamins, pp. 221–244.

Masuda, Hisashi and Terry Joyce (2019). "A database of three-kanji compound words in Japanese, with particular focus on their morphological structures." Poster presentation given as the *'Diversity of writing systems: Embracing multiple perspectives', 12th International Workshop on Written Language and Literacy*, Faculty of Classics, Cambridge University, UK.

Masuda, Hisashi, Terry Joyce, et al. (2014). "A database of semantic transparency ratings for two-kanji Japanese compound words." Poster presentation given at *'Orthographic Databases and Lexicons': 9th International Workshop on Writing Systems and Literacy*, University of Sussex, Brighton, UK.

Miller, Laura (2011). "Subversive script and novel graphs in Japanese girls' culture." In: *Language & Communication* 31.1, pp. 16–26.

Nomura, Masaaki [野村雅昭] (1975). "四字漢語の構造 [The structure of four-kanji Sino-Japanese words]." In: 電子計算機による国語研究 *[Studies in Computational Linguistics]* 7, pp. 36–80.

——— (1988). "二字漢語の構造 [The structure of two-kanji Sino-Japanese words]." In: 日本語学 *[Japanese Studies]* 7.5, pp. 44–55.

Robertson, Wesley C. (2015). "Orthography, foreigners, and fluency: Indexicality and script selection in Japanese manga." In: *Japanese Studies* 35.2, pp. 205–222.

——— (2017). "He's more katakana than kanji: Indexing identity and self-presentation through script selection in Japanese manga (comics)." In: *Journal of Sociolinguistics* 21.4, pp. 497–520.

Shibatani, Masayoshi (1990). *The Languages of Japan*. Cambridge, UK: Cambridge University Press.

Shinmura, Izuru [新村出], ed. (1995). 広辞苑 *[Japanese dictionary]*. 5th ed. 東京 [Tokyo]: 岩波書店 [Iwanami Shoten].

——— ed. (2008). 広辞苑 *[Japanese dictionary]*. 6th ed. 東京 [Tokyo]: 岩波書店 [Iwanami Shoten].

Smith, Janet S. (Shibamoto) (1996). "Japanese writing." In: *The world's writing systems*. Ed. by Peter T. Daniels and William Bright. New York: Oxford University Press, pp. 209–217.

Smith, Janet S. (Shibamoto) and David L. Schmidt (1996). "Variability in written Japanese: Towards a sociolinguistics of script choice." In: *Visible Language* 30.1, pp. 47–71.

Tamamura, Fumio [玉村文郎] (1984). 日本語教育指導参考書*12:* 語彙の研究と教育（上）*[Japanese language education reference guides 12: Lexical research and education 1]]*. 東京 [Tokyo]: 大蔵省印刷局 [Ookurashou Insatsukyoku].

——— (1985). 日本語教育指導参考書*13:* 語彙の研究と教育（下）*[Japanese language education reference guides 13: Lexical research and education 2]]*. 東京 [Tokyo]: 大蔵省印刷局 [Ookurashou Insatsukyoku].

Taylor, Insup and M. Martin Taylor (2014). *Writing and literacy in Chinese, Korean and Japanese*. Vol. 14. Studies in Written Language and Literacy. Amsterdam: John Benjamins.

Tranter, Nicolas (2008). “Nonconventional script choice in Japan.” In: *International Journal of the Sociology of Language* 192, pp. 133–151.