Psychological Research (2017) 81:696–708
DOI 10.1007/s00426-016-0764-3

ORIGINAL ARTICLE

www.kanjidatabase.com: a new interactive online database for psychological and linguistic research on Japanese kanji and their compound words

Katsuo Tamaoka[1] · **Shogo Makioka**[2] · **Sander Sanders**[3] · **Rinus G. Verdonschot**[4]

Received: 21 July 2015 / Accepted: 24 February 2016 / Published online: 16 March 2016

Abstract Most experimental research making use of the Japanese language has involved the 1945 officially standardized kanji (Japanese logographic characters) in the Jōyō kanji list (originally announced by the Japanese government in 1981). However, this list was extensively modified in 2010: five kanji were removed and 196 kanji were added; the latest revision of the list now has a total of 2136 kanji. Using an up-to-date corpus consisting of 11 years' worth of articles printed in the *Mainichi Newspaper* (2000–2010), we have constructed two novel databases that can be used in psychological research using the Japanese language: (1) a database containing a wide variety of properties on the latest 2136 Jōyō kanji, and (2) a novel database containing 27,950 two-kanji compound words (or jukugo). Based on these two databases, we have created an interactive website (www.kanjidatabase.com) to retrieve and store linguistic information to be used in psychological and linguistic experiments. The present paper reports the most important characteristics for the new databases, as well as their value for experimental psychological and linguistic research.

✉ Rinus G. Verdonschot
rinusverdonschot@gmail.com

[1] Graduate School of Languages and Cultures, Nagoya University, Nagoya, Japan

[2] Osaka Prefecture University, Sakai, Japan

[3] Kumulus Centre, Maastricht, The Netherlands

[4] Waseda Institute for Advanced Study (WIAS), Waseda University, 1-6-1 Nishi Waseda, Shinjuku-ku, Tokyo 169-8050, Japan

Introduction

Logographic scripts (used in Japanese and Chinese) have many properties making them attractive to investigate matters that would be difficult to research using alphabetic scripts. For example, logographs are able to convey a particular meaning directly without involving phonology. To date, there has been a long tradition in psychological research that involved the usage of logographic scripts on a variety of topics, such as: dyslexia (e.g. Frith, 1981), word reading (e.g. Yamada, Mitarai, & Yoshida, 1991), sentence processing (e.g. Wang, Verdonschot, & Yang, 2016), memory (e.g. Chen, Cheung, Lau, 1997; Le Bigot, Passerault, & Olive, 2009), stroop tasks (e.g. Luo, & Proctor, 2013), visual perception (e.g. Ono, & Kawahara, 2008) and perception–action coupling (e.g. Yu, Gong, Qiu, & Zhou, 2011). In most research, when constructing experimental materials, important information needs to be collected, for instance to make sure that experimental conditions are balanced or certain requirements are met.

To aid this process, the current paper reports the availability of two novel (and freely available) databases which provide fellow–researchers, especially those who plan to use the Japanese language in their experiments, with an easy-to-access means to control and manipulate important lexical factors to conduct high-quality research. The main highlights of these new databases are that they are based on a recent corpus and reflect the new Jōyō-kanji list that was introduced by the Japanese government in 2010. Additionally, we provide an online tool that allows for an easy lookup and usage of these databases (www.kanjidatabase.com). This paper is organized in the following way: first we give a brief overview on Japanese kanji characters and present some of the currently available databases, then we introduce the new databases, point out how each of the

variables in the current database were calculated, whilst specifying how these particular measures are important in experimental psychological and linguistic research.

What are Japanese kanji?

Kanji are logographic characters (originating from Chinese) that have been adopted over time into the Japanese language. In modern Japanese texts, kanji are typically combined with two other scripts (hiragana and katakana) to form the Japanese writing system. Although there are many content words that are preferably written in hiragana or katakana, typically kanji characters are used to convey the basic meaning of an utterance. For example, in a sentence such as "彼はその理論について論文を書いた" meaning "he wrote an article on the theory" the kanji in the sentence (i.e., 彼 "he", 理論 "theory", 論文 "article" and 書 "to write") contain the meaning and the other script (only hiragana in this case) plays a grammatical role (case/tense marking). When two kanji combine to form words such as 理論 "theory" and 論文 "article" they are called jukugo (meaning two-kanji compound words).

Although the largest Japanese dictionaries list the existence of over 50,000 kanji (e.g. Morohashi, 2000), it is generally thought that about 4000 kanji are used in daily life. To provide an official standard for printed texts, the Japanese government established a list of commonly-used kanji in 1981 which contained 1945 basic Japanese kanji characters (called the Jōyō kanji-list). Since then this list has been used as the standard for Japanese texts including newspapers, magazines, educational materials and research. In 2010, however, the official Jōyō kanji list was extensively revised. Five kanji were removed, and 196 kanji were added (±9 % change). The creation of this new list represents the official adjustment for changed kanji usage over time in modern Japanese and now includes a total of 2136 kanji to serve as the foundation for Japanese written texts. Obviously, this has consequences when selecting stimuli for experimental psychological and linguistic research making use of the Japanese language.

Existing Japanese kanji databases

Various papers report to have employed the (now sold out) kanji database developed by Amano and Kondo (1999, 2000), otherwise known as the NTT database (which costed about 750 USD). This database was created based on a corpus of Asahi Newspaper articles, published between 1985 and 1998. Although it is well constructed, it is likely that databases based on more recent corpora (such as the current Mainichi Shimbun corpus) are more representative of modern Japanese. In addition, the NTT database employed morphological parsing software called sumomo, which, according to the the authors, elicits segmentation parsing error rates around 10 % (Amano, & Kondo, 2000). In comparison, the new database(s) reported in this paper utilized a newer morphological parsing program called MeCab (Kudo, Yamamoto, & Matsumoto 2004) with a segmentation accuracy around 98.96 % (or an error rate of 1.04 %).[1]

Another regularly used database (see e.g. Jincho, Feng, & Mazuka, 2014) is the Balanced Corpus of Contemporary Written Japanese (BCCWJ)[2] by Maekawa et al., (2014) which was generated from a corpus covering a wide range of materials including books, magazines and newspapers. Although there is free access in the form of the simple "shonagon" online search option (which also provides word frequency) other access plans to the database (e.g. chuunagon or DVD versions) are not free and costs depend on the specific license type. The total number of words contained in the BCCWJ is 104.6 million; which is less than a half of the corpus source used for the databases presented in this paper (299.6 million). The sub-corpus of the newspaper in BCCWJ covers the period of 2001–2005.

The last database we would like to mention is the freely available kanji database by Tamaoka and Makioka (2004). The fourth edition (published in 2004) included several novel mathematical indexes such as kanji entropy, redundancy and symmetry and is downloadable as an Excel, Word or PDF file (https://www.lang.nagoya-u.ac.jp/ktamaoka/en/). This database has been amply used in psychological research on the Japanese language (e.g. Hino, Miyamura, & Lupker, 2011; Miwa, Libben, & Baayen, 2012; Miwa, Libben, Dijkstra, & Baayen, 2014; Toyoda, 2009; Verdonschot, La Heij, Tamaoka, Kiyama, You, & Schiller 2013). However, considering the recent change in the Jōyō-kanji list, also this database would need to reflect this change and it would benefit to be based on a more recent corpus as well.

The new Jōyō Kanji and Jukugo Databases

This paper presents the availability of two freely accessible databases: (1) a Jōyō kanji database (partially based on Tamaoka and Makioka 2004) and (2) a jukugo database, allowing for improved control and manipulation of important lexical factors when conducting psychological and linguistic research using Japanese kanji. We will now describe their properties and specific usage for psychological and linguistic research in more detail. Both databases can be freely accessed through a web interface at

[1] For details, see https://mecab.sourceforge.net.

[2] See http://www.ninjal.ac.jp/english/products/bccwj/ for an English explanation.

www.kanjidatabase.com. The manual on how to use this website can be found under the heading "manual".

Description of the Jōyō kanji database

The corpus used in constructing the Jōyō kanji database

To create the current 2136 Jōyō kanji database, we used 11 years' worth of articles from the *Mainichi Newspaper*, beginning in 2000 and ending in 2010. The morphological parsing program MeCab (Kudo, Yamamoto, & Matsumoto, 2004) totaled 477,264 morphological units (type frequency), and a total token frequency of 299,695,840 out of this newspaper corpus (including proper nouns; such as 日産 "Nissan" or 佐藤 "Sato"). Excluding proper nouns, the count was 368,841 for type frequency and 282,816,611 for token frequency. Using the *Mainichi Newspaper* corpus, the present kanji database lists the frequencies for each of the 2136 commonly-used Jōyō kanji. In the following subsections we will now briefly describe each property stored in the database in the order they appear in the look-up function on the www.kanjidatabase.com website and, where necessary, elaborate on their usefulness in experimental psychological and linguistic research.

Strokes (visual complexity)

Kanji strokes refer to the number of brush strokes required to draw a specific kanji. The numbers stored in the database are based on the "Teaching Guides for Kanji Stroke Order" provided in 1958 by the Japanese government (see Horiguchi, 1989). The number of strokes as an index of kanji visual complexity has been shown to exhibit inhibitory effects, indicated by studies on Chinese (Leong, Cheng, & Mulcahy, 1987) and Japanese characters (Tamaoka, & Takahashi, 1999; Tamaoka, & Kiyama, 2013). Leong et al. (1987) found that three major factors influence the processing of logographic characters: (1) orthography, (2) printed frequency, and (3) reading ability. Less visually complex characters were processed faster than those with greater visual complexity (when controlled for printed-frequency and reading ability). However, complex low-frequency characters were processed particularly slowly by less-skilled readers.

Tamaoka and Takahashi (1999) investigated initiation times of drawing two-kanji compound words (jukugo) that were phonetically presented to native Japanese speakers. The initiation time refers to the duration between the time when a participant heard a target two-kanji compound word and the time when the participant started drawing the left-side kanji of the two-kanji compound. This study indicated that there were inhibitory effects of initiation times for visually complex kanji for low frequency kanji while high frequency kanji showed no effects. Likewise, Tamaoka and Kiyama (2013) showed that visually complex kanji with low frequency elicited a heavier cognitive load within single kanji processing. Because visual complexity is an important factor when recognizing kanji (see also: Higuchi, Moriguchi, Murakami, Katsunuma, Mishima, & Uno, 2016), stroke counts must be controlled for conditions within experiments involving kanji.

Grade (age of acquisition)

Attaining kanji competence is much more difficult than learning the other two Japanese scripts, katakana and hiragana. Uno, Wydell, Haruhara, Kaneko and Shinya (2009) identified rates of developmental dyslexia among 495 students from elementary school grades two to six in Japan. They reported the rates of developmental dyslexia as the percentage of the students whose reading and writing scores of hiragana, katakana and kanji fell below the 1.5 standard deviation cut-off. According to this criterion, students displayed difficulty (i.e., those who were beyond the cut-off criterion) in pronouncing (reading) hiragana at 0.2 %, katakana at 1.4 %, and kanji at 6.9 %, while they showed similar difficulties in writing-hiragana at 1.6 %, katakana at 3.8 %, and kanji at 6.0 %. This shows that kanji can be a major obstacle for students at the elementary school level. Additionally, it has repeatedly been shown that words learned earlier in life can be recognized and produced faster than those learned in later in life, all other factors being equal. This phenomenon is often referred to as the age of acquisition (AoA) Effect. The AoA effect on lexical processing is believed to be independent from frequency effects (e.g. Barry, Morrison, & Ellis, 1997; Barry, Hirsh, Johnston, & Williams, 2001; Morrison, & Ellis, 2000). Since the AoA for a particular word is often determined by asking speakers subjective (and difficult to answer) questions (e.g.: "When did you acquire the word <X>?") grade-allocations from grades 1 to 6 for Japanese kanji may serve better as an objective measurement of AoA. The outline of the Japanese language curriculum was released by the Japanese Ministry of Education in 1989 and included a list (called gakushū kanji) containing 1006 kanji taught to students between grades one and six. According to the list, 80 kanji are taught in grade one, 160 kanji are taught in grade two, 200 kanji in grade three, 200 kanji in grade four, 185 kanji in grade five, and 181 kanji in grade six. All these kanji are included in the Jōyō Kanji List. The remaining 1130 kanji are taught in grades 7–9. Since Japanese citizens are obliged to take part in school at least through grade nine, it is reasonable to assume that all 2136 Jōyō kanji are acquired by adult native Japanese speakers educated in Japan.

Kanji classification

Many people believe logographic characters to be predominantly pictographic (e.g. 木 "a tree"). However, such pictographic kanji (Shookei), only comprise 269 kanji (12.6 %) of the Jōyō Kanji List. The majority of kanji (approximately 60 % of the Jōyō Kanji List) are classified as semantic-phonetic composites (Keisei). The next largest classification is semantic composites (Kaii). These are 532 kanji on the Jōyō Kanji list (24.9 %). The most common classification for kanji is called the Rikusho Bunrui or "Six Classifications of Chinese Characters" (for details see: Chikamatsu, 2005; Tamaoka, Kirsner, Yanase, Miyaoka, & Kawakami, 2002 in English; Atsuji, 1988 in Japanese). Values in the database follow the system described by Shirakawa (1994) namely: pictographic (象形), ideographic (指事), compound ideographic (会意), phonetic (形声), loan (仮借) and derivative cognate (転注). In addition to these six standard classifications, the current database also includes a new seventh classification, namely, Japanese-original kanji (i.e., not originally Chinese), termed original (国字; 6 kanji, 0.28 %) for example 峠 ("mountain pass"). Classification frequencies provide useful figures, for instance when one needs to work with a particular kanji class (i.e., pictographs, or compound ideographs).

The Japanese Language Proficiency test (JLPT-test)

The Japanese Language Proficiency Test (hereafter JLPT) for non-native Japanese speakers has been prepared and administered by both the Japan Foundation and the Japan Association of International Education since 1984. As many as 750,000 people took the JLPT in 2009 alone. The highest level is one and the lowest level is four. Although recently the JLPT has been altered to contain five levels (N1–N5); updated kanji lists for the new levels are not provided, therefore we report only the previous levels (1–4) for which official lists are available. Japanese language education, especially outside of Japan, focuses on teaching kanji used in the official JLPT kanji list, as many students who learn Japanese aim to pass the JLPT test(s).

There is a high correlation ($r = -0.67$; $p < .001$) between the grade in which native Japanese children study particular kanji and when those kanji appear in the JLPT prescribed curriculum. The crosstab (containing the 1006 elementary school kanji) across elementary school grade and JLPT levels is shown in Table 1. All 1006 kanji taught in grades one to six are included in JLPT levels four to one. As depicted in Table 1, many kanji taught at the forth (lowest) level of JLPT are also taught in grades one and two. Kanji taught in grades five and grade six are allocated to the second and first (highest) levels of the JLPT. The correlation is negative since the difficulty levels are reversed in JLPT, ranging from the easiest (fourth) level to the hardest (first) level. In research JLPT levels are typically used to control difficulties of lexical items when participants are non-native speakers (e.g. Komori, Tamaoka, Saito, & Miyaoka, 2014; Tanaka, 2015; Yamato, & Tamaoka, 2013).

Name of radical

Kanji are classified in dictionaries according to their main components called radicals. Radicals can be categorized into seven main groups according to their position within a kanji (hen = left, tsukuri = right, kanmuri = on top, ashi = at the bottom, tare = left bottom, go up, go to right top, nyō = left top, down, to right bottom, kamae = enclose like a square). Conversely, kanji dictionaries often use the 214 radical classifications derived from the Chinese Kangxi radicals. These radicals often may give a clue to the meaning or the pronunciation of a kanji (e.g. Leong, & Tamaoka, 1995; Saito, Masuda, & Kawakami, 1998, 1999; Saito, Yamazaki, & Masuda, 2002; Verdonschot et al., 2013). For example, kanji containing the semantic radical 虫 (mushi; "insect") typically represent an insect (e.g. 蜂 "bee", 蛍 "firefly") although there are exceptions (e.g. 蛇 "snake" and 虹 "rainbow"). Additionally, kanji containing 包 ("hoo") such as: 抱, 泡, 胞, 砲, 飽, 咆 mostly sound like "hoo", though this only holds for the (Chinese derived). On-reading and even then only in a selected number of cases. Importantly, kanji radicals have been shown to independently contribute to naming (Flores d'Arcais, Saito, & Kawakami, 1995), semantic classification (Flores d'Arcais, & Saito, 1993) and lexical decision latencies (Miwa et al., 2012). For instance, the latter authors observed semantic radical effects which were separate from whole word and individual kanji frequency effects in processing jukugo.

The current database reports the name of the radical found in each kanji (see Kaiho, & Nomura, 1983; Kess, & Miyamoto, 1999) taken from the radical classification found in the New Nelson's Japanese–English Character Dictionary (Haig, 1997). For instance, if one would like to obtain all the kanji with the insect radical one would type "mushi" at this field when looking in the database.

Radical frequency

Radical frequency refers to the number of kanji in the database sharing the same radical. The top 20 radicals encompass 1132 kanji (or 53.0 %) in the list suggesting that a small number of radicals are frequently used to construct the majority of kanji. It should be noted that the radical frequencies might depend to a certain extent on the way

kanji are classified in a dictionary, although potential changes per dictionary in frequency should be minimal.

Reading within Jōyō

Japanese kanji readings have two ways of reading them, namely: On-readings (derived from Chinese), which are typically associated with sounds, and Kun-readings (original Japanese words), which are typically associated with meanings. Some kanji with On-readings are semantically transparent, but this is quite rare. On-readings are most common within two-kanji compound words. Conversely, Kun-readings are native to Japanese, and usually carry a clear meaning. Most Japanese kanji (over 62 %) have more than one pronunciation. For instance, the kanji 下 for "below", "down" can be pronounced in six different ways in Kun-readings, with inflections as shita, shimo, moto, sa (geru/garu), kuda(su/saru), o(rosu/riru), and two different ways in On-reading, as ka and ge.

The Jōyō kanji list provides all Kun-readings, including inflections of verbs and adjectives. Researchers interested in the effect of the number of sounds in a single kanji should look at the number of Kun-readings without inflections. To meet this need, the present kanji database also provides Kun-readings without inflections. Although the number of On- and Kun-readings is specified in the Jōyō kanji list, there are rare additional readings for many kanji (e.g. one of Kun-readings, /kururi/ for 転). To accommodate this, two different calculations in both the number of pronunciations as well as frequencies are provided in the kanji database termed "within the Jōyō kanji list" and "beyond the Jōyō kanji list" for each of On- and Kun-reading.

Table 2 shows the numbers of kanji, a crosstab between On- and Kun-readings according to the pronunciations specified in the Jōyō kanji list. It, however, excludes overlapping Kun-readings by inflections (Kun-readings are counted without inflections). According to Table 2, 812 kanji, or 38.0 %, only have a single pronunciation, that is: either an On-reading or a Kun-reading. The total number of kanji with multiple On- or Kun-reading is 1324; this stems from the original 2136 kanji minus 69 kanji with a single Kun-reading and 743 with a single On-reading (i.e., 2136 − 69 − 743 = 1324). In other words, after excluding multiple On-readings or Kun-readings, 1324 kanji, or 62.0 %, out of the 2136 kanji have more than one pronunciation. Reading within Jōyō refers to the readings specified in the official Jōyō list. For instance, four readings for the kanji 幸 "happiness" are provided in the list, one On-reading /koo/, and three Kun-readings, namely: /saiwa(i)/, /satyi/ and /siawa(se)/.

Reading beyond Jōyō

The Jōyō kanji list provides only commonly-used readings. The previous example of 幸 'happiness' also has an additional Kun-reading, /miyuki/, which is not listed. To cover all available readings, the present database used the Kanjigen Kaitei Dai-5-ban [Kanji Sources Revised Fifth Version] (Todo, 2010) which is an often used and up-to-date source. Thus, reading beyond Jōyō refers to any reading not provided in the Jōyō kanji list, but included in the kanji dictionary.

# of On and On within Jōyō

An On-reading refers to a pronunciation taken from a Chinese original sound whereas a Kun-reading refers to a pronunciation of an original Japanese sound. Japanese kanji often has multiple readings. For example, the kanji 家 'house' has two On-readings (i.e. /ka/ and /ke/) and two Kun-readings (i.e., /i.e./ and /ya/) in the Jōyō kanji list. In the case of 家, the number of On-reading (# of On) is counted as 2. Thus, the On-readings for 家 within the list (On within Joyo) are/ka/and/ke/.

Kanji ID in Nelson

The New Nelson's Japanese–English Character Dictionary (Haig, 1997) is based on the Japanese–English Character Dictionary by Nelson (1962) and is prevalent within Japanese academic circles around the world. This dictionary contains 7107 kanji entries, each labeled with an entry number. For example, the kanji 貝 (kai, 'seashell') is numbered as 5766 in this dictionary. Kanji can be found by following the sequential numbering of the dictionary, and students can additionally look up the usages for compound words like 貝殻 (kaigara, 'seashell').

# of meanings of On

This parameter refers to the number of meanings associated with an On-reading (Chinese derived reading). A single On-reading occasionally has multiple meanings. For instance, the kanji 脳 is pronounced as noo in its On-reading. This is the only On-reading in this kanji, so the number of On-reading (# of On) is 1, and noo is the only On-reading listed in the Jōyō kanji list (On within Jōyō). However, associated with this On-reading, 脳 has two meanings, specifically: 'brain' and 'memory' according to the new Nelson's dictionary (Haig, 1997). Thus, # of meanings of On is 2. Thus, the English translations of the On-readings (Translation of On) are 'brain' and 'memory'.

# of Kun within Jōyō with inflections

The Japanese-original readings for kanji are called Kun-readings. These kanji are often used as the stem of a verb or an adjective, to which inflectional affixes can be added. In the Jōyō kanji list, the kanji 歌 representing 'a song' or 'to sing' can be either a noun, uta; or a verb, uta(u), which has the verbal inflectional affix—u. Therefore, the number of Kun-readings within the Jōyō kanji list with inflections (# of Kun within Jōyō with inflections) for this kanji (歌) is assessed to be 2; once for the noun of uta and once for the verb uta(u). When the inflection—u is removed, the Kun-reading is the same for both the noun and the verb, uta. The number of Kun-readings within the Jōyō kanji list without inflections (# of Kun within Jōyō without inflections) is counted only uta, and thus, the number of Kun-readings is 1.

Kun within Jōyō

Japanese kanji often have multiple readings in both On- and Kun-readings. For example, the kanji 歩 "to walk" or "a step" has two Kun-readings; aru(ku) for the verb "to walk" and ayu(mu) for the noun "a step" or "a walk" as specified in the Jōyō kanji list. Thus, these two Kun-readings are Kun within Jōyō.

# of meanings of Kun

Many kanji are considered to be morphemic units. Each kanji, therefore, can be combined with another to create multiple kanji compound words. This word construction process can increase the number of meanings associated with a kanji. For instance, the kanji 歩 has two Kun-readings, but three meanings related to these two sounds. Thus, the number of meanings of Kun-readings (# of meanings of Kun) is 3. The translations of two Kun-readings (translation of Kun) are 'walk', 'hike' and 'step'.

Year of inclusion

The first Jōyō kanji list was originally produced by the Japanese government in 1981. At this time, there were 1945 kanji in the list. In 2010, this list was modified to have 2136 kanji. Five kanji were removed, and 196 kanji were added. Kanji which appeared in the old version of the list are indicated by '1981'. The five kanji (i.e., 勺, 錘, 銑, 脹, 匁), which were excluded from the new list, are no longer included in the present database. The 196 kanji (e.g. 曖, 宛, 釜, 鎌, 窟, 熊) added in 2010 are recorded as '2010'. Since only two versions of the list are available, the values of Year of Inclusion take either '1981' or '2010'.

Kanji frequency with proper nouns

Lexical frequency has been repeatedly observed to be a strong factor influencing lexical processing (e.g. Balota, & Spieler, 1999; Brown, & Rubenstein, 1961; Gordon, 1983; Hino, & Lupker, 1998; Jescheniak, & Levelt, 1994; Segui, Mehler, Frauenfelder, & Morton, 1982; Starreveld, La Heij, & Verdonschot, 2013; Taft, 1979). As with lexical frequency, the frequency of a single kanji in a jukugo (e.g. the kanji si 思 "think" in the two-kanji compound sikoo 思考 "thought") has also been shown to influence the processing speed of the entire word (Tamaoka, & Hatsuzuka, 1995). This effect has also been documented for Chinese characters (Taft, Huang, & Zhu, 1994; Taft, & Zhu, 1995, 1997; Wu, Chou, & Liu, 1994; Zhou, & Marslen-Wilson, 1994).

Additionally, proper nouns such as names of people, companies, and places are processed differently from general nouns (Valentine, Moore, & Brédart, 1995). For instance, experimental studies comparing proper nouns and general nouns found reduced response latencies for proper nouns versus general nouns (e.g. Müller, 2010; Proverbio, Mariani, Zani, & Adorni, 2009; Wang, Verdonschot, & Yang, 2016). Wang et al. (2016) explain this difference as follows: the name "Thomas Edison" is connected to its semantic referent only via the knowledge of an individual whereas the general noun "inventor" is connected to a large number of associations representing semantic information. Frequencies for the 2136 kanji of the Jōyō kanji list were calculated using the *Mainichi Newspaper* corpus. Kanji frequencies with proper nouns (including names such as: 佐藤) ranged from 27 to 2817,613, with a mean of 85,823 (SD 173,833). The frequencies of the 2136 kanji with and without proper nouns adhere to a power-law distribution. In this distribution, among the 2136 kanji, a few kanji appear very frequently while the majority of kanji occur infrequently.

It should be noted that the accumulative kanji frequencies are calculated based on all words in the present newspaper corpus. Some words in the corpus cannot be identified as either an On-reading or a Kun-reading. Thus, the accumulative frequency of a kanji (Kanji frequency with/without proper nouns) is not always equal to the sum of a kanji's accumulative frequency of On-readings with/without proper nouns (Acc. Freq. On with/without proper nouns) and a kanji's accumulative frequency of On-readings with/without proper nouns (Acc. Freq. Kun with/without proper nouns). That is, the Kanji frequency with/without proper nouns can be larger than the sum of Acc. Freq. On with/without proper nouns and Acc. Freq. Kun with/without proper nouns.

Acc. Freq. On with proper nouns

We used MeCab (Kudo, Yamamoto, & Matsumoto, 2004) to identify to identify proper nouns (固有名詞) in the 2000–2010 *Mainichi Newspaper* corpus. Once proper nouns were tagged, the accumulative frequency of On-readings with proper nouns (Acc. Freq. On with proper nouns) was calculated for each of the 2136 kanji by selecting words only with On-readings including proper nouns.

Acc. Freq. Kun with proper nouns

Similarly, the accumulative frequencies of Kun-readings with proper nouns (Acc. Freq. Kun with proper nouns) were calculated for each of the 2136 kanji by selecting words only with Kun-readings including proper nouns.

On ratio with proper nouns

The On-reading ratio is defined as a kanji's accumulative frequency of On-readings divided by the kanji's total accumulative frequency of On-readings and Kun-readings. Since the values in Kanji frequency with proper nouns include kanji whose pronunciations cannot be identified as either On- or Kun-readings, the On-reading ratio is calculated by Acc. Freq. On with proper nouns divided by Acc. Freq. On with proper nouns plus Acc. Freq. Kun with proper nouns. The Kun-reading ratio for each kanji is simply calculated by the inverse (i.e., 1—the On ratio). On-reading ratios and Kun-reading ratios are useful for situations which require controlling kanji with multiple readings while still experimentally assessing phonological activations when processing Japanese kanji. For instance, using masked priming, Verdonschot et al. (2013) showed that all pronunciations of a single kanji with around a 50 % On-reading ratio are simultaneously activated upon visual presentation of that kanji (e.g. mizu and sui, both meaning water, are activated upon seeing □). Tamaoka and Taft (2010) also controlled kanji On-reading ratios to investigate On- and Kun-reading sub-lexica.

The present kanji database provides four different On-reading ratio categories by separating kanji with and without proper nouns, and similarly within and beyond standard On-readings in the Jōyō kanji list. For On-readings within the Jōyō kanji list: (1) the On-reading ratios with proper nouns averaged 73.83 % (SD 33.60 %) while (2) the On-reading ratios without proper nouns averaged 74.82 % (SD 32.91 %). For On-readings, including those beyond the standard Jōyō kanji, (3) the On-reading ratios with proper nouns averaged 73.77 % (SD 32.65 %) while (4) On-reading ratios without proper nouns averaged 71.87 % (SD 33.30 %). The second category (i.e., the On-reading ratios without proper nouns) is recommended for usage in most Japanese language processing experiments.

Acc. Freq. On/Kun beyond Jōyō with proper nouns

The Jōyō kanji list provides On- and Kun-readings for each of the 2136 kanji. However, the list excluded some rare readings. For researchers who wish to have a complete kanji frequency overview including all the kanji sounds, we calculated both kanji frequencies within and beyond the Jōyō kanji list. It should be noted that rare readings beyond the Jōyō kanji list are taken from readings described in the Kanji Sources Revised dictionary (Fifth Version; Todo, 2010).

Acc. Freq. Kun beyond Jōyō with proper nouns

The accumulative frequencies of Kun-readings beyond the Jōyō kanji list with proper nouns (Acc. Freq. Kun beyond Jōyō with proper nouns) were calculated for each of the 2136 kanji by selecting words only with Kun-readings beyond the list, including proper nouns. Those kanji which do not have any Kun-reading outside of the Jōyō kanji list have this frequency recorded 0. For instance, the kanji □ has only one Kun-reading with inflection mana(bu) included in the list, so its Acc. Freq. Kun beyond proper nouns is 0. In case that a Kun-reading is found beyond the list, if there is no words found in the newspaper corpus, it's frequency would be 0 as well.

On ratio beyond Jōyō with proper nouns

The On-reading ratio beyond the Jōyō kanji list with proper nouns (On ratio beyond Jōyō with proper nouns) is defined as a kanji's accumulative frequency of On-readings divided by the kanji's total accumulative frequency with proper nouns, including both On-readings and Kun-readings beyond the list. This ratio is considered as an overall grand On-reading ratio of the present corpus. It should be noted that an On-reading ratio beyond the Jōyō kanji list with proper nouns (On ratio beyond Jōyō with proper nouns) is equal to an On-reading ratio within the Jōyō kanji list without proper nouns (On ratio with proper nouns) when either there are no On-/Kun-readings beyond the list or there are no words sounded by On-/Kun-readings beyond the list in the present corpus.

Kanji frequency without proper nouns

See Kanji frequency with proper nouns for details. The same kanji without proper nouns ranged in frequency 6–1,855,755 with a mean of 73,337 (SD 149,730).

Acc. Freq. On without proper nouns

The accumulative frequencies of On-readings without proper nouns (Acc. Freq. On without proper nouns) were calculated for each of the 2136 kanji by selecting words only with On-readings excluding proper nouns.

Acc. Freq. Kun without proper nouns

Likewise, the accumulative frequencies of Kun-readings without proper nouns (Acc. Freq. Kun without proper nouns) were calculated for each of the 2136 kanji by selecting words only with Kun-readings excluding proper nouns.

On ratio without proper nouns

The On-reading ratio without proper nouns (On ratio without proper nouns) was calculated using accumulative frequencies for kanji with On-reading without proper nouns divided by the sum of the kanji's On-reading and Kun-reading accumulative frequency without proper nouns.

Acc. Freq. On beyond Jōyō without proper nouns

The accumulative frequencies of On-readings without proper nouns (Acc. Freq. On without proper nouns) were calculated for each of the 2136 kanji by selecting words beyond On-readings excluding proper nouns. As it is rather rare to find On-readings beyond the list, the frequency will often be zero.

Acc. Freq. Kun beyond Jōyō without proper nouns

The accumulative frequencies of Kun-readings without proper nouns (Acc. Freq. Kun without proper nouns) were calculated for each of the 2136 kanji by selecting words beyond Kun-readings excluding proper nouns. As it is rather rare to find Kun-readings beyond the list, this frequency will often be zero.

On ratio beyond Jōyō without proper nouns

The On-reading ratio beyond the Jōyō kanji list without proper nouns (On ratio beyond Jōyō without proper nouns) is defined as a kanji accumulative frequency of On-readings divided by the total kanji accumulative frequency including On-/Kun-readings beyond the list, but excluding proper nouns. Since proper nouns are excluded, this On-reading ratio is considered unbiased and suitable for the use of experiments.

Left Kanji Prod. and right Kanji Prod

A large majority of Japanese words consists of two kanji. According to Yokosawa and Umeda (1988), two-kanji compound words are extremely common, making up about 70 % of the entries in Japanese dictionaries. In two-kanji compound words, each kanji can be placed on either the left-side or the right-side. Kanji lexical productivity (Kanji Prod.) refers to how frequently a single kanji appears in two-kanji compound words in combination with another kanji. For instance, 水 'water' with the On-reading, sui, can be placed on the left-side of the compound, as 水深 sui + sin 'the depth of the water', 水泳 sui + ei 'swimming', 水洗 sui + sen 'water closet', or on the right-side of the compound, as 渇水 kas-sui 'shortage of water', 湖水 ko-sui 'lake', and 噴水 hun-sui 'fountain'. Kanji productivity refers to two units of kanji being combined to create two-kanji compound words. When the kanji is placed on the left-side of the compound, the number of words which are created by the kanji is the left-side kanji productivity (left Kanji Prod.). Likewise, when the kanji is placed on the right-side, the number of words produced is the right-side kanji productivity (right Kanji Prod.). To calculate these productivities, all 27,950 jukugo (two-kanji compound words) were taken from 2000 to 2010 *Mainichi Newspaper* corpus. The number of left-hand and right-hand kanji productivities are counted for each of the 2136 kanji, using this two-kanji compound word database.

Acc. Freq. left Prod. and Acc. Freq. right Prod

Kanji left-hand and right-hand productivities (left Kanji Prod. and right Kanji Prod.) are simply a count of two-kanji compound words produced by a single kanji with no consideration of word frequency. Accumulative word frequency of all words together is considered to be more accurate in indicating the magnitude of kanji lexical productivity, compared to a simple count of each produced word (Tamaoka, & Makioka, 2004). Accumulative kanji lexical productivities on the left-side (Acc. Freq. left Prod.) and on the right-side (Acc. Freq. right Prod.) are calculated by adding all the frequencies of occurrence for words in all 27,950 jukugo.

Symmetry

As explained in the kanji productivity section, each kanji creates various two-kanji compound words by combining with another kanji on the left-side or right-side. Symmetry indicates a balance tendency of kanji productivity between the left-side and the right-sides. An asymptotic test for

symmetry is performed for each of the 2136 Jōyō Kanji. Consider the number of left-side compounds to be nL, and those of the right-hand side to be nR, and $n\mathrm{L} + n\mathrm{R} = n$. Under the hypothesis of equality of both sides, the expected value is $n/2$. The asymptotic Chi square criterion is:

$$\chi^2 = \frac{(n\mathrm{L} - n\mathrm{R})^2}{n\mathrm{L} + n\mathrm{R}}$$

which is distributed as a Chi square with one degree of freedom. In order to be significant at the 0.05 level, this Chi square value must be greater than 3.84 (for details see, Tamaoka, & Altman, 2004).

When a kanji occurs in fewer than five compounds, it occurs too infrequently to be symmetrical. The symmetry of these kanji was then recoded as '·'. When a kanji was judged to be symmetric (no significance), it was represented by 'S.' When the left-side productivity (left Kanji Prod.) was greater than that of the right-hand side with a Chi square value larger than 3.84, a kanji was judged as progressively asymmetric, indicated by 'P.' When the right-side productivity (right Kanji Prod.) was greater than the left-side, a kanji was judged as regressively asymmetric, represented by 'R.' A large number out of the 2136 kanji in the Jōyō kanji list (939 kanji or 44.00 %) portrayed particular symmetry patterns. Concerning the two asymmetry types, we found 372 kanji (or 17.40 %) of the progressive 'P' type, and 421 kanji (or 19.70 %) of the regressive 'R' type.

Left entropy and right entropy

Entropy in the present database refers to how randomly each kanji produces a two-kanji compound (for more details see, Tamaoka, & Makioka, 2004, p.553.). Entropy is calculated using the formula.

$$H = -\sum p_j \log^2 p_j$$

In this formula, the p in the formula stands for the probability that a specific word will appear among all the compound words combined with multiple kanji on the left-side (left Entropy) or the right-side (right Entropy) of the kanji. If a kanji produces a great variety of two-kanji compound words, its entropy is larger. On the other hand, if the kanji is combined with a small number of kanji to produce two-kanji compound words, its entropy is smaller. The entropies of the 2136 kanji in the Jōyō kanji list indicated a mean of 1.34 (SD 1.15) for the left-side, and a mean of 1.41 (SD 1.13) for the right-side. Entropies depicted an overall similar pattern for both sides when producing two-kanji compound words among the 2136 kanji.

Left/right sound [1–7] and left/right frequency [1–7]

As described earlier, many kanji have multiple pronunciations, which differ depending upon whether the kanji is positioned on the left or right side of a compound. The present database provides frequency order concerning kanji pronunciations for both sides (left/right sound [1–7]) and their accumulative (or token) frequencies for both sides (left/right frequency [1–7]). Occurrences of a given kanji, in which Kun-readings with inflections are counted once, are aggregated in Table 3 (for details see the section of On- and Kun-readings) for calculating the number of Kun-readings available for a given kanji. As depicted in Table 3, five kanji placed on the left side of a compound involved up to six different kanji pronunciations while two kanji placed on the right side had up to seven different pronunciations.

Comparing the kanji frequency of the present database with previous databases

The overlapping kanji between the old and new Jōyō kanji list are exactly 1940. The two previously-created kanji databases were compared using these shared kanji with the present kanji database. Yokoyama, Sasahara, Nozaki and Long (1998) created a database consisting of 4583 kanji and their printed frequencies using the 1993 Tokyo edition of the Asahi Newspaper corpus, consisting of a printed and a CD-ROM version. The printed version (M 8611 times, SD 18,065 times) is smaller in corpus size than CD-ROM version (M 12,514 times, SD 26,122 times). Tamaoka et al., (2002) produced the first web-accessible kanji database in 2002 containing the earlier version of the Jōyō kanji list which culminated in the fourth edition two years later (Tamaoka, & Makioka, 2004).

Pearson's correlations coefficients among these three kanji databases and the present database were calculated using natural logarithms converted from kanji frequencies. The kanji database of the printed 1993 Asashi newspaper version (Yokoyama et al., 1998) was .966 ($p < .001$) with the present database of the 2000–2010 *Mainichi Newspaper* with proper nouns, and .922 ($p < .001$) without proper nouns. Likewise, the kanji database of the Asashi newspaper 1993 CD-ROM version also showed a very high correlation at .965 ($p < .001$) with the present database with proper nouns, and .918 ($p < .001$) without proper nouns. The kanji database encompassing the 1985–1998 Asashi newspaper based on Amano, & Kondo (1999, 2000) (see Tamaoka, & Makioka, 2004) showed a slightly lower, but still very high correlation with the present database at

Table 1 Distribution of the 2136 kanji in the Jōyō kanji list within school grades and JLPT-test levels

School grade	JLPT-test level				Total
	4th	3rd	2nd	1st	
1st	49	19	12	0	80
2nd	31	72	51	6	160
3rd	0	58	133	9	200
4th	0	12	158	30	200
5th	0	2	137	46	185
6th	0	2	98	81	181
Total	80	165	589	172	1006

Table 2 Numbers of On- and Kun-readings in the Jōyō kanji list ($N = 2136$)

Number of Kun-readings	Number of On-readings					Sum
	0	1	2	3	5	
0	0	743	76	2	0	821
1	69	886	126	11	1	1,093
2	6	142	43	3	0	194
3	0	17	4	1	0	22
4	0	1	2	0	0	3
5	0	0	1	0	0	1
6	0	0	2	0	0	2
Sum	75	1789	254	17	1	2136

This table was created using the numbers of pronunciations specified in the 2136 Jōyō kanji list, but the number of Kun-readings in the table excludes overlapped sounds in Kun-readings with inflections (the same sound is counted once)

Table 3 Frequency order of kanji pronunciations and their accumulative (token) frequencies for the 2136 kanji on the Jōyō kanji list

Frequency order of kanji pronunciations	Left-side			Right-side		
	N	Accumulative frequency		*N*	Accumulative frequency	
		M	SD		Mean	SD
1st	1988	23,844	53,992	1912	24,825	53,353
2nd	816	3353	12,659	742	3332	12,695
3rd	213	1763	8779	233	1123	3397
4th	55	291	615	72	524	1310
5th	15	85	122	17	375	872
6th	5	38	63	11	62	84
7th	–	–	–	2	83	98

.847 ($p < .001$) with proper nouns, and .895 ($p < .001$) without proper nouns. The correlation between the present kanji database between with and without proper nouns indicated a high correlation at .948 ($p < .001$). In all, the present kanji database displayed high correlations with the previous two kanji databases.

The novel two-kanji compound word (jukugo) database

Experimental studies involving the Japanese language most often use two-kanji compound words. In our database, two-kanji compound words are defined as

words combing a left-side kanji, or first-kanji, with a right-side kanji, or second-kanji. These compounds are frequently used to depict lexical items. In contemporary Japanese, kanji compound words not only represent lexical items, adapted to Japanese orthography from Chinese (kango), but also various new lexical items formed by the Japanese themselves (e.g. 経済 keizai 'economy') in the process of translating various European and American books especially around the Meiji Restoration in 1868. These new compound words are frequently constructed by two kanji with On-readings. Additionally, words which originated in Japan (wago) mostly contain two kanji which are pronounced in their Kun-readings. With these distinct features, the majority of lexical items in the Japanese lexicon are two-kanji compounds. In order to enhance the usefulness of the kanji database further we created a lexical database containing Japanese jukugo (two-kanji compounds) with their lexical properties (all without proper nouns). All 27,950 compounds were selected from the *Mainichi Newspaper* corpus. Compounds are separately accessible, but all compounds for a given kanji can be searched and stored via a direct query from the website (see the manual section of the website for details).

It should be noted that, according to MeCab (Kudo Yamamoto, & Matsumoto, 2004), some two-kanji compound words are treated as two separate lexical categories. These compounds are written using the same two kanji, therefore we combined the two different frequencies into one frequency. For example, koohee 公平, 'fair', is counted 3659 times as an adjective, and 116 times as a noun. Although this word is more often used as an adjective, we still counted both frequencies together, totaling 3775. The final database of compounds encompassed a total type frequency of 27,950 words, and the total token frequency was 50,813,587.

The jukugo database provides information ordered in nine columns. For instance, inputting a single kanji 玉 'ball' into the "look up jukugo" field on the website will display 46 compounds. In 20 of these compounds, 玉 appears on the left side (e.g. gyokusai 玉砕), and in 26 compounds, 玉 appears on the right side (e.g. medama 目玉). Nine columns displaying information for each compound are subsequently provided by the present database: (1) compound word ID, (2) the two-kanji compound itself (3) compound word frequency, (4) additional grammatical usage for the compound, (5) pronunciation in Roma-ji (Japanese Romanization), and (6) an English translation, (7) a target kanji position (left or right side position in a compound), (8) the target kanji, and (9) the ID of the target kanji.

Conclusion

There are ample psychological and linguistic studies published on a variety of topics which have involved the Japanese language in some way (see "Introduction"). Many of these studies included the usage of kanji (a logographic script derived from Chinese) with a focus on those kanji which are assumed to be known by all Japanese people. These kanji are officially represented in the Jōyō kanji-list which was originally drafted in 1981 by the Japanese government (but has undergone a significant change in 2010).

Currently available databases have not yet included this change into their database. Additionally, many are based on slightly dated corpora and the majority of the databases cost a significant amount of money to use. These issues were the basis for the construction of two new and freely available databases involving Japanese kanji and their compound words. The present paper reports the availability of: (1) a database containing all 2136 kanji from the most recent Jōyō Kanji list and (2) a database containing their 27,950 two-kanji compound words (jukugo). Both databases are based on the extensive (i.e., a token frequency of about 300 million) and recent *Mainichi Newspaper* corpus (ranging from 2000 to 2010). To provide interactive access to these databases we have constructed an easy-to-use online web interface (www.kanjidatabase.com) that allows unrestricted access to all the information stored in both databases.

The primary use of the present databases and its interactive website is to aid psychologists, linguists and other scholars interested in performing research involving the Japanese language, although it is not difficult to imagine other, more educational (and perhaps even recreational) uses. The four most important reasons to consider the current databases are: (1) they are based on the most recent Jōyō kanji list, (2) they are based on a very large and up-to-date corpus, (3) they are freely available (opposed to many other databases), (4) access to the databases is made exceptionally easy through an online graphical user interface (www.kanjidatabase.com).

In conclusion, the databases laid out in this paper can be readily used for the construction of stimuli for psycholinguistic and linguistic experiments or to look up specific properties of a kanji for scientific purposes. We believe they constitute a valuable addition for those working with the Japanese language.

Acknowledgments The present work was supported by the Grant-in-Aid for Challenging Exploratory Research, JSPS Grant number

25580112 (principal researcher: Katsuo Tamaoka), by the Grant-in-Aid for Grant-in-Aid for Scientific Research (C), JSPS Grant Number 15K02656 (principal researcher: Kazuko Komori), and a Grand-In-Aid for JSPS postdoctoral fellows (12F02315) and a JSPS Research Activity Start-Up Grant (15H06687) to Rinus G. Verdonschot.

References

Amano, S., & Kondo, T. (1999). *NTT deeta beesu siriizu: Nihongo no goi tokusei—Dai 1-ki [NTT database series: Lexical properties in Japanese, the first period]*. Tokyo: Sanseido.

Amano, S., & Kondo, T. (2000). *NTT deeta beesu siriizu: Nihongo no goi tokusei—Dai 2-ki [NTT database series: Lexical properties in Japanese, the second period]*. Tokyo: Sanseido.

Atsuji, T. (1988). Kanji-no bunrui: Rikusho-o chuushin toshite [Kanji classification: focusing on six classifications]. In K. Sato (Ed.), *Kanji kooza 1: Kanji towa [Kanji lecture series 1: what is kanji?]* (pp. 49–69). Tokyo: Meiji Shoin.

Balota, D. A., & Spieler, D. H. (1999). Word-frequency, repetition, and lexicality effects in word recognition tasks: beyond measures of central tendency. *Journal of Experimental Psychology: General, 128*, 32–55.

Barry, C., Hirsh, K. W., Johnston, R. A., & Williams, C. L. (2001). Age of acquisition, word frequency, and the locus of repetition priming of picture naming. *Journal of Memory and Language, 44*, 350–375.

Barry, C., Morrison, C. M., & Ellis, A. W. (1997). Naming the Snodgrass and Vanderwart pictures: effects of age of acquisition, frequency and name agreement. *Quarterly Journal of Experimental Psychology, 50A*, 560–585.

Brown, H., & Rubenstein, C. R. (1961). Test of response bias explanation of word-frequency effect. *Science, 133*, 280–281.

Chen, H. C., Cheung, H., & Lau, S. (1997). Examining and reexamining the structure of Chinese–English bilingual memory. *Psychological Research, 60*(4), 270–283.

Chikamatsu, N. (2005). L2 Japanese kanji memory and retrieval: An experimental on the tip-of-the-pen (TOP) phenomenon. In V. Cook & B. Bassetti (Eds.), *Second language writing* (pp. 71–96). New York: Multilingual Matters Ltd.

Flores d'Arcais, G. B., & Saito, H. (1993). Lexical decomposition of complex Kanji characters in Japanese readers. *Psychological Research, 55*, 52–63.

Flores d'Arcais, G. B., Saito, H., & Kawakami, M. (1995). Phonological and semantic activation in reading kanji characters. *Journal of Experimental Psychology Learning Memory and Cognition, 21*, 34–42.

Frith, U. (1981). Experimental approaches to developmental dyslexia: an introduction. *Psychological Research, 43*(2), 97–109.

Gordon, B. (1983). Lexical access and lexical decision: mechanisms of frequency sensitivity. *Journal of Verbal Learning and Verbal Behavior, 22*, 24–44.

Haig, J. H. (1997). *The new Nelson Japanese–English character dictionary: based on the classic edition by Andrew N. Nelson*. Tokyo: Tuttle Publishing.

Higuchi, H., Moriguchi, Y., Murakami, H., Katsunuma, R., Mishima, K., & Uno, A. (2016). Neural basis of hierarchical visual form processing of Japanese Kanji characters. *Brain and Behavior,*. doi:10.1002/brb3.413.

Hino, Y., & Lupker, S. J. (1998). The effects of word frequency for Japanese Kana and Kanji words in naming and lexical decision: can the dual-route model save the lexical-selection account? *Journal of Experimental Psychology Human Perception and Performance, 24*, 1431–1453.

Hino, Y., Miyamura, S., & Lupker, S. J. (2011). The nature of orthographic–phonological and orthographic–semantic relationships for Japanese kana and kanji words. *Behavior Research Methods, 43*, 1110–1151.

Horiguchi, J. (1989). Kanji no hitsujun [Stroke order of kanji]. In Y. Takebe (Ed.), *Nihongoto nihongo kyooiku: Dai-8-kan. Nihongono moji hyooki (Joo) [Japanese and Japanese education: Vol. 8. Japanese writing system, No. 1]* (pp. 97–124). Tokyo: Meiji Shoin.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology Language Memory and Cognition, 20*, 824–843.

Jincho, N., Feng, G., & Mazuka, R. (2014). Development of text reading in Japanese: an eye movement study. *Reading and Writing, 27*(8), 1437–1465.

Kaiho, H., & Nomura, Y. (1983). *Kanji joohoo shori no shinrigaku [Psychology of kanji information processing]*. Tokyo: Kyoiku Shuppan.

Kess, J. F., & Miyamoto, T. (1999). *The Japanese mental lexicon: psycholinguistic studies of kana and kanji processing*. Amsterdam: John Benjamins.

Komori, K., Tamaoka, K., Saito, N., & Miyaoka, Y. (2014). Dai-2-gengo tosite Nihongo-o manabu chuugokugo wasya no nihongo no kanjigo no shuutoku ni kansuru koosatsu. Acquisition of Japanese kanji compound words by Chinese native speakers learning Japanese as a second language. *Chuugoku-go washa no tameno nihongo kyooiku kenkyuu [Studies on Japanese language education for native Chinese speakers], 5*, 1–16.

Kudo, T., Yamamoto, K., & Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. In: Proceedings of the 2004 conference on empirical methods in natural language processing (EMNLP-2004) (pp. 230–237).

Le Bigot, N., Passerault, J. M., & Olive, T. (2009). Memory for words location in writing. *Psychological Research, 73*(1), 89–97.

Leong, C. K., & Tamaoka, K. (1995). Use of phonological information in processing kanji and katakana by skilled and less skilled Japanese readers. *Reading and Writing, 7*, 377–393.

Leong, C. K., Cheng, P.-W., & Mulcahy, R. (1987). Automatic processing of morphemic orthography. *Language and Speech, 30*, 181–196.

Luo, C., & Proctor, R. W. (2013). Asymmetry of congruency effects in spatial stroop tasks can be eliminated. *Acta Psychologica, 143* (1), 7–13.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., … Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation, 48*, 345–371.

Miwa, K., Libben, G., & Baayen, R. H. (2012). Semantic radicals in Japanese two-character word recognition. *Language and Cognitive Processes, 27*(1), 142–158.

Miwa, K., Libben, G., Dijkstra, T., & Baayen, R. H. (2014). The time-course of lexical activation in Japanese morphographic word recognition: evidence for a character-driven processing model. *Quarterly Journal of Experimental Psychology, 67*, 79–113.

Morohashi, T. (2000). *Dai Kanwa Jiten [The great Japanese kanji dictionary]*. Tokyo: Taishukan.

Morrison, C. M., & Ellis, A. W. (2000). Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology, 91*, 167–180.

Müller, H. M. (2010). Neurolinguistic findings on the language lexicon: the special role of proper names. *Chinese Journal of Physiology, 53*(6), 351–358.

Nelson, A. N. (1962). *The original modern reader's Japanese–English character dictionary* (Classic ed.). Tokyo: Tuttle Publishing. **(the former Charles E. Tuttle Company)**.

Ono, F., & Kawahara, J. I. (2008). The effect of false memory on temporal perception. *Psychological Research, 72*(1), 61–64.

Proverbio, A. M., Mariani, S., Zani, A., & Adorni, R. (2009). How are 'Barack Obama' and 'President Elect' differentially stored in the brain? An ERP investigation on the processing of proper and common noun Pairs. *PLoS One, 4*(9), e7126.

Saito, H., Masuda, K., & Kawakami, M. (1998). Form and sound similarity effects in kanji recognition. In C. K. Leong & K. Tamaoka (Eds.), *Cognitive processing of the Chinese and Japanese languages* (pp. 169–203). London: Kluwer Academic Publishers.

Saito, H., Masuda, K., & Kawakami, M. (1999). Subword activation in reading Japanese single kanji character words. *Brain and Language, 68*, 75–81.

Saito, H., Yamazaki, O., & Masuda, H. (2002). The effect of number of Kanji radical companions in character activation with a multi-radical-display task. *Brain and Language, 81*, 501–508.

Segui, J., Mehler, J., Frauenfelder, U., & Morton, J. (1982). The word frequency effect and lexical access. *Neuropsychologia, 20*, 615–627.

Shirakawa, S. (1994). *Jitoo [Kanji etymology]*. Tokyo: Heibonsha.

Starreveld, P. A., La Heij, W., & Verdonschot, R. G. (2013). Time course analysis of the effects of distractor frequency and categorical relatedness in picture naming: an evaluation of the response exclusion account. *Language and Cognitive Processes, 28*, 633–654.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory and Cognition, 7*, 263–272.

Taft, M., Huang, J., & Zhu, X. P. (1994). The influence of character frequency on word recognition responses in Chinese. In H.-W. Chang, J.-T. Huang, C.-W. Hue, & O. J. L. Tzeng (Eds.), *Advances in the study of Chinese language processing* (Vol. 1, pp. 59–73). Taipei: Department of Psychology, National Taiwan University.

Taft, M., & Zhu, X. P. (1995). The representation of bound morphemes in the lexicon: a Chinese study. In L. B. Feldman (Ed.), *Morphological aspects of language processing* (pp. 293–316). Hillsdale: Lawrence Erlbaum Associates.

Taft, M., & Zhu, X. P. (1997). Submorphemic processing in reading Chinese. *Journal of Experimental Psychology Learning Memory and Cognition, 23*, 761–775.

Tamaoka, K., & Altmann, G. (2004). Symmetry of Japanese kanji lexical productivity on the left- and right-hand sides. *Glottometrics, 7*, 68–88.

Tamaoka, K., & Hatsuzuka, M. (1995). Kanzi niji jyukugo no shori niokeru kanji siyoohindo no eikyoo [The effects of kanji printed-frequency on processing Japanese two-morpheme compound words]. *Dokusho Kagaku [The Science of Reading], 39*, 121–137.

Tamaoka, K., Kirsner, K., Yanase, Y., Miyaoka, Y., & Kawakami, M. (2002). A Web-accessible database of characteristics of the 1945 basic Japanese kanji. *Behavior Research Methods Instruments and Computers, 34*, 260–275.

Tamaoka, K., & Kiyama, S. (2013). The effects of visual complexity for Japanese kanji processing with high and low frequencies. *Reading and Writing, 26*(2), 205–223.

Tamaoka, K., & Makioka, S. (2004). New figures for a Web-accessible database of the 1945 basic Japanese kanji, fourth edition. *Behavior Research Methods, Instruments and Computers, 36*, 548–558.

Tamaoka, K., & Taft, M. (2010). The sensitivity of native Japanese speakers to On and Kun kanji readings. *Reading and Writing, 23*, 957–968.

Tamaoka, K., & Takahashi, N. (1999). Kanji niji jyukugo no shoji koodoo niokeru goi siyoo hindo oyobi shojiteki hukuzatsusei no eikyoo [The effects of word frequency and orthographic complexity on the writing process of Japanese two-morpheme compound words]. *Sinrigaku Kenkyuu [The Japanese Journal of Psychology], 70*, 45–50.

Tanaka, M. (2015). Japanese Kanji word processing for Chinese Learners of Japanese: a study of homophonic and semantic primed lexical decision tasks. *Theory and Practice in Language Studies, 5*(5), 900–905.

Todo, A. (2010). *Kanji-gen Kaitei Dai-5-ban [Kanji Sources Revised Fifth Version]*. Tokyo: Gakken.

Toyoda, E. (2009). An analysis of L2 readers' comments on kanji recognition. *Electronic Journal of Foreign Language Teaching, 6*, 5–20.

Uno, A., Wydell, T. N., Haruhara, N., Kaneko, M., & Shinya, N. (2009). Relationship between reading/writing skills and cognitive abilities among Japanese Primary-School Children: normal readers versus poor Readers (dyslexics). *Reading and Writing, 22*, 755–789.

Valentine, T., Moore, V., & Brédart, S. (1995). Priming production of people's names. *The Quarterly Journal of Experimental Psychology Human Experimental Psychology, 48*, 513–535.

Verdonschot, R. G., La Heij, W., Tamaoka, K., Kiyama, S., You, W.-P., & Schiller, N. O. (2013). The multiple pronunciations of Japanese kanji: a masked priming investigation. *Quarterly Journal of Experimental Psychology, 66*, 2023–2038.

Wang, L., Verdonschot, R. G., & Yang, Y. (2016). The processing difference between person names and common nouns in sentence contexts: an ERP study. *Psychological Research, 80*, 94–108.

Wu, J.-T., Chou, T.-L., & Liu, I.-M. (1994). The locus of the character/word frequency effect. In H.-W. Chang, J.-T. Huang, C.-W. Hue, & O. J. L. Tzeng (Eds.), *Advances in the study of Chinese language processing* (Vol. 1, pp. 31–58). Taipei: Department of Psychology, National Taiwan University.

Yamada, J., Mitarai, Y., & Yoshida, T. (1991). Kanji words are easier to identify than katakana words. *Psychological Research, 53*(2), 136–141.

Yamato, Y., & Tamaoka, K. (2013). Chuugokujin nihongo gakushu-usha niyoru gairaigo shori eno eigo rekisikon no eikyoo [Effects of English knowledge on the reading of Japanese texts via Japanese loanwords performed by native Chinese speakers learning Japanese]. *Lexicon Forum, 6*, 229–267.

Yokosawa, K., & Umeda, M. (1988). Processes in human Kanji-word recognition. In: Proceedings of the 1988 IEEE international conference on systems, man, and cybernetics, August 8–12, 1988, Beijing and Shenyang, China, pp. 377–380.

Yokoyama, S., Sasahara, H., Nozaki, H., & Long, E. (1998). *Shinbun denshi media-no kanji: Asahi shinbun CD-ROM-ni yoru kanji hindo hyoo [Japanese kanji in the newspaper media: Kanji frequency index from the Asashi Newspaper on CD-ROM]*. Tokyo: Sanseido.

Yu, H., Gong, L., Qiu, Y., & Zhou, X. (2011). Seeing Chinese characters in action: an fMRI study of the perception of writing sequences. *Brain and Language, 119*(2), 60–67.

Zhou, X., & Marslen-Wilson, W. (1994). Words, morphemes and syllables in the Chinese mental lexicon. *Language and Cognitive Processes, 9*, 393–422.