Glottometrics 10
2005

Corpus Studies on Japanese Kanji

Guest Editor

Katsuo Tamaoka

Hiroshima University, Japan

Hituzi Syobo and RAM-Verlag

Tokyo, Japan/ Lüdenscheid, Germany

ISSN 2625-8226

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitdchrift (2-3 Ausgaben pro Jahr) für die quantitative Erforschung von Sprache und Text.

Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden.

Glottometrics kann aus dem **Internet** heruntergeladen werden (**Open Access**), auf **CD-ROM** (PDF-Format) oder als **Druckversion** bestellt werden.

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals (2-3 times a year).

Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors.

Glottometrics can be downloaded from the **Internet (Open Access)**, obtained on **CD-ROM** (as PDF-file) or in form of **printed copies.**

Herausgeber – Editors

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an

Orders for CD-ROM or printed copies to RAM-Verlag RAM-Verlag@t-online.de

Herunterladen/ Downloading: https://www.ram-verlag.eu/journals-e-journals/glottometrics/

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 10 (2005), Lüdenscheid: RAM-Verlag, 2005. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet (Open Access) unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar.
Bibliographische Deskription nach 10 (2005) **ISSN 2625-8226**

Contents

Kess, J. K.
On the History, Use, and Structure of Japanese Kanji 1-15

Tamaoka, K., Altmann, G.
Mathematical Modelling for Japanese Kanji Strokes in Relation to Frequency, Asymmetry and Readings 16-29

Masuda, H., Joyce, T.
A Database of Two-Kanji Compound Words Featuring Morphological Family, Morphological Structure, and Semantic Category Data 30-44

Miyaoka, Y., Tamaoka, K.
A Corpus Investigation of the *Right-hand Head Rule* Applied to Japanese Affixes 45-54

Long, E., Yokoyama, Sh.
Text genre and kanji frequency 55-72

Tamaoka, K., Matsuoka, Ch., Sakai, H., Makioka, S.
Predicting Attachment of the Light Verb *–suru* to Japanese Two-kanji Compound Words Using Four Aspects 73-81

Joyce, T.
Constructing a Large-Scale Database of Japanese Word Associations 82-98

Maruyama, N.
Sizuo Mizutani (1926)
The Founder of Japanese Quantitative Linguistics 99-108

Glottometrics 10, 2005, 1-15

On the History, Use, and Structure of Japanese Kanji

Joseph F. Kess[1]
University of Victoria, Victoria, Canada

Abstract: This paper traces the historical development of kanji, the Chinese characters used in the Japanese orthographic system. The paper outlines the structural principles which underlie their composition, both in respect to single kanji and to their combination in compound words. Discussion also pays attention to their usage and frequency, as well as to the various script reforms that have affected their number and deployment. Lastly, commentary on their role in the development of Japanese psycholinguistics and the relevance of this work to psychological studies of language in general is offered.

Keywords: kanji, hanzi, Chinese characters, orthography, writing systems, Japanese

1. Introduction

The Japanese writing system is one of the most intricate, most elegant, and yet most complex writing systems in the world. Japanese orthography has its origins in a millennium-old adaptation of Chinese characters and is still subject to manipulation in steps toward legislated literacy (see Kabashima, 1977; Koizumi, 1991). The current system consists of two *kana* syllabaries, which are limited to the syllabic structure of the language, and a large inventory of logographic *kanji* which represent many, but by no means all, Japanese content words. Japanese orthography has evolved into a system which freely mixes these three script types, deploying Chinese characters known as *kanji* in conjunction with the two syllabaries, *hiragana* and *katakana*.

Japanese *kanji* are the logographic symbols imported and adapted into Japanese orthography from Chinese *hanji* over several distinct historical periods. In Chinese, *hanji* are the only writing system, but in Japanese, the use of these Chinese characters came to be complemented by other orthographic script types. Because Japanese borrowed in separate and distinct historical periods, different pronunciations accompanied those kanji, thus complicating reading in the modern writing system. As a result, Japanese is a very mixed system; it not only employs the two syllabaries which match the relatively simple syllabic structure of the language, but also a large inventory of logographic kanji which can have varying pronunciations derived from either borrowed Chinese readings or native Japanese readings.

There are many characters commonly in use in written Chinese, but in Japan script reforms have settled upon 1,945 officially sanctioned characters in non-specialist writings in the popular

[1] Address correspondence to: Joseph F. Kess, Department of Linguistics and Centre for Asia-Pacific Initiatives, University of Victoria, Victoria, British Columbia, Canada, V8W 2Y2. E-MAIL: KOE@UVVM.UVIC.CA

press and government documents. However, this number swells as general texts add characters to represent personal and place names, and as publishing houses and newspaper firms incorporate their own specialized in-house inventories. The number of potential characters is far larger, as can be glimpsed in the coverage provided by specialized dictionary inventories. For example, the 12-volume *Daikanwa Jiten* lists 49,964 characters, sufficient to read even the classic Japanese texts of the past, while the more compact *Daijiten* dictionary lists 14,924, and the *Shinjigen* dictionary 9,921 kanji (see Gottlieb, 1995).

Reading written Japanese, unfortunately, does not mean simply memorizing a limited number of single characters which represent single words with single readings. Unlike Chinese or Korean, Japanese kanji can have two possible types of reading for a given character: there are *on*-readings based on the Chinese forms that came in during different episodes of borrowing from China; there may also be *kun*-readings based on the native Japanese pronunciations for the same kanji. Multiple pronunciations (or *readings*, as they are called) for a given kanji are common, and one or more *on*-readings and one *kun*-reading of a character is not uncommon.

There is some unpredictability in whether a kanji will have Chinese or Japanese readings, and how many readings there will be. Homophones are not only possible, but are in fact very common. Kanji characters with different meanings often have the same pronunciation; for example, the syllable /ki/ could be any one of the following: 木 'tree', 気 'feeling', 機 'chance', 輝 'to glow', or 期 'time period'. When kanji appear as singles, that is, in isolation, they usually take a *kun*-reading, if they have one. For example, 頭 'head' is read as *atama* in isolation, but as *too* when in compounds with other kanji, as in 頭髪 *toohatsu* 'hair of the head', 頭数 *toosuu* 'the number of head of cattle', or 馬五頭 *uma-gotoo* 'five (head of) horses'. A few kanji only have an *on*-reading, but no *kun*-reading associated with them; for example, the two kanji which form the compound word 気候 *ki-koo* 'climate, weather' only have those *on*-readings of 気 *ki* and 候 *koo*. In general, when kanji occur in compounds, there is tendency for the pronunciation to have an *on*-reading, but this is by no means a given. But the Chinese *on*-readings can also vary; a given kanji can have several *on*-readings (or even several *kun*-readings) which correspond to the period of historical borrowing from China they arrived in. For example, the kanji 頭 for 'head, chief, top, beginning', can be read /too/, /do/, /zu/, /ju/, in the Chinese way, or as /saki/, /atama/, /kashira/, /kobe/, /kaburi/, /tsumuri/ in the Japanese way, depending upon the context (Coulmas, 1989).

The result is an enormous inventory of configurations that must be mastered, stored, and recalled in order to access the written form of Japanese content words. Simplified kanji lists numbering below 2,000 have resulted from various orthography reforms, but mastering even 2,000 items is a far different task than mastering an inventory of alphabetic symbols numbering between 20 and 30. Thus, the structuring of the mental dictionary for Japanese kanji will inevitably differ from that of European languages in many ways.

2. Kanji History

Chinese orthography itself dates from the second millennium B.C. (see Boltz, 1994), and the Japanese began borrowing features of this system as early as the Fourth Century A.D. Chinese elements were actually borrowed in several distinct historical periods, but the influence of the 'Han characters' (*han-zi* in Chinese) associated with the Han Dynasty period, as well as aspects of Confucian scholarship, has been the most pervasive. As a result, this inventory of Chinese *hanzi*, now become Japanese 漢字 *kanji*, remains the most productive set of readings when one encounters Chinese characters in Japanese, and is also the most productive set in creating new words in the contemporary Japanese lexicon.

But Chinese is very different from Japanese in its word structure, and the morphological principles which favored this script type in Chinese did not occur in Japanese. One result is that the core vocabulary of Japanese developed an entirely new set of word formation procedures based on the imported Chinese readings of characters, and these procedures came to underlie both learned, technical words and even common vocabulary words which compete with native Japanese words. This situation is not unlike the flood of Latin and Greek borrowings which came into English in Early Modern English and gave rise to a new method of word coinage. If we compare Chinese and Japanese, two-character compound words can account for as much as 65% of the vocabulary in Chinese (Liu, Chuang & Wang, 1975), while compound words in Japanese might run around 40% and single kanji words 45% (Hatta & Kawakami, 1996). At first, the Chinese-based forms provided the basis of a written language in texts in Japan for only a literate few while Japanese remained the spoken, unwritten vernacular. And when Japanese did develop an orthography of its own, that writing system came to also rely heavily on the Chinese logographic system.

The Japanese went on to make up their own characters, simply by analogy with the traditional ones. Some of these have only *kun*-readings (for example, 辻 *tsuji* 'crossroads' and 峠 *tooge* 'mountain pass'), while others have only an *on*-reading (癌 *gan* 'cancer'). Still other unique Japanese creations have both *kun*- and *on*-readings, as for example, 働く *hataraku* 'work, labor' and 働 *doo* 'work, labor' (see Martin, 1972). And some of these Japanese creations reflect unique Japanese assignments of *on*-readings to create a reading for a native Japanese word; for example, 風 *fu* and 呂 *ro* are put together to form a kanji for 風呂 *furo* 'bath'.

3. Kanji Policies

The competition between reformist and conservative social philosophies has influenced language policies since the turn of the last century, but the general drift of orthographic reform has resulted in the number of commonly used kanji with official approval dipping below the 2000 mark. Official support for major orthographic revisions was particularly evident after the end of the Second World War, as the enthusiasm for democraticization reforms in the politics and education provided a platform upon which orthographic changes could also be realized. In 1946, the Ministry of Education whittled the kanji inventory down to a set of 1,850 *Tooyoo kanji* ('current/-temporary use' kanji) characters (see Seeley, 1984), and in 1981 this set was further modified to a slightly expanded set of 1,945 *Jooyoo kanji* ('common/everyday use' kanji) characters to be used in the educational system. Both sanctioned lists were official attempts to restrict the number of kanji that could be learned effectively. The *Jooyoo kanji* were not as restrictive as the *Tooyoo kanji* list, and reflected contemporary use of kanji in government and the mass media, but they did not include academic, professional, and artistic specializations (see Yasunaga, 1981). Although *Jooyoo Kanji* and *Tooyoo Kanji* do not differ dramatically in number, the new list strove to more accurately reflect actual kanji usage in modern society, as for example, kanji use in laws, ordinances, official texts, newspapers, magazines, and the media. The *Jooyoo Kanji* are presented in graded waves to elementary school students, with an expectation of mastery of the complete set by Grade 10 of Junior High School (Shimamura, 1990; Hayashi, 1991).

Although settled for now, this orthographic drama may not yet have reached its denouement, since this is one of the fronts where skirmishes take place between linguistic conservatives and language reformers. In every country, language policy is constantly evolving, reflecting changes and moods in the socio-political Zeitgeist, and Japan is no exception to this rule. However, it is worth noting that despite nostalgia on the part of some for an earlier 'golden age' of kanji erudition, surveys of kanji abilities between the early and late part of the Twentieth Century show better kanji knowledge scores on average for post-war children as a whole (Shimamura, 1997; see also Gottlieb, 1995). In sum, kanji writing is a difficult skill to acquire and maintain (see Kaiho, 1987); even college graduates make errors in their informal writing or produce cursive script in which the specific kanji cannot be deciphered (Hatta, Kawakami & Hatasa, 1997; Hatta, Kawakami & Tamaoka, 1998). There are also generational differences in respect to kanji reading and writing abilities which can be attributed to kanji familiarity, differences in schooling and educational experience, and even the fact that certain kanji become familiar because of their repeated appearance in the popular media only at a given point in time (Ukita et al., 1996). Learning how to read and write in Japanese takes up a good deal of the time spent in the educational system, and acquiring and maintaining kanji is a life-long process, at least insofar as mastery of new and specialized kanji inventories is concerned.

4. Kanji Frequencies

Any discussion of kanji usage inevitably reflects three practical facts: their absolute numbers, their frequency of usage, and their ease of access from memory. As mentioned, the number of extant kanji characters is enormous, and 'comprehensive' kanji dictionaries contain anywhere between 12,000 and 50,000 entries (Morohashi, 1989; Kindaichi, 1991). But dictionary entries do not carry the same weight as relative frequencies in understanding usage, as well as the way in which the mental dictionary is accessed. Researchers at the National Language Research Institute (now the National Institute for Japanese Language) analyzed a one-year CD ROM collection of Asahi Shimbun morning and evening runs from 1993 (Asahi Newspaper, Kinokuniya & Nichigai Associates, 1994), providing frequency data for the 24 million kanji total (Chikamatsu et al., 1998; Nozaki & Yokoyama, 1996; Yokoyama & Nozaki, 1996; Nozaki et al., 1996; Yokoyama et al., 1998). The 1000 most frequent characters account for about 95% of total kanji usage, and expanding the set to include the 1600 most frequent characters accounts for 99% of the total usage. Another 3000 characters takes in the remaining 1% of kanji usage. While this suggests that Japanese readers must know 4600 kanji to master the average newspaper, in reality knowing the most frequent 1600 will cover most of the words they will encounter. The same general picture emerges in kanji counts for magazines and periodicals, where the top 500 kanji account for around 75% of kanji usage, the top 1000 for 90%, the top 2180 for 99% (Nozaki, Yokoyama & Chikamatsu, 1997). Such insights further imply that such information, as well as the role of familiarity (see the comprehensive four-volume inventory by Amano & Kondoo, 2003), should be built into teaching and learning exercises (Nozaki et al., 1996).

Statistical data suggest a variable decline in the overall use of kanji; for example, novels written in 1900 employed text which was 39.3% kanji, while those written in 1950 employed only 27.5% (Nomura, 1984). The same decline is noted in charting the frequency of kanji usage in major Japanese newspapers published during the Meiji (1868-1911), Taisho (1912-1925), and Showa (1926-1989) eras (see Kajiwara, 1982). The use of kanji in the newspapers aimed at bureaucrats and intellectuals was at first extremely high, with kanji percentages as high as 65% of the text. At the extreme end of the continuum were government notices published in such papers, with a kanji ratio which at times went as high as 95%. High frequencies for kanji were observed even in the earlier versions of the newspapers aimed at the common people, where percentages were not uncommonly around 55%. Throughout the last century, however, kanji have somewhat decreased in Japanese newspapers, influenced by governmental decrees reducing the number of officially approved kanji, but also affected by the commercial motive of making newspapers available to a larger readership and the stylistic drift toward using more colloquial language (Kajiwara, 1982).

Even with a downward drift over time, the ratio of kanji to kana in modern Japanese newspaper text can run as high as 42% at times (see Kaiho & Nomura, 1983). But the centrality of kanji in the orthography cannot be overlooked. Research on eye movements in reading Japanese texts demonstrates the critical role of kanji in processing text longer than a single word, and for coming to grips with the cognitive processes in reading (Osaka, 1987, 1991; Osaka & Oda, 1991). Experiments using Japanese texts show that kanji-based texts are both easier and faster to read than kana-only texts, for in normal kanji-based, mixed texts, the eye skips from kanji to kanji, using them like high-profile stepping stones which stand out in relief in the field of kana. This arrangement allows the reader to organize the textual sentences in a top-down processing sweep as one jumps from kanji to kanji. The peripheral vision for kanji-based texts also allows a wider range of 6 words, as well as longer saccade patterns (Osaka, 1992), suggesting that the practice of having logographic writing incorporated into normal *kana/kanji majiribun* 'mixed text' gives rise to efficient processing strategies which are realized in shorter fixation patterns followed by longer saccades.

The type of outlet, however, has an effect on kanji percentages; for example, the highest percentage (38.2%) appears in political and economic weeklies. Weekly magazines and newspapers favored exclusively by male readers or female readers have the lowest kanji ratios, at 23.6% and 25.2%, respectively. The lower figures here also reflect the specialized, often imported vocabularies that accompany coverage of activities aimed at a specifically male or female readership.

However, when we look at the specifics of kanji frequencies in newspaper corpora, the exemplars of high frequency kanji seem to have not changed dramatically over the past thirty years (Chikamatsu et. al, 1998; Yokoyama et al., 1998). A comparison of the first 3000 frequency ranks between the 1966 and 1993 kanji frequency rankings reveals an extremely high correlation between the two usage periods, such that the overall pattern of kanji usage has not significantly changed between those points. The correlation for kanji usage between those two chronological points was r = .95 (n = 4543), even when including all the lower frequency kanji. This means that 445 kanji ranked in the top 500 kanji in the 1966 analysis of frequencies were still ranked in the top 500 in 1993. And the remaining 55 kanji were still in the top 1000 in the 1993 corpus

(Yokoyama & Nozaki, 1996). Other overviews suggest similarly close correlations for the 1,945 basic *Jooyoo Kanji* between those time periods (Tamaoka and Makioka, 2004; Tamaoka et al., 2002). Where there are shifts in kanji frequency, it is typically in the low frequency group, suggesting that high frequency kanji tend to stay in as high frequency kanji, while low-frequency kanji are more likely to shift in respect to their usage over time (Nozaki, Yokoyama & Chikamatsu, 1997).

Lastly, the modern availability of word processors, electronic dictionaries, and personal computers, and their impact on current kanji awareness, cannot be overlooked here. Such ready access has made kanji easily available through facile input strategies, and helps the average reader cope with the fact that commercial publishers of magazines, books, and newspapers often maintain larger inventories of proper names, place names, and specialized or technical vocabulary for their specialized uses. Indeed, it may come to pass that the question of how many kanji there are will no longer matter as a practical question, as increasing numbers of users simply plug into the more and more sophisticated mechanical devices. There may even come a time in Japan when there will be two kanji sets in actual practice: one used in reading comprehension, a 'read-only' set, while the other set will be a 'write-only' set, the set that literate Japanese will be required to re-produce in actual handwriting tasks (see Takata, 1991; Kess & Miyamoto, 2001).

5. Kanji Structure

The compositional principles that form Chinese characters were borrowed and then kept in their original design features. Thus, the production of Chinese characters, simple or complex, generally proceeds from left-to-right, top-to-bottom, and outside-to-inside. Chinese characters are traditionally classified into four major groupings; over three-quarters are phonetic-semantic in their composition, while ideographic, diagrammatic, and compound semantic kanji show smaller percentages (Ito, 1979; Sato, 1973; Saito, Inoue & Nomura, 1979). Ideographic kanji picture a stylized version of some object or concept, as in the peaks and valleys suggested by 山 *yama* 'mountain'. Diagrammatic kanji portray a logical, geometric, or conceptual relationship by the way the character is arranged, as in the up-down spatial relationships shown by 上 *ue* 'up' and 下 *sita* 'down'. Compound-semantic kanji combine two simple characters to suggest a transparent meaning based on the additive value of the individual units, as in 明るい *akarui* 'light, bright' based on the component parts of 日 *hi* 'sun' and 月 *tsuki* 'moon'.

But such ideographic, diagrammatic, and compound-semantic kanji represent no more than 20% of the kanji inventory. Approximately 80% of kanji are phonetic-semantic kanji, and often contain a semantic 'radical' (*hen*) on the left-hand side of the phonetic-semantic kanji, and/or a phonetic radical (*tsukuri*) on the right-hand side. The semantic radical supposedly gives a rough approximation of where the kanji fits in the semantic scheme of things, while the phonetic radical suggests a Chinese reading for the kanji. An example of a semantic 'radical' (*hen*) might be the left-hand side of the following kanji 晴, which has the meaning of 'sun, day', while an example of the phonetic radical (*tsukuri*) might be the right-hand side of the same kanji 晴, which has the same reading as that piece when it appears as a separate stand-alone kanji within another kanji compound word, namely, 青 *sei* 'blue, green', as in *sei-nen*, 'a youth'.

But these are only two of the seven categories that phonetic-semantic compounds are divided into, according to the compositional elements that can enter into the construction process for kanji. The other five elements, and their placement, are as follows: the *kanmuri* 'crown' in 草; the *ashi* 'leg' in 先; the *kamae* 'structure' in 国; the *tare* 'hanging' in 庭; and the *nyoo* 'entering' in 道 (Morton & Sasanuma, 1984; Tamaoka, 1991; Kaiho & Nomura, 1983).

The presentation of a Chinese character is meant to fit within an idealized box-like configuration which dedicates the same relative space to the individual characters in running text. The expectation of equidimensionality implies that each character, regardless of composition or number of strokes, occupies the same box-like space of an identical size. Kanji characters can be simple or they can be complex. That is, a simple character is one that can stand alone, and like the original Chinese, there are many of these simple basic characters. Some of these may also be used as a 'radical' which forms part of a single larger and more complex character in the way that the original Chinese historically deployed 214 such radicals to reflect supposedly semantic categories to classify kanji. It may appear in its original, but reduced shape. For example, 木 can stand alone as *ki*, the word for 'tree', or it can enter into the formation of larger, complex characters such as tree names (松, 椿, 梅, 杉 *matsu*, *tsubaki*, *ume*, *sugi* 'pine, camelia, plum, Cryptomeria cedar') or 'woody' by-products (枝, 根, 板 *eda*, *ne*, *ita* 'branch, root, board', and so forth (see Morton & Sasanuma, 1984; Leong & Tamaoka 1995). The radical may appear in a somewhat altered shape, but such radicals are usually not stand-alone radicals. For example, the 'water' radical 氵 is different from the character for 'water' 水 *mizu*, though it is an abbreviated form of the full character. This is the left-hand component in characters such as 池, 河, 波, 海, 泳ぐ, 酒, 涙, 港, 深い *ike*, *kawa*, *nami*, *umi*, *oyo-gu*, *sake*, *yu*, *namida*, *minato*, *fuka-i* 'pond, river, wave, sea, swim, sake, hot water, tear, port, deep', but you will never see this abbreviated component stand alone.

In the abstract, the system seems cognitively ideal, but the reality is that this kind of semantic categorization has severe limitations in reading modern Japanese. In an age when pillows and bridges were made of wood, such assignments of the radicals made sense; but circumstances change, and the complex characters 橋, 枕, 机, 村 for *hashi*, *makura*, *tsukue*, *mura* 'bridge, pillow, desk, village' don't seem transparently woody anymore. In some cases, it is not always obvious which element is the radical in a complex character (Chen, Allport, & Marshall, 1996), and in other cases, the same homophonic reading conveys the same semantic assignment, but a completely different kanji is used (contrast 作 and 造, both *tsukuru* 'to make, manufacture, produce').

Orthographic simplicity versus orthographic complexity is not necessarily reflected in processing difficulty. In fact, orthographic complexity in terms of a large number of strokes in certain kanji may facilitate processing by underscoring the uniqueness of a kanji configuration. High frequency kanji are typically read faster, because of the likelihood of familiarity; but when frequency is held constant, more complex kanji are easier to read than less complex kanji, simply because the distinctiveness of a kanji like 龍 *ryuu* 'dragon' in terms of strokes facilitates reading it. For kanji of 13 strokes or less, difficulty in kanji processing increases proportionally to the number of strokes; however, after this point, increase in the number of strokes actually facilitates kanji processing (Kaiho, 1979), probably because the kanji is taken as a chunk with its visual signature, thus eliciting shorter recognition times.

6. Phonological vs. Semantic Information in Kanji Words

The majority of Chinese logographs, the source from which Japanese kanji are derived, are phonographs (Wang, 1981) which typically exhibit two possible constituent parts. In traditional terms, there is a radical or *signific*, usually on the left side of the character, which refers to meaning, and on the right side of the character, there is often a *phonetic* which refers to the pronunciation of the character (Chen & Yuen, 1991). Although the characters imported from China into Japanese often retained the physical shape of these phonetic radicals, such phonetic radicals are neither as reliable nor as useful in reading Japanese kanji as they are in reading Chinese hanzi. Although various figures have been given for these phonetic radicals (Ito, 1979; Saito, 1981), an inventory (Saito, Kawakami & Masuda, 1995) demonstrates that a third of the complex kanji in the JIS/Japanese Industrial Standard set of 2965 kanji characters have the same pronunciation as their right-hand phonetic radicals. This seems to suggest that the phonological information embedded in some Japanese kanji by virtue of such phonetic radicals are an integral part of the identity of those lexical items. For example, the right hand component in the following kanji carries the pronunciation of the kanji: /kai/, as in 悔 'to regret' and 海 'sea'; /sei/, as in 晴 'clear weather' and 清 'clear water'; /sen/, as in 銭 'money' and 浅 'shallow'; and /shin/, as in 侵 'to invade' and 浸 'to soak'; respectively (see Morton & Sasanuma, 1984; Kess & Miyamoto, 1999).

The traditional belief has been that kanji represent words, not sounds, and that the semantic radicals cue meanings. It is true that Japanese readers can use the cues provided by the component parts of kanji to guess at the meaning of a new kanji when those cues are reliable, and a good example of this is seen in the way skilled Japanese readers infer the meaning of unfamiliar technical words (Hatano, Kuhara & Akiyama, 1981), much as skilled English readers do with Latin- or Greek-derived technical terms. However, the pervasive belief that the semantic radicals provide a built-in conceptual categorization system simply is not tenable. The architecture of Japanese kanji is more complicated than that, for the embedded components in complex characters are more variable than left- versus right-hand placement. The compositional elements that generate kanji can be found to the left, above, below, or even inside the stem, as well on the left- or right-hand side. One recent inventory calculated the configurational possibilities for the JIS set of 2965 kanji, and found that more than half of these (1668 of 2965) exhibited components on both the left and right sides (Saito et al., 1995, 1997; Saito, Kawakami, & Masuda 1998). Of these, 760 groupings cluster around right-hand *tsukuri* radicals, while only 247 groupings cluster around the left-hand *hen* radicals. The processing consequence of this asymmetry in this basic set of complex kanji is that if the right-hand radical is correctly identified, the field of choice narrows to an average of 2.2, that is, 1668 divided by 760. The field of choice is wider for the left-hand radical cluster, for even if it is correctly identified, the average here still works out to 6.8, or 1668 divided by 247. Obviously, the two placements offer different degrees of information, with the right-hand *tsukuri* offering a better set of clues in respect to reducing the potential range of possible kanji groupings. If anything, such insights underscore the contribution of other elements besides the traditionally cited left-hand *hen* semantic radical.

And just because an element is considered a radical for historical reasons does not mean it is commonly used. Martin (1972) once observed that although there are roughly 200 radicals, more than half of all kanji include one of the 20 most frequent radicals. Saito, Kawakami, and Masuda (1995) point out that the average subject in a psycholinguistic experiment, let alone the average reader, relies more on configurational considerations than on historically-derived etymological considerations. In fact, they extracted 857 basic radicals from the JIS first set, by focussing on their configurational similarity, a figure different than the 214 historical categories. Of these, the majority (610 or 71%) appear on the right-hand side of the complex characters; only 97 (or 11%) appear in the left-hand position. And they found some radicals could appear on either the left- or the right-hand side; this moveable set of 'free-floating radicals' was actually larger than the set of left-hand radicals, 150 in number or 18% of the total 857 (Saito et al., 1997).

But even these numbers do not tell the final story, for the three radical types have different collocational possibilities with other components in creating complex kanji of this left-right kind. The right-hand *tsukuri* radicals combine with an average of 2.0 left hand components, the left-hand *hen* radicals with an average of 8.6 right-hand components, and the floating radicals with an average 5.5 components on the right side and 2.9 on the left side. Amazingly, when Saito et al. (1997) tested subjects to see whether they could estimate the number of kanji characters that could be formed with a particular radical, they found that subjects could correctly evaluate the number of possible characters as a function of the number of possible collocations with that radical. Not only did they find that subjects could indeed evaluate the number of possible characters as a function of the number of possible collocations with that radical, but they also found that the subjects were better in estimating that number for left-hand radicals than with right hand radicals. This might of course reflect kinetic storage considerations that arise from countless practice sessions in which writing kanji is mastered through left-to-right, top-to-bottom practice sequences. In a sense, there appears to be a contrast between two aspects of subjects' knowledge, recognition versus recall. In recognition, the right-hand radical *tsukuri* are more informative, because they collocate with a smaller number of companions on the left side to make a complete kanji. This makes sense when one considers that there being fewer candidates facilitates recognition because there are fewer to sort through. On the other hand, in recall, the left-hand radical *hen* is more useful in evaluating its number of companions in a specific kanji family. Whether this is because the left-hand radicals recall a larger number of kanji with a more centralized semantic pivot, or because of kinetic associations in writing practice, is uncertain. But the fact is that the two component types each play a role in the knowledge base subjects have about complex kanji, and which they call into play in recall and recognition tasks.

We might compromise by admitting that the real role of the 'semantic' radicals is not so semantic in recognition of many complex characters; the recognition units which contribute to access procedures are simply not isomorphic with the semantic radicals as traditionally conceived. Upper and lower configurations, as well as repeated or parallel elements, are also critical. Where the semantic radicals may be of particular significance is in those cases where a kanji is unfamiliar or unknown, and where one searches for any and all clues that may give some indication of its range and identity.

7. Compound Kanji

Many common words in modern Japanese are not represented by a single character, but are compound words composed of two or more characters. Some estimates suggest that compound kanji words comprise at least 50% of most dictionaries, and Yokosawa and Umeda (1988) report figures as high as 70%. These compound words, or *jukugo* 熟語, are a composite of two to four kanji and usually carry an *on*-reading. There are many such words in common use, words such as 緊張, 誕生日, 新幹線, 証券会社 *kinchoo*, *tanjoobi*, *shinkansen*, *shookengaisha*. 'stress, birthday, bullet train, securities firm'.

Some compound kanji have an unexpected, irregular *kun*-reading which portrays some vague semantic identity in their reading. For example, 土産 *miyage* 'souvenir' is derived from 'soil' and 'produce', and attempts to convey the flavor of 'a local product brought back as a souvenir' in its choice of characters; so does 田舎 *inaka* 'countryside', with its characters for 'rice field' and 'reside'. But these compounds definitely require mastery as unique orthographic events. Other kanji shift between the *on* and the *kun*-readings in compounds, with the correct reading conditioned by the other kanji in the same compound. For example, 父親 *chichi-oya* 'father' and 両親 *ryo-shin* 'parents' both contain the kanji 親, but it has a *kun*-reading of *oya* in the first compound and an *on*-reading of *shin* in the second.

Not surprisingly, some experimental reports have concluded that lexical access for Japanese words written in kanji employs the whole word as the basic element in searching the mental lexicon, and not the kanji character units or their analyzable parts (for example, Sakuma, Ito & Sasanuma, 1989). It has also been suggested that the whole word as such is the most resistant to neurological impairment through damage or disease (Sasanuma, 1992). But other reports offer support for the role of the first kanji in the storage and retrieval of kanji compounds from memory (Tamaoka & Takahashi, 1999; Yamada & Kayamoto, 1998). Here the frequency of the first kanji, as well as of the whole kanji compound itself, is a good predictor of ease of word recognition and even priming results (Hirose, 1992). Thus, we are left with two different possibilities as to the relevant recognition units for accessing kanji compounds. There is evidence that the kanji compounds are themselves the recognition units as whole words, but there is also evidence that the individual component kanji are the relevant recognition units. Frequency is, of course, a crucial factor in both arguments because frequency is after all the key to how often such judgments are made, and is also reflected in the asymmetrical nature of lexical productivity in the creation of two-kanji compound words (see Tamaoka & Altmann, 2004). High frequency kanji in the first position in kanji compounds facilitate access for naming, because this is where access procedures are first initiated, and successful naming will depend upon the accuracy of matching phonological information to the kanji embedded in the compound. But the number of possible readings must also be addressed before the final correct reading comes up for an accurate naming response. High frequency kanji in the second position in kanji compounds facilitate access for lexical decision tasks because this is the end point at which the complete information for real kanji is finalized. At this point, one is deciding whether a given compound is in fact a real kanji compound, and that decision can only be taken once the final pieces of the processing puzzle are in place. The frequency factor is not a contradictory element after all, because the high frequency of an individual kanji also reflects its frequency of appearance in kanji compounds. The final arbiter of successful naming responses is in fact the contextual level for kanji compounds, for this is where final decisions have to be made for the correct reading of a particular kanji, and thus the correct pronunciation of the compound. But accuracy in deciding whether a kanji compound is in fact a real compound word inevitably reflects knowledge about how often this configuration appears and whether a lexical address for this configuration is commonly accessed.

8. Relevance to Psychological Studies of Language

The history of Japanese psycholinguistics can be essentially linked to psychological interests in the long-standing Japanese interest in the specific requirements of *kanji* processing (see Kess & Miyamoto, 1999), and beyond that, to the even longer Japanese tradition of meditation on the nature of their language and the unique status it is assumed to have among the world's languages (see Kess and Miyamoto, 1994). More recently, the computer age has sharpened that focus with the practical demand for computational devices which can read, parse, and translate between languages. This explosion of interest has gone hand in hand with the development of the discipline of *cognitive science* and its exploration of how the mind deals with linguistic information. One important area of research within cognitive science focusses on written language, and attempts to explain word recognition, priming and association, the mental lexicon, and indeed, the entire reading process. An understanding of the cognitive mechanisms by which humans process linguistic information through the printed medium in Japanese (Leong & Tamaoka, 1998; Ukita et al., 1996; Yokoyama, 1997) will add much to these research concerns and will unquestionably inform our grasp of the universals of language and language processing.

REFERENCES

Amano, S., & Kondoo, T. (2003). *Nihongo no goi tokusei Dai 1-ki* [*Lexical properties of Japanese, Vol. 1*], CD-ROM edition. NTT Komyunieeshon Kagaku Kiso Kenkyujo. Tokyo: Sanseidoo.

Asahi Shimbun. (1994). CD ROM, 1993 Editions of the *Asahi Shimbun*. Tokyo: Kinokuniya & Nichigai Associates.

Boltz, W. G. (1994).*The origin and early development of the Chinese writing system*. New Haven: American Oriental Society.

Chen, M. J., & Yuen, J. C. (1991). Effects of Pinyin and script type on verbal processing: Comparisons of the China, Taiwan, and Hong Kong experience. *International Journal of Behavioral Development*, *14*, 429-448.

Chen, Y. P., Allport, D. A., & J. C. Marshall. (1996). What are the functional orthographic units in chinese word recognition: The stroke or the stroke pattern? *Quarterly Journal of Experimental Psychology, 49A*(4), 1024-1043.

Chikamatsu, N., Yokoyama, S., Nozaki, H., Long, E, Sasahara, H., & Fukuda, S. (1998). Development of a Japanese kanji character frequency list. *Proceedings of the 12th International Unicode Conference* (*Part 1*), held in Tokyo, Japan. April, 1998.

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

Gottlieb, N. (1995). *Kanji politics: Language policy and Japanese script*. London and New York: Kegan Paul International.

Hatano, G., Kuhara, K., & Akiyama, M. (1981). Kanji help readers of Japanese infer the meaning of unfamiliar words. *The Quarterly Newsletter of the Laboratory of Comparative Human Cognition, 3*(2), 30-33.

Hatta, T., & Kawakami, A. (1996). Lexical and naming processes of non-prototypical kanji: Evidence of the component parts activation. *Asia Pacific Journal of Speech, Language and Hearing*, *1*, 55-64.

Hatta, T., Kawakami, A., & Tamaoka, K. (1998). Writing errors in Japanese kanji: A study with Japanese students and foreign learners of Japanese. *Reading and Writing*, *10*, 457-470.

Hatta, T., Kawakami, A., & Hatasa, Y. (1997). Kanji writing errors in Japanese college students and American Japanese students. In H-C. Chen (ed.), *Cognitive processing of Chinese and related Asian languages* (pp. 401-416). Hong Kong: The Chinese University Press.

Hayashi, O. (1991). Kanji no kaiho to moji kyoiku [Kanji liberation and character education]. *Gengo, 20*,44-49.

Hirose, H. (1992). Jukugo no ninchi katei ni kansuru kenkyu: Puraimingu-ho ni yoru kento [Using the priming paradigm to investigate word recognition for kanji compound words]. *Shinrigaku Kenkyu, 63*, 303-309.

Ito, K. (1979). Keisei moji to kanji shido [Character formation and kanji teaching]. *Gengo Seikatsu* [*Language Life*], *326*,68-79.

Kabashima, T. (1977). Kanji kara romaji made [From kanji to romaji]. In A. Sakakura (ed.), *Nihongo no Rekishi* [*The History of the Japanese Language*] (pp. 114-152). Tokyo: Taishukan.

Kaiho, H. (1979). Kanji joho shori kisei o megutte [Information processing for kanji]. Keiryo Kokugogaku, 11, 331-340.

Kaiho, H. (1987). Ningen ni okeru Kanji Joho Shori [Kanji Information Processing in Humans]. *Keiryo Kokugogaku to Nihongo Shori* [*Mathematical Linguistics and Japanese Language Processing*] (pp. 49-61). Tokyo: Akiyama Shoten.

Kaiho, H., & Nomura, Y. (1983). *Kanji joho shori no shinrigaku* [*The psychology of kanji processing*. Tokyo: Kyoiku Shuppan.

Kajiwara, K. (1982). Shinbun no kanji ganyuritsu no hensen: Meiji, Taisho, Showa o tsujite [Shifts in occurrence rate for newspaper kanji: Through the Meiji, Taisho, and Showa Eras]. *Kokuritsu Kokugo Kenkyusho Hokoku* [*Bulletin of the National Language Research Institute*], *71*, 209-236.

Kess, J. F., & Miyamoto, T. (1994). *Japanese psycholinguistics: A classified and annotated research bibliography*. Amsterdam: John Benjamins Publishers.

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

Kess, J. F., & Miyamoto, T. (2001). Kanji knowledge as Read-only vs. Write-only: The effect of the computer age. In M. Nakamura (ed.), *Japan in the Global Age* (pp. 175-187). Centre for Japanese Research, University of British Columbia.

Kindaichi, K. (1991). *Shineikai kokugo jiten* [*New Japanese word dictionary*]. Tokyo: Sanseido.

Koizumi, T. (1991). Nihon ni okeru moji seisaku no rekishi [The history of character policy in Japan]. *Gengo, 20*, 38-43.

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*: *An Interdisciplinary Journal, 7*, 377-393.

Leong, C. K. & Tamaoka, K. (1998). *Cognitive processing of the Chinese and Japanese languages*. Boston: Kluwer Academic Publishers.

Liu, I. M., Chuang C. J., & Wang, S. C. (1975). *Frequency Count of 40, 000 Chinese Words*. Taipei: Lucky Books.

Martin, S. E. (1972). Non-alphabetic writing systems: some observations. In J. F. Kavanaugh & I. G. Mattingly (eds.), *Language by Ear and by Eye* (pp. 81-102). Cambridge, MA: MIT Press.

Morohashi, T. (1989). *Daikanwa Jiten* [*Japanese kanji character dictionary*], Vol. 1-12. Tokyo: Taishukan.

Morton, J., & Sasanuma, S. (1984). Lexical access in Japanese. In L. Henderson (ed.), *Orthographies and Reading: Perspectives From Cognitive Psychology, Neuropsychology, and Linguistics* (pp. 25-42). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nomura, M. (1984). Kanji no tokusei o hakaru: kanji no keiryo kokugogaku [Measuring the characteristics of kanji: Mathematical linguistics and kanji]. In H. Kaiho (ed.), *Kanji o kagaku suru* [*Making kanji scientific*] (pp. 1-34). Tokyo: Yuhikaku.

Nozaki, H. & Yokoyama, S. (1996). Shinbun to zasshi ni okeru kanji shiyohindo no bunseki [The analysis of kanji frequency in newspapers and magazines]. *Proceedings of the 24th Meeting of the Behaviometric Society of Japa*, 266-267.

Nozaki, H., Yokoyama, S., Isomoto, Y., & Yoneda, J. (1996). Moji shiyo ni kansuru keiryoteki kenkyu: Nihongo kyoiku shien no kanten kara [A study of character frequency: From the point of view of Japanese language education]. *Nihon Kyoiku Kogaku Zasshi* [*Journal of Japanese Educational Technology, 20*(3),141-149.

Nozaki, H., Yokoyama, S., & Chikamatsu, N. (1997). Shinbun to zasshi ni okeru kanji shiyo hindo no bunseki [The analysis of kanji frequency in newspapers and magazines]. *Nihon Kyoiku Kogaku Zasshi* [*Journal of Educational Technology, 21*, 21-24.

Osaka, N. (1987). Effect of peripheral visual field size upon eye movements during Japanese text processing. In J. K. O'Regan & A. Levy-Schoen (eds.), Eye movements from physiology to cognition: Selected/Edited proceedings of the third European conference on eye movements, Dourdan, France, September 1985 (pp. 421-429). Amsterdam: North-Holland.

Osaka, N. (1991). Yomi no seishin butsurigaku: Yuko shiya no yakuwari o chushin ni [Psychophysics of Japanese text reading: Role of effective visual field]. *Tetsugaku Kenkyu* [*The Journal of Philosophical Studies*], *48*, 588-612.

Osaka, N. (1992). Size of saccade and fixation duration of eye movements during reading: psychophysics of Japanese text processing. *Journal of the Optical Society of America*, *9*(1), 5-13.

Osaka, N. & Oda, K. (1991). Effective visual field size necessary for vertical reading during Japanese text processing. *Bulletin of the Psychonomic Society, 29*, 345-347.

Saito, H. (1981). Toward Comparative Studies of Reading Kanji and Kana. *The Quarterly Newsletter of the Laboratory of Comparative Human Cognition,3*(2), 33-36.

Saito, H., Inoue, M., & Nomura, Y. (1979). Information processing of kanji (Chinese characters) and kana (Japanese characters): The close relationship among graphemic, phonemic, and semantic aspects. *Psychologia, 22*, 195-206.

Saito, H., Kawakami, M., & Masuda, H. (1995). Kanji koosei ni okeru buhin (bushsu) no shutsugen hindohyo [Frequencies for semantic and phonetic radica components in complex kanji radicals]. *Bulletin of the Nagoya University Graduate School of Informatics and Science, 1*,113-134.

Saito, H., Kawakami, M., & Masuda, H. (1998). Form and sound similairity effects in kanji recognition. *Reading and Writing, 10*, 323-357.

Saito, H., Kawakami, M., Masuda, H., **& Flores d'Arcais G. B**. (1995). Conjunctionability effects on radical migration with kanji characters. Paper presented at the *Seventh International Conference on the Cognitive Processing of Chinese and other Asian Languages*. Hong Kong.

Saito, H., Kawakami, M., Masuda, H., **& Flores d'Arcais G. B**. (1997). Contributions of radical components to kanji character recognition and recall. In H. C. Chen (ed.), *Cognitive processing of chinese and related asian languages* (pp.109-140). Hong Kong: The Chinese University Press.

Sakuma, N., Ito, M., & Sasanuma, S. (1989). Puraimingu paradaimu ni yoru kanji tango no ninchi yunitto no kento [Recognition units of kanji words: Priming effects on kanji recognition. *The Japanese Journal of Psychology, 60*, 1-8.

Sasanuma, S. (1992). Neuropsychology of reading: Universal and language-specific features of reading impairment. In P. Bertelson, P. Eelen & G. d'Ydewalle (eds.), *International Perspectives on Psychological Science. Vol. 1*: *Leading Themes* (pp. 105-125). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sato, K. (1973). *Japanese ideomatic letter kanji: Design of Japanese letters*. Volume 5. Tokyo: Maruzen.

Seeley, C. (1984). The Japanese script since 1900. In C. Seeley (ed.), *Aspects of the Japanese writing system.* Special Issue of *Visible Language, Visible Language*, *18*(3), 267-302.

Shimamura, N. (1990). Kanji no shutokuritsu: Haito kanji ni yoru chigai [Acquisition rates of kanji by school children: Differences in kanji by grade level]. *Keiryo Kokugogaku*, *17*, 273-279.

Shimamura, N. (1997). Senzen no kodomo no kanji [Kanji ability of pre-war children]. *Dokusho Kagaku*, *41*, 124-128.

Takata T. (1991). Kanji no Unmei [The Fate of Kanji]. *Gengo, 20,* 52-58.

Tamaoka, K. (1991). Psycholinguistic nature of the Japanese orthography. *Studies in Language and Literature, 11,* 49-82. Matsuyama: Matsuyama University.

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

Tamaoka, K., Kirsner, K., Yanase, Y., Miyaoka, Y., & Kawakami, M. (2002). A Web-accessible database of characteristics of the 1,945 basic Japanese kanji. *Behavior Research Methods, Instruments, & Computers 34*, 260-275.

Tamaoka, K., & Makioka, S. (2004). New figures for a Web-accessible database of the 1,945 basic Japanese kanji, fourth edition. *Behavior Research Methods, Instruments, & Computers, 36*(3), 548-558.

Tamaoka, K.. & Takahashi, N. (1999). Kanji niji jyukugo no shoji kodo ni okeru goi shiyo hindo oyobi shojiteki fukuzatsusei no eikyo [The effects of word frequency and orthographic complexity on the writing process of Japanese two-morpheme compound words]. *Japanese Journal of Psychology, 70*(1), 45-50.

Ukita, J., Sugishima, I., Minagawa, M., Inoue, M., & Kashu, K. (1996). *Nihongo no hyoki keitai ni kansuru shinrigakuteki kenkyu* [*Psychological research on orthographic forms in Japanese*]. *Psychological Monograph No. 25*. Tokyo: Japanese Psychological Association.

Wang, W. S. Y. (1981). Language structure and optimal orthography. In O. J. L. Tzeng & H. Singer (eds.), *Perception of Print: Reading Research in Experimental Psychology* (pp. 223-236). Hillsdale, NJ: Lawrence Erlbaum.

Yamada, J., & Kayamoto, Y. (1998). Valency, secondary frequency, and lexical access: A Japanese study. *Applied Psycholinguistics, 19,* 87-97.

Yasunaga, M. (1981). Joyo Kanjihyo ga umareru made [Until Joyo Kanji were born]. *Gengo Seikatsu, 355*, 24-31.

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

Yokoyama, S. (1997). *Hyoki to kioku* [*Orthography and memory*]. *Psychological Monograph No. 26*. Tokyo: Japanese Psychological Association.

Yokoyama, S. & Nozaki, H. (1996). Asahi Shimbun CD-ROM ni yoru kanji hindo kijunhyo no sakusei to suryo bunseki [The statistical analysis of kanji frequency in the Asahi Shinbun]. *Simposiumu* "*Jinmon Kagaku ni okeru Suryoteki Bunseki*" [*The Symposium on Quantitative Analysis in the Human Sciences, March 11, 1996*]. Statistics and Mathematics Research Institute, Ministry of Education. pp. 1-4.

Yokoyama, S., Sasahara, H., Nozaki, H. & Long, E. (1998). *Shimbun denshi media nokanji: Asahi Shimbun CD-ROM ni yoru kanji hindohyo* [*Kanji in the electronic newspaper media: Kanji frequency tables from the Asahi Newspaper CD-ROM*]. *Kokuritsu Kokugo Kenkyujo purojekuto sensho 1* [*National Language Research Institute project, Special publication 1*]. Tokyo: Sanseido.

Glottometrics 10, 2005, 16-29

Mathematical Modelling for Japanese Kanji Strokes in Relation to Frequency, Asymmetry and Readings[1]

Katsuo Tamaoka, Hiroshima, Japan[2]
Gabriel Altmann, Lüdenscheid, Germany

Abstract: The present study investigates the relationship between of Japanese kanji strokes and their printed-frequencies of occurrence, compositional asymmetry and kanji multiple readings. First, distributions of kanji strokes in both samples of the 1,945 basic kanji and of 6,355 kanji appearing in the *Asashi Newspaper* published between 1985 and 1998 followed a negative hypergeometric distribution as demonstrated by Figure 1. The distribution of strokes of the 1,945 kanji with their printed-frequencies is rather rhapsodic, as shown in Figure 2, but a rough-fitting model is drawn in Figure 3. Mathematical modelling for kanji strokes with lexical compositional asymmetry reveals the interesting tendency of *regressive compounding*; that is, that the greater the number of strokes in a kanji, the more it tends to produce two-kanji compound words by adding a kanji on the right side of the target kanji, as shown in Figure 4. A kanji may often have multiple readings; this study also examines the number of readings in relation to the number of kanji strokes. As shown in Figure 6, the greater the number of kanji strokes, the fewer the number of readings. In other words, the more visually complex the kanji is, the more specialised its reading becomes. As such, kanji strokes, as one of the central characteristics of kanji, are closely related to other properties such as frequency, asymmetry and readings. The present study uses mathematical modelling to indicate these relations.

Key words: mathematical modelling, kanji strokes, kanji frequency, lexical compositional asymmetry, multiple readings, regressive compounding, self-regulation cycle

1. A self-regulation cycle of linguistic properties

A single property of a lingustic entity can hardly be isolated from its other properties. The associated properties make up a self-regulating cycle in which one change creates a chain of alterations. The change comes gradually, and its reactions follow with some delay. These series of changes, as a whole, maintain a balanced 'equilibrium' state. Although this conjecture is considered to be applicable to linguistic properties, it is still a preliminary hypothesis which must be validated empirically. The present study on Japanese kanji strokes attempts to explore this conjecture.

A kanji represents the smallest unit of meaning, the 'morpheme'. Unlike alphabetic languages, Japanese kanji are drawn in horizontal and vertical dimensions (for details on kanji in general, see Kaiho & Nomura, 1983; Morton & Sasanuma, 1984; Tamaoka, 1991, 2003, 2005a, 2005b; Tamaoka, Kirsner, Yanase, Miyaoka & Kawakami, 2002; Tamaoka & Makioka, 2004; Tamaoka & Yamada, 2000; Wydell, Butterworth & Patterson, 1995; Wydell, Patterson & Humphreys, 1993). Among the 1,945 basic Japanese kanji, the simplest kanji is

[1] This study is supported by the Research Institute of Science and Technology for Society (RISTEX) at the Japan Science and Technology Corporation (JST), Kawaguchi, Japan.

[2] Address correspondence to: Katsuo Tamaoka, International Student Center, Hiroshima University 1-1, 1-chome, Higashihiroshima, Japan 739-8524. E-mail: ktamaoka@hiroshima-u.ac.jp

drawn as a single horizontal line: − meaning 'one'; whereas the most complex kanji requires 23 strokes to write: 鑑 meaning 'appreciate'. As such, the number of strokes roughly indicates the visual complexity of a kanji. The question then is whether or not the number of kanji strokes (or kanji visual complexity) has some relation to other properties of a kanji. In the present study, three kanji properties are examined as candidates for possibly being related to the number of kanji strokes; namely, the printed-frequency of a kanji (hereafter, refer to 'kanji frequency'), asymmetry for two-kanji compound word production, and multiple readings.

2. Mathematical modeling for kanji strokes

The lexical database sampled from the *Asahi Newspaper* published between 1985 and 1998 (Amano & Kondo, 2000) contains a total of 6,355 different kanji. Kobayashi (1981) reports that the 1,000 most frequently used kanji represent 93.9 percent of all kanji printed in newspapers, and 90.0 percent of all kanji appearing in magazines. Knowledge of 2,000 kanji covers 99.6 percent of those found in newspapers and 98.6 percent in magazines. Knowledge of 3,000 kanji covers about 99.9 percent of kanji used in newspapers and magazines. Considering these figures, 6,355 kanji must encompass almost all the kanji seen in Japanese written texts. Then, the question is whether the number of kanji strokes has some influence on the kanji frequency.

As can be seen in Table 1, the distribution of the number of strokes is very regular. Using the possibilities furnished by the Wimmer-Altmann theory (for details, see Wimmer, Altmann, 2005) the 1-displaced negative hypergeometric distribution was chosen:

$$(1) \qquad P_x = \frac{\binom{M+x-2}{x-1}\binom{K-M+n-x}{n-x+1}}{\binom{K+n-1}{n}}, \qquad x = 1,2,3,\ldots,n+1.$$

This formula describes the distribution of stroke numbers. Using the kanji database of 1,945 basic kanji (Tamaoka et al., 2002; Tamaoka & Makioka, 2004), the distribution of stroke numbers is obtained as seen with the above 1-displaced negative hypergeometric distribution with parameters $K = 11.6517$, $M = 4.9490$, $n = 22$, DF = 18, $X^2 = 22.53$, P = 0.21, $C = 0.01$. If the distribution is stable, it is automatically assumed that a larger sample must display the same behaviour. As can be seen in Table 1, iterative fitting to the inventory of 6,355 kanji shows that the distribution of strokes is, in fact, relatively stable and the fit is satisfactory. As the chi-square grows with increasing sample size, the values obtained for the larger sample were $X^2 = 45.63$ with 26 DF, corresponding to $P = 0.01$ and $C = 0.0072$. The result is displayed graphically in Figure 1. The increasing values of the parameters indicate that perhaps a limiting distribution would yield better results (e.g., Poisson or negative binomial) but an investigation of this issue will be postponed until data from other writing systems such as Chinese, Assyrian, Sumerian, and so on has also been examined. However, the chi-square value decreases for n between 29 and 40, and it increases again above 40. These figures indicate that the negative hypergeometric model is a very stable estimator for this purpose.

There exist different kinds of frequency counts: (a) using individual texts, (b) a list which represents a part of a dictionary, (c) the complete dictionary, (d) a frequency dictionary and a corpus representing a mixture of texts and being identical with a frequency dictionary. Here, we analyze case (b).

Table 1
The negative hypergeometric model for the distribution of strokes in 6,355 kanji

X (Strokes)	f_x (Number of kanji)	NP_x
1	5	3.61
2	29	17.46
3	57	48.11
4	109	98.91
5	146	168.45
6	202	250.76
7	312	336.85
8	436	416.84
9	469	481.90
10	539	525.74
11	579	545.25
12	577	540.53
13	537	514.34
14	432	471.26
15	427	416.76
16	377	356.37
17	300	295.05
18	205	236.71
19	168	184.12
20	136	138.88
21	110	101.57
22	67	72.00
23	47	49.43
24	38	32.84
25	20	21.08
26	12	13.06
27	8	7.79
28	6	4.46
29	3	2.44
30	2	2.43
	K = 24.6877, M = 6.8335, n = 41, DF = 26 X^2 = 45.63 , P = 0.0100, C = 0.0072	

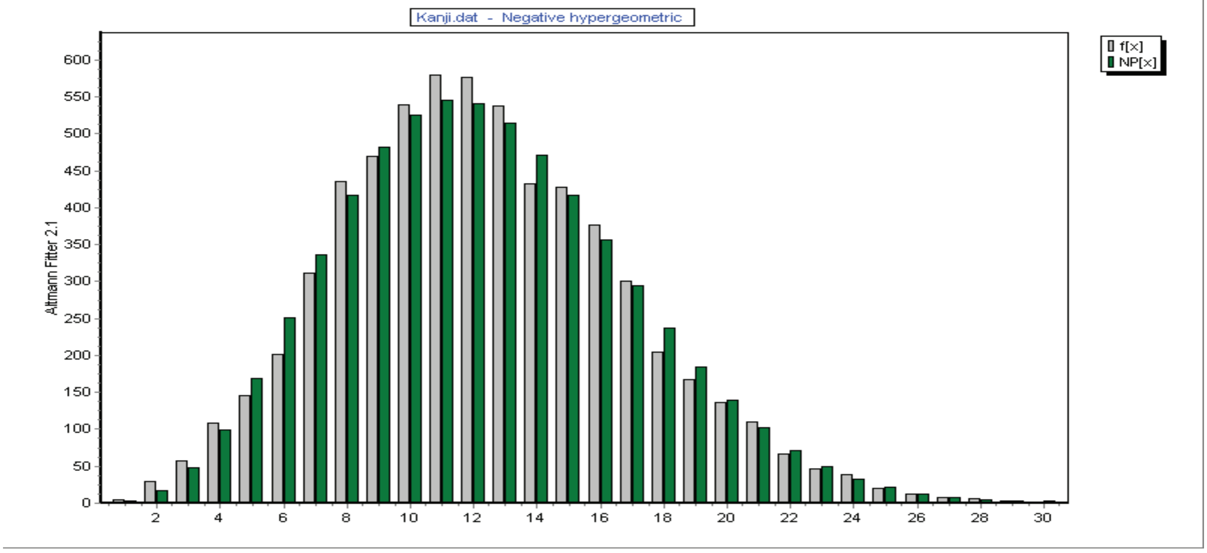


Figure 1. The negative hypergeometric model for the distribution of strokes of 6,355 kanji

3. Modelling the relationship between kanji strokes and kanji frequency

A further question is whether the above model continues to be suitable if the frequencies of the given kanji *in a corpus* are taken into consideration. This question is of a rather methodological nature. The first problem is the sample size. In our case, 6,335 kanji were taken from a very large corpus of 86,542,349 word occurrences where no classical statistical test can rescue the model from rejection. In other words, even if the model could perhaps be accepted, *cum grano salis*, the characteristics of the Chi-square test, destroy our hope. Its linear increase with increasing sample size is fatal for this investigation. The second problem is that language laws hold for homogeneous data. However, a corpus is as heterogeneous as it can be. It contains a mixture of frequencies which, in spite of statistical practice, do not level out with the increasing sample size. On the contrary, the irregularities can escalate. We might erroneously assume that a corpus is a sample from a population, but it has frequently been claimed that there are no populations in language (cf. Orlov, Boroda, Nadarejšvili, 1982). There is no population that can be called "Shakespeare" or "Akutagawa", nor any population as "the language of the *Asahi Newspaper*" or "the language of the *Times*". Likewise, there is no population that can be called "English texts", "Japanese texts", "the word stock of German" or even "the word stock of Ainu". Great dictionaries contain about one to two hundred thousand lexical items, but the German word stock with all of its special dictionaries (terminology) is estimated to be approximately twenty million. A student of German found in one volume of a German magazine "*Der Spiegel*" more than 8,000 compounds that were not found in any German dictionary. Do they belong to the word stock if they are constructed ad hoc and not in regular use? Is the size of the word stock a real or a potential number? Even though some models yield good results if the lexicon size is considered infinite (cf. Kornai, 2002), the usual models of vocabulary growth are of an exponential type, having finite limits (e.g., Piotrowski, Bektaev, Piotrowskaja, 1985).

As can be seen in Figure 2, the frequency of kanji with a given number of strokes is very rhapsodic. It is multimodal. The main irregularities are in the middle of the distribution. Deviations from any "honest" distribution are so enormous that in this case, no usual model is adequate.

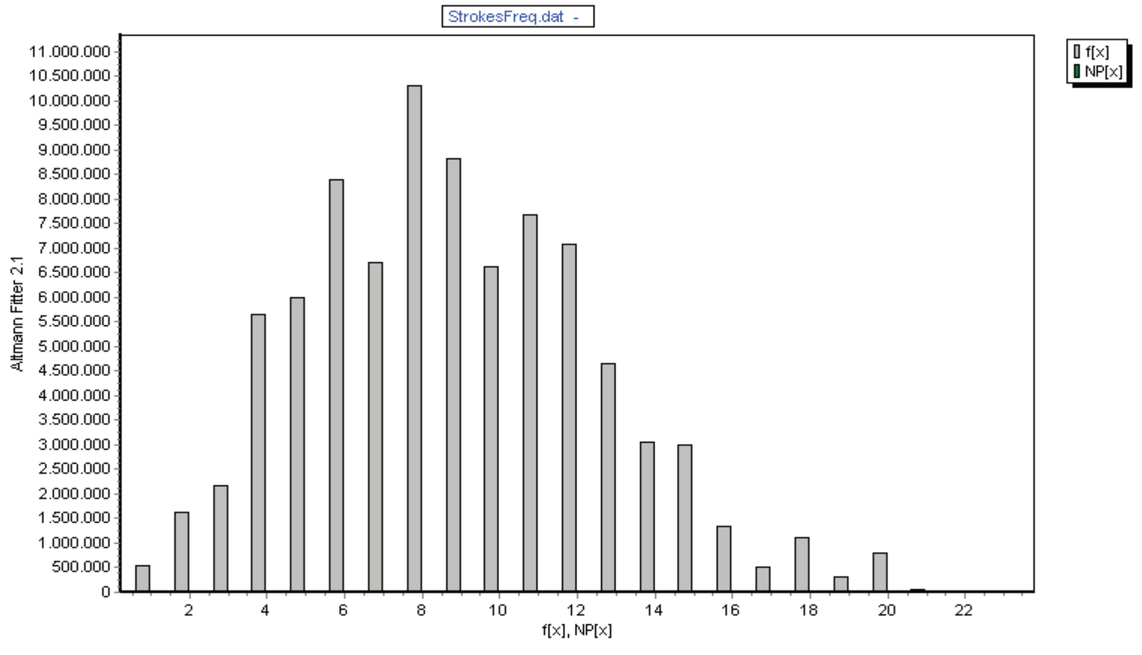


Figure 2. Frequency of kanji with X strokes

Of course, there are different possibilities to capture this irregularity. One of them is the mixing of distributions in different proportions. This technique, however, multiplies the number of parameters and it is not known how many components must be used. The second possibility is smoothing by pooling two or more frequency classes. In this case, the character of a discrete distribution is lost. This is, of course, merely a question of approximation, and in principle it is not relevant what kind of model – discrete or continuous – is used, as the data are distorted. Even if the law behind their generation was known, any model of this law would be forced to take account of many subsidiary conditions, which consequently yield a very special case. Therefore, it is best to consider modelling as a conceptual activity, taking into account simplifications, and even conscious distortions, at the beginning (Bunge, 1967, p. 388).

Table 2 presents the original data corresponding to Figure 2 and the data pooled by 2 using the mean X of the pertinent frequency classes (column 3). It can be seen that in the latter case the "curve" is smooth and can be well approximated by either a continuous distribution or a simple continuous curve.

The second case is preferable on several grounds. First, there is no need to derive a model for false, distorted data – since mixed data are false data. It would be more correct to take each article of the *Asahi Newspaper* separately. Second, the first approximation should be as concise as possible. Third, there is no problem in switching from discrete models to continuous ones or vice versa (cf. Mačutek, Altmann, in press). Since there is no software for a battery of continuous distributions, non-linear regression is simply used to find a preliminary model. Fitting is simpler if the absolute frequencies are transformed into relative ones, as shown in the fifth column of Table 2. Starting from the above mentioned Wimmer-Altmann theory (2001) the simplest case of its continuous version was chosen, namely

$$(2) \qquad \frac{dy}{y} = \left(\frac{b}{x} - c\right)dx$$

which says that the relative rate of change of frequency (*dy/y*) is proportional to the relative rate of change of stroke number (*bdx/x*) to which a disturbing constant *c* is added, originating from distortions caused by mixing of texts and pooling of frequency classes. The solution of the differential equation (2) yields

$$(3) \qquad y = ax^{b}e^{-cx}$$

and the results given in the last column of Table 2. As can be seen, the high value of $R = 0.97$ corroborates this kind of approach. Of course, this is merely the first, not the last word.

Based on values in Table 2, formula (3) is calculated as:

$$y = 0.0040x^{3.8342}e^{-0.5058x}$$

This formula is depicted graphically in Figure 3. It shows that after smoothing very heterogeneous data, the relation between stroke number and frequency becomes increasingly rigorous, and almost bell-shaped (with a slight asymmetry).

Table 2
Frequencies of kanji with *x* strokes

X (Stroke)	f_x (Kanji frequency)	x*	f_x*	p^*_x	p_{theor}
1	530246	1.5	2153263	.024881	0.008890
2	1623017				
3	2172880	3.5	7840187	.090594	0.083269
4	5667307				
5	5983444	5.5	14368733	.166031	0.171322
6	8385289				
7	6722157	7.5	17026988	.196747	0.204631
8	10304831				
9	8835897	9.5	15466409	.178715	0.184202
10	6630512				
11	7687320	11.5	14765449	.170615	0.139357
12	7078129				
13	4663567	13.5	7718553	.089188	0.093718
14	3054986				
15	3010830	15.5	4343000	.050183	0.057885
16	1332170				
17	520867	17.5	1647099	.019032	0.033523
18	1126232				
19	321022	19.5	1127029	.013023	0.018460
20	806007				
21	65617	21.5	77071	.000890	0.009761
22	11454				
23	8568	23.0	8568	.000099	0.005920
	86542349		86542349		a = 0.0040 b = 3.8342 c = 0.5058 R = 0.9712

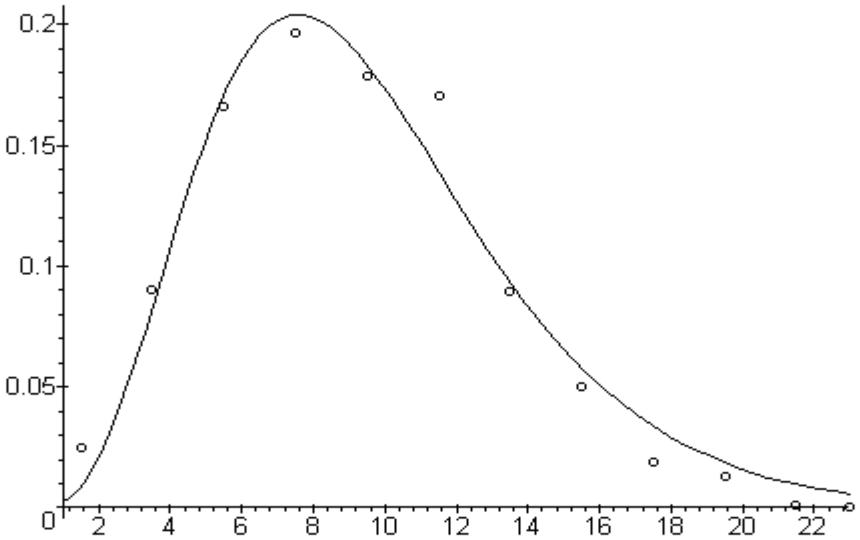


Figure 3. Fitting (3) to the relationship between number of strokes and kanji frequency

5. Modelling the relationship between kanji strokes and lexical compositional symmetry

In a previous study by Tamaoka and Altmann (2004), the symmetry between left-hand and right-hand side compounding has been examined both for individual kanji and for the whole field of 1,945 kanji. For instance, the kanji 学 /gaku/ meaning 'to learn' or 'learning' is combined with another kanji on the right-hand side position such as in 学校 (/gaQ koR/, 'school'), 学生 (/gaku sei/, 'student') and 学者 (/gaku sja/, 'scholar'). Combinations with other kanji on the left-hand side position are also possible such as in 入学 (/njuR gaku/, 'school admission'), 文学 (/bun gaku/, 'literature') and 私学 (/si gaku/, 'private school'). Tamaoka and Altmann investigated two questions regarding how two-kanji compound words were produced from a single kanji by adding another kanji on either the left or the right side, and how symmetric they are on both sides. Furthermore, the present study examines whether the number of strokes has some influence on the symmetry of compound building. Tamaoka and Altmann (2004) tested asymptotically the symmetry using the chi-square criterion

$$(4)\qquad X^2 = \frac{(n_L - n_R)^2}{n_L + n_R}$$

which is distributed as a chi-square with 1 degree of freedom. Formula (4) does not show directly whether there is left or right symmetry. Therefore for use as a coefficient of symmetry we express (4) in the form

$$(5)\qquad S = \frac{(n_L - n_R)}{\sqrt{n_L + n_R}}$$

indicating left asymmetry if $S > 0$, and right asymmetry if $S < 0$, without evaluating its significance. Of course, any other coefficient that shows the direction of asymmetry and does not allow 0 in the denominator would also be appropriate. Eliminating all cases of kanji having no compounding, we obtained 1,934 cases for which we computed the mean asymmetry at each $x = 1,\ldots,23$. As shown in Figure 4, the results indicated that the greater the number of strokes, the greater the left asymmetry of the kanji.

Table 3
Number of strokes of a kanji and its asymmetry

Number of strokes	Formula (3)	Lin.reg.	Lin.reg.
1	-9.52823359	-2.431076	--------
2	-0.93903791	-2.240516	-0.949802
3	-1.65643326	-2.049956	-0.851436
4	-0.81334384	-1.859395	-0.753069
5	-0.59039237	-1.668835	-0.654703
6	-0.37989045	-1.478275	-0.556337
7	0.02421970	-1.287714	-0.457970
8	0.20483855	-1.097154	-0.359604
9	-0.21886048	-0.906594	-0.261237
10	0.22139170	-0.716034	-0.162871

11	-0.18819254	-0.525473	-0.064504
12	0.05195306	-0.334913	0.033862
13	-0.00252829	-0.144353	0.132229
14	0.28074040	0.046208	0.230595
15	0.28992592	0.236768	0.328962
16	0.02195363	0.427328	0.427328
17	-0.00939287	0.617888	0.525695
18	0.50832419	0.808449	0.624061
19	0.85177389	0.999009	0.722427
20	1.13805872	1.189569	0.820794
21	0.59888953	1.380129	0.919160
22	0.65786894	1.570690	1.017527
23	1.78854380	1.761250	1.115893

		R = 0.3688
		a = -2.6216
		b = 0.1906
			R = 0.7611
			a = -1.1465
			b = 0.0984

The values (number of strokes vs. formula (3)) are given in Table 3 (first three columns). Again, x = 1 displays a fully anomal value and can be left out from further computation. There are only two kanji with one stroke, 一 and 乙. The kanji 乙 is rather uncommon, occurring only 418 times in the *Asahi Newspaper* corpus of 287,792,797 words (Amano & Kondo, 2000). This kanji only produced 11 two-kanji compound word-types with a joint token frequency of 418 (this happened to be the same as the single kanji frequency); by contrast, the kanji 一 produces 381 two-kanji compound word-types with a joint token frequency of 529,828 (Tamaoka & Makioka, 2004). This discrepancy of the two kanji with one stroke created a deviated value of -9.52823359 in the formula (3). Excluding the case of these two one-stroke kanji, we obtain a rather clearer picture in the fourth column of Table 3 and Figure 4. The F-test for linear regression is highly significant in both cases (with or without x = 1), but we prefer the determination coefficient even if it is not so high (R = 0.76).

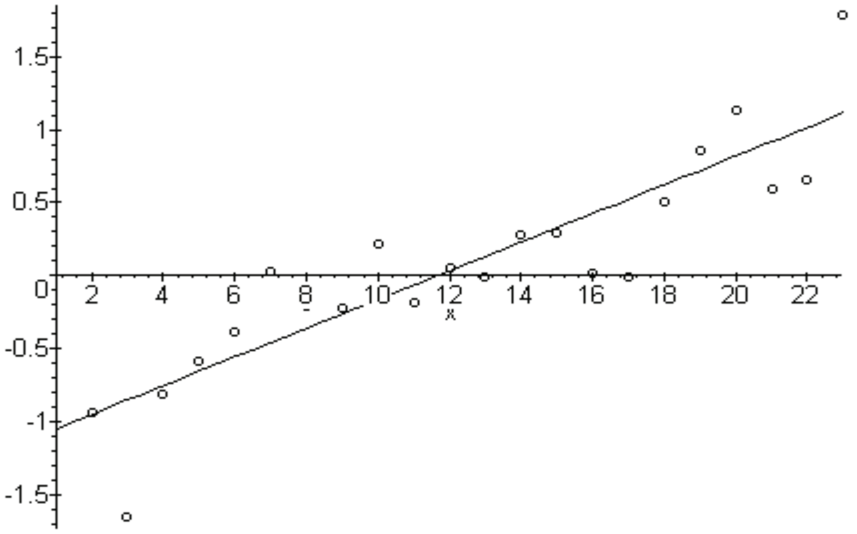


Figure 4. Kanji strokes vs. asymmetry (without x = 1)

This result shows that the greater the number of strokes in a given target kanji, the more it tends to produce two-kanji compound words by adding a kanji on the right side (with the target kanji positioned on the left side). This tendency was called *regressive compounding* in the kanji symmetry study of Tamaoka and Altmann (2004). In the case of a kanji with two strokes, the kanji 大 meaning 'big' appears 503,023 times in the *Asahi Newspaper* (Amano & Kondo, 2000). When this kanji is positioned on the right side, 399 two-kanji compound words are produced by adding other kanji on the left side, e.g. 巨大 meaning 'huge'. The same kanji produces 70 different two-kanji compound words by adding other kanji on the right side, e.g. 大学 meaning 'a university'. Roughly five times more compounds were produced when the target kanji is positioned on the right side (adding kanji on the left side) than on the left side (adding kanji on the right side). The token frequencies of these compounds show a similar difference; left-side compounds occur 374,376 times and right-side compounds 128,647 times.

In the case of a kanji with many strokes, this tendency reverses. For example, the kanji 警 meaning 'warn' or 'admonish' appears 91,377 times in the *Asahi Newspaper*. When this kanji is positioned on the right side, 10 different two-kanji compound words are produced by adding other kanji on the left side; for instance 夜警 meaning 'night watch'. The same kanji produces 27 different two-kanji compound words by adding other kanji on the right side, for example 襲来 meaning 'invasion'. The number of compounds for the left side position of the target kanji (adding kanji on the right side) was 2.7 times greater than for the right side position of the target kanji (adding kanji on the left side). The difference in token frequencies of these compounds is not so great but still maintains a similar trend: left-side compounds occur 66,454 times and right-side compounds occur 91,377 times.

Although this overall tendency of *regressive compounding* between kanji strokes and kanji compositional or compounding asymmetry is clearly shown in Figure 4, it is a surprising discovery which lacks a clear explanation.

5. Mathematical modelling for kanji strokes with multiple readings

Kanji pronunciations can be divided into two types: the On-reading derived from the original Chinese pronunciation, and the Kun-reading originating from the Japanese way of reading kanji (for details see Kaiho & Nomura, 1983; Morton & Sasanuma, 1984; Tamaoka, 1991, 2003, 2005a, 2005b; Wydell, Butterworth & Patterson, 1995; Wydell, Patterson & Humphreys, 1993). The mixture of material with two phonological origins created multiple pronunciations of a single Japanese kanji. In dispensing with the Chinese tones, and adapting three different sound systems from China, the Japanese created a great number of multiple readings for the kanji. According to the *Database for the 1,945 Basic Japanese Kanji (2nd edition)* produced by Tamaoka, Kirsner, Yanase, Miyaoka and Kawakami (2002), the total number of kanji which have only one pronunciation is 699 (667 kanji with a single On-reading and 32 kanji with a single Kun-reading). This is 35.94 percent of the 1,945 basic Japanese kanji. Kanji which have only On- or only Kun-readings, regardless of the number of pronunciations, total 779 (739 for On-reading and 40 for Kun-reading), or 40.05 percent of the 1,945 basic Japanese kanji. In other words, although it is a commonly-held notion that a kanji has both an On-reading and a Kun-reading, only 1,166 kanji or about 59.95 percent of the 1,945 basic Japanese kanji have both types of pronunciations.

Starting from the well-known dependence between word length and word polysemy, we conclude that the same relation must hold between the number of strokes (length) of a kanji and the number of readings, which may be an analogue to polysemy. In Table 4 we see the relevant numbers based on 1,945 kanji (Tamaoka, et al., 2002). The usual relation is here (cf. Köhler 1986); it can be derived from the differential equation (2) setting c = 0 :

$$(6) \qquad y = ax^{-b} .$$

In order to obtain a more solid result, we leave out all mean values based on less than 10 kanji in each case of strokes, i.e. in Table 4, two kanji with 1 stroke, 5 kanji with 21 strokes, 2 kanji with 22 strokes and 1 kanji with 23 strokes. A total of 10 kanji were excluded from the calculation of Formula (6). Iterative computing of (6) yields

$$y = 3.6382x^{-0.2454}$$

giving a preliminarily satisfactory result. Again, we would like to point out that the choice of 1,945 kanji is not random but is based on specific criteria. The graph of observed and computed values is in Figure 6.

Table 4
Dependence of the number of readings on the number of strokes

Number of strokes	Number of kanji	Number of readings (On + Kun)	Mean number of readings	Computed means (4)
1	2	5	2.50	-
2	12	36	3.00	3.07
3	30	95	3.17	2.78
4	66	153	2.32	2.59
5	93	220	2.37	2.45
6	111	258	2.32	2.34
7	141	305	2.16	2.26
8	187	398	2.13	2.18
9	179	379	2.12	2.12
10	199	410	2.06	2.07
11	195	396	2.03	2.02
12	199	431	2.17	1.98
13	147	296	2.01	1.94
14	105	199	1.90	1.90
15	104	180	1.73	1.87
16	68	142	2.09	1.84
17	34	57	1.68	1.82
18	33	64	1.94	1.79
19	21	30	1.43	1.77
20	11	21	1.91	1.74
21	5	9	1.80	-
22	2	5	2.50	-
23	1	1	1.00	-
			a = 3.6382, b = 0.2454, R = 0.81	

In order to show that this is parallel to Köhler's approach, we draw the respective parts of his scheme using our variables *L-Strokes* meaning the logarithm of stroke number and *L-readings* meaning the logarithm of reading numbers. Since the relationship has a decreasing character, we use a negative proportionality coefficient *–b* and add a factor *A* representing the specification requirement. The scheme is shown in Fig. 5.

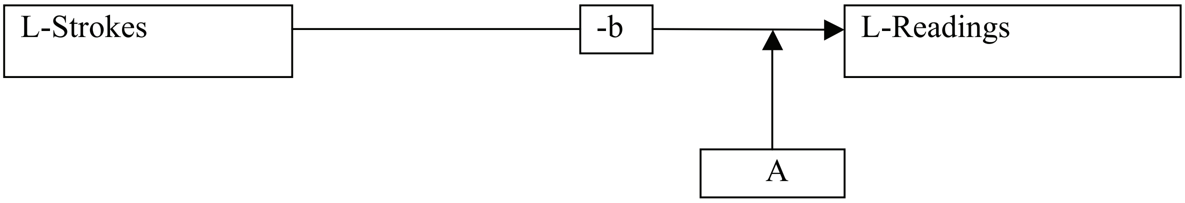


Fig. 5. Systems theoretical scheme of the relationship between the number of strokes and the number of readings of a kanji

The above scheme is analogous to that of the relationship between word length and meaning: in order to specify the meaning, one adds to the word an affix or another word (to build a compound) or reduplicates it. In case of kanji, the reading will be more specific the more complex the kanji is, i.e. the more elementary strokes it contains.
Using the above scheme we have

$$L\text{-}Readings = A - b(L\text{-}Strokes)$$

or $$\log y = A - b \log x.$$

Taking antilogarithms and denoting $e^A = a$, we obtain formula (6).

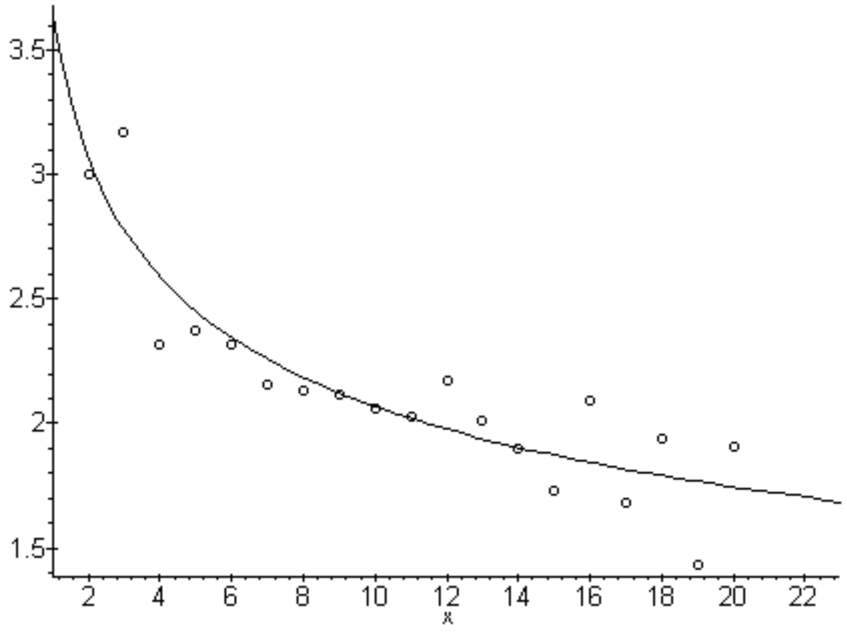


Figure 6. Kanji strokes with mean numbers of kanji readings

Thus, taking pairs of kanji properties and examining their links, we can discover a partial coincidence with Kohler's cycle; but in general we must expect also other types of relationships which are characteristic of kanji only. A more general view of this issue might be achieved by incorporating analyses of Korean and even Assyrian script, both having analogous properties. In Köhler's original self-regulating cycle, only four properties are set in relation: frequency, length, polysemy and polytexty; and all are linked in the way given by formula (6). With the aid of Köhler's cycle, formulas like (3) and (6) can be set up without the use of differential equations; at the same time they yield a systems theoretical explanation. The formulae express all the requirements imposed on language by its users (cf. Köhler 1986, 1987, 1990a,b,c, 2005). Usually they are captured by the parameters, which must be interpreted; later on their value must be stated *a priori*.

6. Conclusions

It is known that word length is strongly related to frequency in many languages (e.g., Köhler 1986; Breiter 1994; Hammerl 1991; Leopold 2000; Miyajima 1992; Strauß, Grzybek, Altmann 2005). In the case of Japanese kanji, word length can be described as *visual complexity* represented by kanji strokes which require drawing a whole picture of a kanji. The number of kanji strokes is one of a kanji's central properties. This study has investigated the relationships of the number of strokes in Japanese kanji to their printed frequencies of occurrence, compositional asymmetry and kanji multiple readings. Distributions of kanji strokes in *both* samples (the 1,945 basic kanji, and the 6,355 kanji appearing in the *Asashi Newspaper* between 1985 and 1998) produced a negative hypergeometric distribution. The distribution of strokes of the 1,945 kanji with their printed frequencies is rather rhapsodic. Looking at the relationship between kanji strokes and the kanji's lexical compositional asymmetry uncovers the interesting tendency of *regressive compounding*, which fits in with a previous proposal by Tamaoka and Altmann (2004). It has been ascertained that the greater the number of strokes in a kanji, the more strongly it tends to produce two-kanji compound words by adding a kanji on the right side (i.e. with the target kanji positioned on the left side). Furthermore, we have also examined the number of readings of a kanji in relation to the number of kanji strokes. A kanji often has multiple On- and Kun-readings; but we found that the larger the number of kanji strokes, the fewer the number of readings. In other words, the more visually complex the kanji is, the more specialised its reading becomes. As such, the number of strokes, as one of the central characteristics of the kanji, has close relationships with other properties of the kanji: frequency, asymmetry and reading. The present study has demonstrated these relationships. Since the number of kanji strokes can be considered an integral feature of the system of morphemes and lexical properties, we suggest that further investigations into this area will strongly enhance our overall picture of the mutual relationships among different properties of the kanji.

References

Amano, N., & Kondo, K. (2000). *Nihongo-no goi tokusei [Lexical properties of Japanese].* Tokyo: Sanseido.

Breiter, M. A. (1994). Length of Chinese words in relation to their other systemic features. *Journal of Quantitative Linguistics 1, 224-231.*

Bunge, M. (1967). *Scientific research I.* Berlin: Springer.

Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells.* Trier, WVT.

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

Kobayashi, I. (1981). *Kanji kyooiku-no kiso kenkyuu [Fundametal studies on teaching kanji].* Tokyo: Meiji Tosho.

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Köhler, R. (1987). Systems theoretical linguistics. *Theoretical Linguistics 14, 241-257.*

Köhler, R. (1990a). Linguistische Analyseebenen, Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation. *Glottometrika 11, 1-18.*

Köhler, R. (1990b). Zur Charakteristik dynamischer Modelle. *Glottometrika 11, 39-46.*

Köhler, R. (1990c). Elemente der synergetischen Linguistik. *Glottometrika 12, 179-187***.**

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Handbook of Quantitative Linguistics: 760-774.* Berlin: de Gruyter.

Kornai, A. (2002). How many words are there? *Glotometrics 4. 61-86.*

Leopold, E. (2000a). Length-distribution of words with coinciding frequency. In: *Proceedings of the fourth conference of the International Quantitative Linguistic Association, Prague, August 24-26: 76-77.*

Mačutek, J., Altmann, G. (in press). Discrete and continuous modeling in quantitative linguistics. (to appear in Journal of Quantitative Linguistics 2006).

Miyajima, T. (1992). Relationship in the length, age and frequency of Classical Japanese words. *Glottometrika 13, 219-229.*

Morton, J., & Sasanuma, S. (1984). Lexical access in Japanese. In L. Henderson (Ed.), *Orthographies and reading: 25-42*. London: Lawrence Erlbaum Associates.

Orlov, J.K., Boroda. M.G., Nadarejŝvili, I.Š. (1982). *Sprache, Text, Kunst. Quantitative Analysen.* Bochum: Brockmeyer.

Piotrowski, R.G., Bektaev, K.B., Piotrowskaja, A.A. (1985). *Mathematische Linguistik.* Bochum: Brockmeyer.

Strauss, U., Grzybek, P., Altmann, G. (2005). Word length and word frequency. In: Grzybek, P. (ed.), *Word length studies and related issues: 255-272.* Boston/Dordrecht: Kluwer.

Tamaoka, K. (1991). Psycholinguistic nature of the Japanese orthography. *Studies in Language and Literature* (Matsuyama University), *11(1), 49-82***.**

Tamaoka, K. (2003). Where do statistically-derived indicators and human strategies meet when identifying On- and Kun-readings of Japanese kanji? *Cognitive Studies*, 10(4), 1-28.

Tamaoka, K. (2005a). Meimei kadai ni oite kanji 1-ji no shoji to on'in no tan'i wa itti suru ka [Is an orthographic unit of a single Japanese kanji equivalent to a kanji phonological unit in the naming task?] *Cognitive Studies*, *12(2), 47-73.*

Tamaoka, K. (2005b). The effect of morphemic homophony on the processing of Japanese two-kanji compound words. *Reading and Writing*, *18, 281-302.*

Tamaoka, K., & Yamada, H. (2000). The effects of stroke order and radicals on the knowledge of Japanese kanji orthography, phonology and semantics. *Psychologia*, *43, 199-210.*

Tamaoka, K., & Altmann, G. (2004). Symmetry of Japanese kanji lexical productivity in the left- and right-hand sides. *Glottometrics*, *7, 65-84.*

Tamaoka, K., & Makioka, S. (2004). New figures for a Web-accessible database of the 1,945 basic Japanese kanji, fourth edition. *Behavior Research Methods, Instruments & Computers*, *36(3), 548-558.*

Tamaoka, K., Kirsner, K., Yanase, Y., Miyaoka, Y., Kawakami, M. (2002). A Web-accessible database of characteristics of the 1,945 basic Japanese kanji. *Behavior Research Methods, Instruments & Computers, 34(2), 260-275.*

Wimmer, G. & Altmann, G. (2001). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Handbook of Quantitative Linguistics: 791-807.* Berlin: de Gruyter.

Wydell, T.N., Patterson, K.E., & Humphreys, G.W. (1993). Phonologically mediated access to meaning for kanji: Is a rows still a rose in Japanese Kanji? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, *491-514.*

Wydell, T. N., Butterworth, B., & Patterson, K. E. (1995). The inconsistency of consistency effects in reading: The case of Japanese kanji. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, *1155-1168.*

Glottometrics 10, 2005, 30-44

A Database of Two-Kanji Compound Words Featuring Morphological Family, Morphological Structure, and Semantic Category Data

Hisashi Masuda, Hiroshima Shudo University
Terry Joyce, Tokyo Institute of Technology[1]

Abstract: One of the most fundamental issues for all models of the mental lexicon is how to represent essential information about the morphological structure of polymorphemic words. This paper describes the construction of a large-scale database of two-kanji compound words, which supplements a central component of data relating to 78,426 compound headwords from the Kōjien dictionary with several components focusing on morphological family, morphological structure, and semantic category data. The database will be a particularly valuable resource in terms of supporting and extending research into the lexical retrieval and representation of two-kanji compound words within the Japanese mental lexicon from the perspective of compound word morphology, such as the series of constituent-morpheme priming experiments (Joyce, 1999, 2002, 2003a, 2003b, 2004; Joyce & Masuda, 2004) that are discussed briefly.

Keywords: Two-kanji compound words, morphological family, morphological structure

1 Introduction

As an important part of our linguistic knowledge, the representation of morphological information concerning the structure of polymorphemic words is a fundamental issue for all models of the mental lexicon (e.g., Feldman, 1995; Jarema, Kehayia, & Libben, 1999; Sandra & Taft, 1994; Taft, 1991). This is clearly true not only because of the vast numbers of polymorphemic words that exist in all languages and because of the relative ease with which language users produce and comprehend both existing and novel forms (Sandra, 1994), but also because the issue has profound implications for lexical processing and for the organization of lexical representations within the mental lexicon.

Indeed, the involvement of morphological information in the mental lexicon has been one of the most researched and debated topics within visual word recognition research over the last 30 years or so. The debate has focused mainly on comparing competing models of lexical representation and their assumptions concerning lexical processing. For example, in contrast to full-listing models (e.g., Butterworth, 1983), that assign no role to morphology, decomposed storage models, such as the extremely influential 'prefix-stripping' model of Taft and Forster

[1] Address correspondence to: Hisashi Masuda, Ph.D., Hiroshima Shudo University, 1-1-1 Ozuka-Higashi, Asaminami-ku, Hiroshima, 731-3195 Japan. Email: hmasuda@shudo-u.ac.jp or to Terry Joyce, Ph.D., Tokyo Institute of Technology, W9-29, 2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan. Email: terry@valdes.titech.ac.jp

Part of this work has been made possible by a grant to the first author from the Institute of Advanced Studies at Hiroshima Shudo University (Japan). The second author's contributions to the database are part of a research project under the 21st Century COE Program "Framework for Systematization and Application of Large-scale Knowledge Resources" (Program leader; Professor Sadaoki Furui) at the Tokyo Institute of Technology, Japan.

(1975, 1976), regard morphological parsing as an obligatory stage in lexical access. Occupying the middle ground, there are also models that propose the existence of both whole-word and morpheme representations, but which adopt different approaches to lexical access, such as the augmented addressed morphology model (Caramazza, Laudanna, & Romani, 1988) and the parallel dual route model of morphological processing (Schreuder & Baayen, 1995), which both assume separate parsing routes, or the multilevel interactive- activation framework (Taft, 1991; 1994), which treats morpheme representations as intermediate-level units.

While most of this research has been concerned primarily with the inflectional and derivational morphology of relatively few languages, such as English, Italian, and Dutch, that all use alphabetic writing systems, research into the nature of morphological involvement within the Japanese mental lexicon can undoubtedly make very valuable contributions to this body of research for two simple but extremely important reasons. The first reason relates the complex nature of the Japanese writing system which, in addition to two syllabographic, or more precisely moraic, kana scripts, continues to extensively use kanji, which are most appropriately characterized as a morphographic writing system. The second reason is that, because of extensive lexical borrowing from Chinese and native word-formation processes, compounding is highly productive in Japanese (Kageyama, 1982), with the two-kanji compound word being the most common word structure in the Japanese language (Nomura, 1988; Yokosawa & Umeda, 1988). Apart from a few notable exceptions (e.g., Hirose, 1992; Joyce, 1999, 2002, 2004; Joyce & Masuda, 2004; Tamaoka & Hatsuzuka, 1998), however, there has, rather surprisingly, been relatively little research that has focused specifically on the lexical retrieval and representation of two-kanji compound words within the Japanese mental lexicon from the perspective of compound word morphology. While the relative lack of research into the morphological aspects of two-kanji compound words may simply be because researchers have been preoccupied with orthographic (e.g., Kawakami, 1997, 2000; Ogawa & Saito, 2001) and phonological (e.g., Fushimi, Ijuin, Patterson, & Tatsumi, 1999; Masuda, 2002a; Wydell, Patterson, & Humphreys, 1993) aspects, we believe that it also reflects the fact that there have been very few databases dedicated to the lexical properties of two-kanji compound words and, in particular, large-scale databases featuring morphological family and structure data.

This paper reports on the construction of a database of two-kanji compound words featuring morphological family, morphological structure, and semantic category data,[2] which the present authors are building in order to conduct, and hopefully encourage, further research into the morphological aspects of two-kanji compound words in the Japanese mental lexicon. A central component of the database is a list of 78,426 two-kanji compound-word headwords, of which both constituents belong to the 2,965 Japanese Industrial Standard level 1 (JIS1) kanji list, that was extracted from Kōjien, an authoritative desktop dictionary of the Japanese language (Shinmura, 1995). The database also consists of a number of other components that emphasize various morphological and semantic aspects of two-kanji compound words. After briefly discussing the theoretical implications of extending the concepts of *orthographic neighbors* (Coltheart, Davelaar, Jonasson, & Besner, 1977) and *morphological families* (Schreuder & Baayen, 1997) to the Japanese writing system, Part 2 of the paper introduces the morphological family data components of the database, which combine counts for the Kōjien list with usage-based cumulative frequency data (Joyce & Ohta, 2002). Part 3 starts with a selective review of some studies employing the constituent-morpheme priming paradigm before outlining the morphological structure components of the database. In addition to noting word-

[2] The database of two-kanji compound words featuring morphological family, morphological structure, and semantic category data (version 1.0) may be accessed at the following websites: http://ns1.shudo-u.ac.jp/-~hmasuda/cwdb.htm or http://www.valdes.titech.ac.jp/~terry/cwd.html. As detailed in this paper, and at the websites, the database presently consists of a number of Excel files which may be downloaded for research purposes, on the condition that use of the database is acknowledged by citing this article.

formation classification data collected by the second author, Part 3 introduces the first stage of an ongoing large-scale psychological survey concerning native Japanese speaker awareness for the morphological structure of the two-kanji compound words. Finally, Part 4 briefly describes the inclusion of semantic category data for 24,519 two-kanji compound words based on the National Institute for Japanese Language's (2004) recently revised word list according to semantic principle.

2 Morphological Family Data

There is considerable evidence suggesting that recognition of a target word is influenced by orthographically similar words, usually referred to as orthographic neighbors (Coltheart et al., 1977), and by morphologically-related or family words (Schreuder & Baayen, 1997). Much of the research into neighborhood effects has adopted Coltheart et al.'s (1977) straightforward definition of an orthographic neighbor—any word that can be generated by changing just one letter of a given word while preserving letter positions (e.g., *mice* and *race* are both neighbors of *rice*), with the neighborhood being the set of such neighbors. However, while this definition is simple enough, there has been much controversy surrounding neighborhood effects, especially over whether these are inhibitory or facilitatory in nature (e.g., Andrews, 1992; Grainger, 1990). In contrast to the purely visual overlap of orthographic neighbors, Schreuder and Baayen's (1997) notion of morphological family recognizes the semantic connections between sets of words sharing a constituent morpheme. Accordingly, a morphological family includes singular and plural forms (e.g., *table*, *tables*), as well as words sharing a stem formed either by derivation (*tablet*, *tabular*) or compounding (*tablespoon*, *timetable*). Looking at word frequency effects for monomorphemic, or simplex, Dutch nouns, Schreuder and Baayen reported an effect of morphological family size, but not for cumulative family frequency; a finding that has also been observed for English simplex nouns (Baayen, Lieber, & Schreuder, 1997).

Although the notion of orthographic neighbors has been extended to two-kanji compound words (e.g., Kawakami, 1997, 2000; Saito, 1997), in an analogy of equating 'one letter' with 'one character', as Joyce and Ohta (2002) point out, the analogy completely overlooks the fact that orthographically letters and characters function at different levels. In contrast to cenemic, or phonographic, writing systems where the graphic units represent either phonemes (i.e., alphabetic letters), or syllable-/mora-sized phonological units (i.e., Japanese kana), the graphic units of pleremic writing systems are semantically-informed denoting both sounds and meanings, which is the case with kanji (Coulmas 1996; Haas, 1976, 1983). While the term logographic is often used for kanji, this is undoubtedly misleading for it implies that only lexemes are represented and, as Joyce and Ohta suggest, a far more accurate term is morphographic, reflecting the fact that kanji represent both free and bound morphemes. In this light, we believe that morphological family is the more appropriate concept for thinking about the relationships between a set of two-kanji compound words that have a constituent kanji in common.[3]

Setting aside such theoretical issues for the moment, we turn now to introduce our morphological family data. There are a couple of important differences between Kawakami's

[3] While claiming that morphological family is the appropriate concept for the Japanese writing system, we acknowledge that our database only covers the two-kanji compound words members of a family. Complete family data would also include the frequencies of a morpheme as a word stem (i.e., in verbs, such as 化 as the stem of 化ける /bakeru/ 'turn, change') and as constituents of longer compound words (i.e., 化 as a suffix of the meaning '-ize' in 近代化 /kindaika/ 'modernize').

(1997, 2000) data, a similar database by Ogawa, Saito, and Yanase (2005),[4] and the morphological family data components of our database that require some comment. The first major difference is that while Kawakami and Ogawa et al. only present data based on the Kōjien dictionary (editions 4 and 5, respectively), our database also includes usage-based type and token counts (Joyce & Ohta, 2002). For example, the corresponding counts in Ogawa et al.'s database, referred to as companions, are based solely on the Kōjien list of 78,426 two-kanji compound words, but the problem with only having dictionary-based counts is that the counts can be inflated by rarely used words. While highlighting the difficult issues faced by researchers seeking to quantify the mental lexicon, the inclusion of low-frequency words entails, at least implicitly, the untenable assumption that they are actually stored in the average mental lexicon. Accordingly, our database provides both Kōjien-based counts and usage- based counts (Joyce & Ohta, 2002) to assist interested researchers in making the appropriate comparisons.

The second significant difference relates to the sources and use of frequency data. While Kawakami (2000) provides cumulative frequency data for constituent kanji (but not frequency data for the compound words themselves) based on the floppy disk version (1997) of the National Language Research Institute's (NLRI) (1962) magazine survey, Ogawa et al. (2005) provide compound word frequency data (but not cumulative frequency data for constituents) based on the NLRI's (1970) newspaper survey. However, the major concern with both of these as appropriate measures of present-day word frequencies stems from the fact that the relevant surveys were conducted more than 35 years ago and at least a decade prior to the promulgation in 1981 of the Jōyō Kanji List, the official guideline specifying 1,945 kanji for daily use. In contrast, our database has both compound word frequency and cumulative constituent kanji frequency data that Joyce and Ohta (2002) compiled from a six-year period (1993-1998) of newspaper frequency data included in the NTT database (Amano & Kondō, 2000).

Table 1
Morphological Family Data for the First Five JIS1 Kanji as a Function of Position

		First constituent				Second constituent			
Code	Kanji	K	U-TTy	U-ATy	U-ATo	K	U-TTy	U-ATy	U-ATo
16-01	亜	16	10	5.5	63.5	5	3	3.0	55.2
16-02	唖	6	1	1.0	9.2	3	2	1.0	2.0
16-03	娃	0	0	0	0	0	0	0	0
16-04	阿	39	7	3.8	12.3	10	0	0	0
16-05	哀	29	18	13	138	1	1	1.1	46.3

Note: K = morphological family count based on the Kōjien list; U-TTy = the total type count based on usage (Joyce & Ohta, 2002); U-ATy = the average type count based on usage; U-ATo = average token count based on usage. This table is based on the presentation of the data in the 'Morphological family data-Constituents' Excel file.

Table 1 shows morphological family counts for the first five JIS1 kanji as a function of their position within compound words. Sorted according to the kuten code for the JIS1 kanji, the morphological family data consists of four kinds of data for the kanji as a first constituent of two-kanji compounds words and the corresponding counts as the second constituent. The

[4] Ogawa et al's (2005) database would seem to be more focused on the constituent kanji and, particularly, their pronunciations, rather than on the two-kanji compound words themselves. Their notion of phonological neighbors, based on the pronunciations of the constituent kanji, is certainly much more restrictive than what the traditional definition would encompass.

first family count (K) is the type count based for the Kōjien list after adjustment for orthographic repetitions.[5] The remaining three counts are usage-based cumulative frequency counts calculated by Joyce and Ohta (2002).[6] The first (U-TTy) is the total type count for the six-year period, while the second (U-ATy) is the average type count over the period. The last count (U-ATo) is the average token count.

Table 2
Morphological Family Data for 亜 as First Constituent

First constituent	Compound	Pronunciation	Usage	U-ATy	U-ATo
亜	亜鉛	あえん	1	1.00	34.17
亜	亜欧	あおう	1	0.17	0.17
亜	亜科	あか	1	0.17	0.17
亜	亜綱	あこう	1	0.17	0.17
亜	亜将	あしょう	0	0	0
亜	亜流	ありゅう	1	1.00	9.17
亜	亜種	あしゅ	1	0.83	7.83
亜	亜聖	あせい	0	0	0
亜	亜相	あしょう	0	0	0
亜	亜族	あぞく	0	0	0
亜	亜炭	あたん	0	0	0
亜	亜父	あふ	0	0	0
亜	亜麻	あま	1	0.33	0.33
亜	亜目	あもく	1	0.50	0.50
亜	亜門	あもん	1	1.00	10.5
亜	亜鈴	あれい	1	0.33	0.50
亜	16		10	5.5	63.5

Note: Usage indicates whether the compound word is included in the usage counts (Joyce & Ohta, 2002); U-ATy = average type count (max. 1.00); U-ATo = average token count. This table is based on the presentation of the data in the ‘Morphological family data-Compound words-First’ Excel file.

Table 2 presents part of the morphological family data for 亜 /a/ ‘come after; sub-; Asia’, showing the 16 two-kanji compound word members of which 亜 is the first constituent, together with their pronunciations. As it is difficult to present both sides of a constituent kanji’s complete morphological family in a single Excel file, the full family listings are split between two files (‘Morphological Family Data-First constituent’ and Morphological Family Data-Second constituent’). The usage column indicates whether the compound word is included in

[5] Because Kōjien treats cases where an identical orthographic form is associated with more than one pronunciation or more than one meaning as separate headwords, the total of 78,426 must be adjusted when counting words based on orthographic form. Adjusting for orthographic repetitions (7,614 types and 17,048 tokens), the total number of orthographic types is actually 68,992.

[6] It should be noted that Joyce and Ohta (2002) excluded proper nouns from their data. Although the treatment of proper nouns is problematic, especially for kanji, they were omitted because proper nouns are not normally used in word recognition research and because of their special distributional characteristics (while proper nouns represented 49% of the type counts in the newspaper corpus, they only accounted for about 13% of the tokens).

Joyce and Ohta’s counts, while U-ATy and U-ATo are the average type and the average token counts over the six-year period, respectively. The average type count indicates how frequently the particular compound word appeared over the 6-year period of the newspaper corpus; so, for example, 1.00 means every year, while 0.5 indicates that the compound word appeared in three out of the six years. Note that the total line in Table 2 corresponds to the first constituent counts for 亜 in Table 1.

3 Morphological Structure Data

One of the most useful experimental paradigms for investigating the extent of morphological involvement in the lexical retrieval and representation of polymorphemic words, particularly compound words, within the mental lexicon is what has been referred to as constituent-morpheme priming—comparing the facilitation on lexical decision responses to a compound word due to prior presentation of a constituent morpheme relative to a control condition (Joyce, 2002; see also Drews, 1996).[7] The constituent-morpheme priming paradigm has been used to investigate compound words in a number of European languages. For instance, Monsell (1985) has employed the paradigm in a study of both semantically- transparent (e.g., *tightrope*) and opaque (e.g., *butterfly*) English compound words, finding facilitation in both constituent prime conditions for both types of compounds. Sandra (1990) has also used a variation of the paradigm, presenting primes that are associatively related to a constituent, in a study of Dutch compound words. However, while he also observed facilitation for both constituent conditions for semantically-transparent compounds, there was no priming for opaque compounds. More recently, Kehayia, Jarema, Tsapkini, Perlak, Ralli, and Kadzielawa (1999) have conducted constituent-morpheme priming experiments with transparent noun-noun and adjective-noun compound words in Greek and Polish, reporting priming for both constituent conditions in both languages.

The constituent-morpheme priming paradigm has also been employed in a series of studies that specifically address the nature of lexical retrieval and representation of two-kanji compound words within the Japanese mental lexicon from a morphological perspective (Joyce, 1999, 2002, 2003a, 2003b; Joyce & Masuda, 2004). Given the rich diversity in the morphological structure of two-kanji compound words, which must be captured by models of the Japanese mental lexicon, Joyce (1999, 2002) investigated the patterns of constituent-morpheme priming across five word-formation principles.[8] The principle conditions were modifier + modified (M+M) (e.g., 山桜 /yamazakura/ ‘mountain cherry’), verb + complement (V+C) (e.g., 登山 /tozan/ ‘mountain climbing’), complement + verb (C+V) (e.g., 外食 /gaishoku/ ‘eat out’), associative pairs (AP) (e.g., 男女 /danjo/ ‘man and woman’), and synonymous pairs (SP) (e.g., 山岳 /sangaku/ ‘mountains’). Across two experiments varying the stimulus onset asynchronicity (SOA) between the primes and target compound words, the results were very consistent, with both constituent conditions facilitating lexical decision responses across all

[7] The constituent-morphemic priming paradigm may be seen as a version of what is sometimes referred to as (partial) repetition priming, particularly in studies of derivational morphology. For instance, Fowler, Napps, and Feldman (1985) use the term repetition priming in their study that showed that affixed words (e.g., unhappy) facilitate responses to the word stem alone (e.g., happy) at similar levels to the repetition condition.

[8] While a number of classifications of the word-formation principles, or morphological structure, exist (e.g., Kageyama, 1982; Nomura, 1988), most recognize about nine main types. The other principles of affixation, repetition, abbreviation, and phonetic borrowing are, however, for varying reasons less suitable for the constituent-morpheme priming paradigm.

five word-formation conditions and, in the majority of cases, at similar levels, clearly suggesting that morphological information plays an important role in the lexical retrieval of two-kanji compound words.

The results also indicated a possible effect of verbal morphology, because the only word-formation condition with a significant difference between the first and second constituent conditions was in the V+C condition, where responses in the verbal constituent condition were faster. To further investigate that possibility, Joyce (2003a; 2003b) calculated positional ratios (PR) (i.e., how often a given kanji appears as the first constituent or as the second), based on the cumulative frequency data (Joyce & Ohta, 2002) discussed in Part 2, in order to contrast low and high PRs for the verbal constituents of V+C and C+V compound words. The main finding from those experiments was a reversed pattern of priming across the high-PR V+C and C+V compound word conditions; with greater priming for the verbal constituents than for the respective complement conditions. Additional evidence for the notion of verb morphology effects has also come from a recent experiment conducted by Joyce and Masuda (2004), with three short SOA conditions (60 ms, 150 ms, and 250 ms) to examine the time courses of morphological and semantic activation for two-kanji compound words, where again a reversed pattern of priming was observed between the V+C and C+V compound words across the two shortest SOA conditions.

This series of constituent-morpheme priming experiments, providing important evidence concerning the involvement of morphological information within the Japanese mental lexicon, has relied on the results of word-formation classification surveys conducted by the second author to establish the experimental contrasts between the word-formation conditions. While there is generally clear consensus about the various word-formation principles, the task of classifying a given two-kanji compound word under the appropriate principle can be more problematic. Accordingly, the surveys collected native Japanese speaker evaluations (on a 7-point scale) concerning the appropriateness of classifying a given two-kanji compound word according to a particular word-formation principle for a corpus of 1,561 two-kanji compound words.[9] The obtained classification evaluations are included in the present database as part of the morphological structure data component. Table 3 shows 10 example compound words, two from each of the five word-formation principles, with high classification evaluations.

Table 3
Examples of Two-Kanji Compound Words
with High Word-Formation Classification Evaluations

Compound word	Word-formation principle	Classification Evaluation
暖冬 /dantō/ ‘mild winter’	Modifier + modified	7.0
旧友 /kyūyū/ ‘old friend’	Modifier + modified	7.0
飲酒 /inshu/ ‘drink alcohol’	Verb + complement	7.0
乗馬 /jōba/ ‘horse riding’	Verb + complement	7.0
急増 /kyūzō/ ‘rapid increase’	Complement + verb	7.0
早退 /sōtai/ ‘leave early’	Complement + verb	7.0
男女 /danjo/ ‘man and woman’	Associative pairs	6.9

[9] Joyce and Ohta (1999) report on the first survey which included 200 compound words for each of five word-formation principles (1,000 compound words in total), and the classification evaluations for the additional 561 items (97 M+M, 205 V+C, 176 C+V, and 83 SP compound words) were collected in two smaller unpublished surveys.

左右 /sayū/ 'left and right'	Associative pairs	6.9
河川 /kasen/ 'rivers'	Synonymous pairs	6.8
燃焼 /nenshō/ 'combustion'	Synonymous pairs	6.8

Note: Participants were asked to evaluation the appropriateness on classifying the compound words according to a particular principle on a 7-point scale, with 1 representing bad examples and 7 good examples. This table is based on the presentation of the data in the 'Word-formation principle classifications' Excel file.

While these word-formation classification evaluations have proved to be extremely valuable in supporting the series of Japanese constituent-morpheme priming experiments, they are not, however, without certain limitations. Principal among these is the fact that the corpus of 1,561 compound words only covers a small proportion of all two-kanji compound words. Moreover, because the word-formation classification surveys focused on relatively high-familiarity two-kanji compound words that are quite transparent semantically,[10] the word-formation classification data alone cannot be used to investigate the extent of morphological involvement in the processing of low-familiarity and semantically-opaque two-kanji compound words. Accordingly, the present authors have recently started conducting a large-scale psychological survey about native Japanese speaker awareness for the morphological structure of two-kanji compound words, in order to support further visual word recognition research into the morphological aspects of two-kanji compound words. The results of our first morphological structure survey involving 11,308 two-kanji compound words, which will be supplemented with future survey results, form the core of the morphological structure component of the database.

The morphological structure survey corpus consists of 11,308 two-kanji compound words selected from the Kōjien headword list, of which both constituents belong to the 1,945 Jōyō kanji list, and have an average frequency of 10 or more over a six-year period of newspaper articles. These compound words were divided into 11 lists (1,028 words per list), and all 11 lists were presented to the native Japanese speaker participants. In contrast to the simpler task in the word-formation classification surveys, where the respondents were merely asked to evaluate the appropriateness of classifying a particular compound word according to a single principle, in this survey respondents were asked to classify the compound words according to five morphological structure categories (M+M, V+C, C+V, SP, and other), and in the cases of the first four categories to evaluate the appropriateness of the classification on a 5-point scale (with 1 corresponding to 'fits this category more than the others' and 5 corresponding to 'definitely this category'). Respondents were also asked to evaluate their familiarity for the pronunciation of the compound word on a 3-point scale (0 = 'not known', 1 = 'known - low confidence', and 2 = 'known - high confidence'). The participants in the first stage of the survey were 9 native Japanese undergraduate and graduate students, who were paid a fee for their efforts. The participants were requested to complete one list of classifications and evaluations a day over an 11-day period, with the presentation order for the lists being counter-balanced among the respondents.

Table 4 shows the numbers of two-kanji compound words classified under the same principle by more than 50 percent of the respondents as a function of morphological structure

[10] Although some of the V+C and C+V compound words have familiarity ratings of 5.0 or above according to the NTT database (Amano & Kondō, 1999), the majority of the surveyed compound words have ratings over 5.5 (on 7-point scales). The generally high classification scores for most of the compound words also indicate these compound words are rather semantically-transparent.

category. In total, 7,593 compound words (67.1% of the 11,308 corpus items) were classified under the same principle by more than 50 percent of the respondents. However, looking at this result from the other perspective, the fact that 3,715 compound words (32.9%) were not consistently classified clearly indicates that the classification task is quite difficult, and that there are relatively few words for which there is a clear consensus about the morphological structure among native Japanese speakers.

Table 4
The Numbers of Two-Kanji Compound Words Classified under the Same Principle by More than 50 Percent of the Respondents as a Function of Morphological Structure Category

Morphological structure	Example	Number	Percentage	Appropriateness rating	Pronunciation familiarity
M+M	熱風	4,596	40.6	4.06	1.98
V+C	止血	1,047	9.3	4.07	1.97
C+V	骨折	1,240	11.0	3.81	1.96
SP	金銭	95	0.8	3.75	1.96
Other	白黒	615	5.4	-	1.91

These points are also reflected in the average appropriateness ratings, presented in Table 4, which show that the respondents did not always have full confidence in their classifications across all the morphological structure categories. It is interesting to note in this context, that apart for a few exceptions, the respondents highly rated their familiar for the pronunciations of the compound words; with 9,605 compound words (84.9%) being rated known with high confidence by all respondents and no items had an average rating of less than 1.[11] Clearly, however, the level of familiarity for the compound words themselves was not a factor behind the general lack of consensus concerning the morphological structure of the compound words. These findings suggest that, similar to the semantic transparency-opaqueness continuum, the distinctions between morphological structure categories are not based on clear discrete boundaries, and that native Japanese speaker awareness for the morphological structures of two-kanji compound words is actually quite fuzzy in nature.

Table 5 shows examples of the morphological structure data component of the database, based on the respondent data collected to date.[12] The table includes five high and five low frequency (based on average newspaper counts) compound words together with the morphological structure classifications (% of respondents), the average appropriateness ratings, and pronunciation familiarity ratings. This morphological structure data will be very useful in supporting further research into the involvement of morphological information in the lexical retrieval and representation of two-kanji compound words, particularly research focusing on the interactions between familiarity, semantic transparency, and morphological structure.

[11] There are two possible factors behind the high pronunciation familiarity ratings; one is that the minimum average newspaper frequency of 10 is quite high, and the second is that because the survey compound words consist of Jōyō kanji, the ratings may be reflecting familiarity for the constituent readings more than for the pronunciation of the compound word.

[12] For the first stage of the morphological structure survey, we sought to establish a large survey corpus, but this has, inevitably, involved fewer respondents. The morphological structure component of the database at the websites will be regularly updated as new survey data is processed.

Table 5
Examples of the Morphological Structure Data Component of the Database with Average Frequency Counts, Percentages of Respondents Classifying the Compound Words under a Morphological Structure Category and Average Appropriateness Ratings, Together with Pronunciation Familiarity Ratings

	Average	Morphological structure (%)					Average appropriateness rating				Pronunciation
Compound	Frequency	M+M	V+C	C+V	SP	Others	M+M	V+C	C+V	SP	familiarity
問題	22,168	22.2	33.3	0	0	44.4	4.5	4.0	-	-	1.9
政府	16,856	66.7	0	0	11.1	22.2	4.0	-	-	1.0	2.0
首相	14,956	33.3	11.1	0	0	55.6	4.3	4.0	-	-	2.0
昨年	11,783	88.9	0	0	0	11.1	4.3	-	-	-	2.0
企業	11,339	33.3	33.3	11.1	0	22.2	3.7	4.0	4.0	-	2.0
余熱	10	33.3	33.3	11.1	0	22.2	3.7	2.3	2.0	-	2.0
両様	10	66.7	0	11.1	0	22.2	4.0	-	4.0	-	1.9
老境	10	66.7	0	11.1	0	22.2	3.7	-	2.0	-	1.7
論集	10	33.3	0	55.6	0	11.1	3.7	-	3.8	-	2.0
和合	10	0	0	33.3	33.3	33.3	-	-	3.3	2.3	1.9

4 Semantic Category Data

Complementing the morphological family data outlined in Part 2, which emphasizes shared constituent morphemes, and the morphological structure data described in Part 3, concerned with the relationships between constituent morphemes, the semantic category data in the present database focuses on compound word meaning. Specifically, this component of the database consists of semantic category codes from the National Institute for Japanese Language's (NIJL) (2004) semantically-classified word list for 24,519 two-kanji compound words (35.54% of the 68,992 Kōjien orthographic words).

The NIJL (2004) word list classifies approximately 96,000 modern Japanese words and expressions according to 895 semantic categories. In addition to semantic themes, such as abstract relations, human activity, and products and implements, the words are also classified in terms of word class, distinguishing nouns, verbs, modifiers, and other parts of speech, with the corresponding codes prefixed with 1, 2, 3, and 4 respectively. In order to add these codes the present database, the two-kanji compound word entries in the NIJL word list were input into an Excel file together with the corresponding code (or codes in the cases of polysemous words). Although NIJL entries consisting of a two-kanji compound word and the dummy verb する /suru/ 'do' were included, phrasal entries (i.e., where the compound word was part of a longer expression) were not. As a result of comparing this list with the Kōjien compound words, it was found that 24,519 of the Kōjien items are assigned a semantic category code in the NIJL word list.

Table 6
Examples of Two-Kanji Compound Words in Semantic Sub-Categories Adjacent to the Sub-Category 1.5110.15 Containing the Compound Word 亜鉛 /aen/ 'Zinc'

Category Code	Two-Kanji Compound Word Members
1.5110 元素 'elements'	
1.5110.09	鉄分 鋼鉄 砂鉄 銑鉄 鋳鉄 鉄鋼 軟鉄 練鉄 錬鉄
1.5110.10	水銀
1.5110.15	亜鉛
1.5110.25	黄燐 赤燐
1.5110.26	硫黄

Note: The table only includes the nearest sub-categories (two prior and two subsequent) to the sub-category 1.5110.15 that have two-kanji compound word members. This table is based on the presentation of the data in the 'Semantic category data-Category' Excel file.

These two-kanji compound words are listed together with the codes in the 'Semantic category data-Compounds' file at the database web site. Table 6 presents examples of compound words in the neighboring sub-categories to the sub-category 1.5110.15 which contains the compound word 亜鉛 /aen/ 'zinc'. The semantic category code data makes it easy to group these two-kanji compound words by general semantic themes, opening up interesting possibilities for investigating the contribution that compound word meaning makes to the word recognition process and how this interacts with the meanings of the constituents according to the morphological structure of the compound words. For instance, Masuda (2002b) has reported that both 信号 /shingō/ 'signal' and 信仰 /shinkō/ 'faith, belief, creed', which both have 信 /shin/ 'believe' as a constituent, facilitated responses to 宗教 /shūkyō/ 'religion'. While the phonological overlap between these primes may have been a factor in those results, the semantic category data would be very useful in teasing apart the orthographic, phonological and semantic contributions to the visual word recognition of compound words, because although both 信仰 and 宗教 are classified under the category 1.3047 信仰・宗教 'faith/religion', 信号 is classified under the different category of 1.3121 合図 'sign/signal'.

In summary, this paper has reported on the construction of large-scale database of two-kanji compound words, highlighting in particular the morphological family data, the morphological structure data and the semantic category data components of the database. The authors are building this database to support and extend research into the lexical retrieval and representation of two-kanji compound words within the Japanese mental lexicon from the perspective of compound word morphology, such as the series of constituent-morpheme priming experiments discussed, in the hope of deepening our understanding of how morphological information relating the structures of polymorphemic words is represented within the mental lexicon.

References

Amano, S., & Kondō, T. (1999). *Nihongo no goitokusei* [Lexical properties of Japanese] (Vols. 1-6, NTT database series). Tokyo: Sanseidō.

Amano, S., & Kondō, T. (2000). *Nihongo no goitokusei* [Lexical properties of Japanese] (Vol. 7; Frequency, NTT database series). Tokyo: Sanseidō.

Andrews, S. (1992) Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 234-254.

Baayen, R.H., Lieber, R., & Schreuder, R. (1997). The morphological complexity of simplex nouns. *Linguistics*, *35*, 861-877.

Butterworth, B. (1983). Lexical representation. In B. Butterworth, (Ed.), *Language production: Volume 2 Development, writing and other language processes* (pp. 257-294). London, England: Academic Press.

Caramazza, A., Laudanna, A., & Romani, C. (1988). Lexical access and inflectional morphology. *Cognition*, *28*, 297-332.

Coltheart, M., Davelaar, E., Jonasson, J.T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.). *Attention and performance VI* (pp. 535-555). Hillsdale, NJ: Lawrence Erlbaum Associates.

Coulmas, F. (1996). *The Blackwell encyclopedia of writing systems*. Oxford, England: Blackwell.

Drews, E. (1996). Morphological priming. *Language and Cognitive Processes*, *11*, 629-634.

Feldman, L.B. (Ed.). (1995). *Morphological aspects of language processing*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Fowler, C.A., Napps, S.E., & Feldman, L. (1985). Relations between regular and irregular morphologically related words in the lexicon as revealed by repetition priming. *Memory & Cognition*, *13*, 241-255.

Fushimi, T., Ijuin, M., Patterson, K., & Tatsumi, I.F. (1999). Consistency, frequency, and lexicality effects in naming Japanese kanji. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 382-407.

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language*, *29*, 228-244.

Haas, W. (1976). Writing: The basic options. In W. Haas (Ed.), *Writing without letters* (pp. 131-208). Manchester, England: Manchester University Press.

Haas, W. (1983). Determining the level of a script. In F. Coulmas, & K. Ehlich (Eds.), *Writing in focus* (pp. 15-29) Berlin, Germany: Mouton.

Hirose, H. (1992). Jukugo no nichi katei ni kan suru kenkyū: Puraimingu hō ni yoru kentō [An investigation of the recognition process for jukugo by use of priming paradigms]. *The Japanese Journal of Psychology*, *63*, 303-309.

Jarema, G., Kehayia, E., & Libben, G. (Eds.). (1999) Mental lexicon [Special issue]. *Brain and Language, 68(1/2).*

Joyce, T. (1999). Lexical access and the mental lexicon for two-kanji compound words: A priming paradigm study. *Proceedings of the 2nd International Conference on Cognitive Sciences and 16th Annual Meeting of the Japanese Cognitive Science Society Joint Conference*, 27-30 July, Tokyo, Japan, 511-514.

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

Joyce, T. (2003a). Frequency and verb-morphology effects for constituents of two-kanji compound words. Poster session presented at the *4th Tsukuba International Conference on Memory* (Human Learning and Memory: Advanced in Theory and Application), 11-13 January, Tsukuba, Japan.

Joyce, T. (2003b). Kanji niji jukugo ni okeru dōshi kōzō yōso no ichiteki hindo [Positional frequency of verbal constituents within two-kanji compound words]. *Proceedings of the 67th Meeting of the Japanese Psychological Association*, 13-15 September 2003, Tokyo University, Tokyo, Japan, p. 590.

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

Joyce, T., & Masuda, H. (2004). Kōsei keitaiso toshite no kanji no tanjikan senkō teiji ga kanji niji jukugo no goihandan ni oyobosu puraiminngu kōka [Priming effects from brief presentations of constituent kanji on lexical decisions for two-kanji compound words] *Proceedings of the 68th Meeting of the Japanese Psychological Association*, 12-14 September 2004, Kansai University, Osaka, Japan, p. 613.

Joyce, T. & Ohta, N. (1999). The morphology of two-kanji compound words: Data from a word-formation classification survey. *Tsukuba Psychological Research*, *22*, 45-61.

Joyce, T., & Ohta, N. (2002). Constituent morpheme frequency data for two-kanji compound words. *Tsukuba Psychological Research*, *24*, 111-141.

Kageyama, T. (1982). Word formation in Japanese. *Lingua*, *57*, 215-258.

Kawakami, M. (1997). JIS isshu kanji 2965 ji o mochiite sakusei sareru kanji niji jukugo sūhyō: Macintosh ban iwanami kōjien daiyonban ni motozuku ruijigosū chōsa [Numerical data for two-kanji compound words formed from 2965 JIS level 1 kanji characters: Survey of synonyms based on Iwanami's Kojien dictionary (4th edition CD for Macintosh)]. *Bulletin of the School of Education* (Nagoya University), *44*, 243-299.

Kawakami, M. (2000). JIS isshu kanji 2965 ji o mochiita kanji niji jukugo no ruiseki ruijigo hindo hyō [A table of cumulative frequencies of Japanese neighbor-kanji-compound words] *The Science of Reading*, *44*, 150-159.

Kehayia, E., Jarema, G., Tsapkini, K., Perlak, D., Ralli, A., & Kadzielawa, D. (1999). The role of morphological structure in the processing of compounds: The interface between linguistics and psycholinguistics. *Brain and Language*, *68*, 370-377.

Masuda, H. (2002a). Phonological effect on false recognition of Japanese kanji compounds in two-word displays. *Abstracts of the Third International Conference on the Mental Lexicon,* 103. The Banff Centre, Banff, Canada.

Masuda, H. (2002b). Semantic activation of component kanji characters in reading Japanese two-kanji compounds [Abstract]. *Proceedings of 10th International Conference on the Cognitive Processing of Chinese and Other Related Asian Languages*, 87. National Taiwan University, Taipei, Taiwan.

Monsell, S. (1985). Repetition and the lexicon. In A. W. Ellis (Ed.), *Progress in the Psychology of Language, Vol. 2*. London: Lawrence Erlbaum Associates.

National Institute for Japanese Language (2004). *Bunrui goi hyō—Zōhokaiteipan* [Word list by semantic principles: Revised and enlarged edition]. (Source 14). Tokyo: Dainihon Tosho.

National Language Research Institute (1962). *Gendai zasshi kyūjū shu no yōgo yōji* [Vocabulary and characters in 90 current magazines] (Research reports 21, 22, and 25). Tokyo: Shuei Shuppan.

National Language Research Institute (1970). *Denshi-keisanki ni yoru shinbun no goi-chōsa* [Studies on the vocabulary of modern newspapers, Volume 1]. (Research report 37). Tokyo: Shuei Shuppan.

National Language Research Institute (1997). *Gendai zasshi kyūjū shu no yōgo yōji* [Vocabulary and characters in 90 current magazines] Floppy disk version (National Language Research Institute Language Processing Data Vol. 7). Tokyo: Sanseidō.

Nomura, M. (1988). Niji kango no kōzō [The structure of 2 kanji Sino-Japanese words]. *Nihongogaku*, *7*, 5, 44-55.

Ogawa, T., & Saito, H. (2001). Kanji niji jukugo no shikakuteki ninchi ni okeru kinbōgogun no kasseika katei — Zenkinnteki kaijo kadai o mochiita kenntō [The activation of neighbors in the visual recognition of two-kanji compound words: A progressive demasking task study]. *Proceedings of the 65th Meeting of the Japanese Psychological Association*, 7-11 November 2001, University of Tsukuba, Tsukuba, Japan, p. 215.

Ogawa, T., Saito, H., & Yanase, Y. (2005). Niji jukugo no gokeisei ni okeru JIS daiichi suijun ni zokusuru kanji 2,965 ji no ketsugoo tokusei [The combinatory characteristics of the 2,965 JIS level 1 kanji in the word formation of two-kanji compound words]. *The Japanese Journal of Psychology*. 76, 269-275.

Saito, H. (1997). Shinteki jisho [Mental lexicon]. In Y. Matsumoto, T. Kageyama, M. Nagata, H. Saito, T. Tokunaga, *Iwanami Kooza Gengo no kagaku 3 Tango to jisho* (pp. 94-153). Tokyo: Iwanami Shoten.

Sandra, D. (1990). On the representation and processing of compound words: Automatic access to constituent morphemes does not occur. *Quarterly Journal of Experimental Psychology*, *42A*, 529-567.

Sandra, D. (1994). The morphology of the mental lexicon: Internal word structure viewed from a psycholinguistic perspective. *Language and Cognitive Processes*, *9*, 227-269.

Sandra, D., & Taft, M. (Eds.) (1994). Morphological structure, lexical representation and lexical access [Special issue] *Language and Cognitive Processes*, *9(3)*.

Schreuder, R., & Baayen, R.H. (1995). Modeling morphological processing. In Laurie Beth Feldman, (Ed.), *Morphological aspects of language processing* (pp. 131-154). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schreuder, R., & Baayen, R.H. (1997). How complex simplex words can be. *Journal of Memory and Language*, *37*, 118-139.

Shinmura, Izuru, (Ed.). (1995). Kōjien. (Fifth edition). Tokyo: Iwanami.

Taft, M. (1991). *Reading and the mental lexicon*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Taft, M. (1994). Interactive-activation as a framework for understanding morphological processing. *Language and Cognitive Processes*, *9*, 271-294.

Taft, M., & Forster, K.I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*, 638-647.

Taft, M., & Forster, K.I. (1976). Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior*, *15*, 607-620.

Tamaoka, K., & Hatsuzuka, M. (1998). The effects of morphological semantics on the processing of Japanese two-kanji compound words. *Reading and Writing*, *10*, 293-322.

Wydell, T.N., Patterson, K.E., & Humphreys, G.W. (1993). Phonologically mediated access to meaning for kanji: Is a rows still a rose in Japanese kanji? *Journal of Experimental Psychology Learning, Memory, and Cognition*, *19*, 491-514.

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

Glottometrics 10, 2005, 45-54

A Corpus Investigation of the *Right-hand Head Rule* Applied to Japanese Affixes [1]

Yayoi Miyaoka [2]
Hirohsima University of Economics, Japan
Katsuo Tamaoka [3]
Hiroshima University, Japan

Abstract: The present study investigates differences between Japanese prefixes and suffixes using editions of the *Asashi Newspaper* published between 1985 and 1998 (Amano & Kondo, 2000). The *right-hand head rule* (e.g., Kageyama, 1982; Kageyama, 1999; Namiki, 1982; Nishigauchi, 2004; Williams, 1981) predicts that prefixes would be attached to a wide variety of nouns while suffixes would be regularly attached to a smaller group of nouns. Twenty-four frequently-used affixes consisting of 12 prefixes and 12 suffixes were compared according to 7 corpus features, including printed-frequency, productivity, accumulative productivity, commonality, coalescence degree, Herdan's logarithmic function of type-token ratio (log TTR), and entropy. Although a series of Mann-Whitney *U*-tests calculated for the six corpus features of printed-frequency, productivity, accumulative productivity, commonality, coalescence degree and log TTR did not reveal any differences between the 12 prefixes and the 12 suffixes, the *t*-test for entropy indicated a significant difference. This suggests that the prefixes were more randomly or chaotically attached to nouns than the suffixes. Although the present findings are limited only to the selected 24 affixes, the result supported the *right-hand head rule*.

Keywords: Japanese affix, prefix, suffix, right-hand head rule, coalescence degree, Herdan's logarithmic function of type-token ratio (log TTR), entropy

1. Background of Japanese prefixes and suffixes

In general, there are three types of morphological word formation: compounding, derivation, and inflection (Kageyama, 1993). Morphological elements having independent units of meaning are called 'bases', while those that are connected to bases are called 'affixes'. Furthermore, there are four types of affixes in the languages of the world: 'prefixes', which are placed at the head of the base (e.g. *dis* in '*dis*like' in English); 'infixes', which appear within a word (e.g. *um* in 'k*um*ain' for 'ate' in Tagalog), 'suffixes' which are added to the end of the base (e.g. *er* in 'report*er*' in English), and 'circumfixes' (e.g. *baik* meaning 'good', and *ke*baik*an* 'goodness' in Indonesian). The only two types of affixes used in modern Japanese are prefixes and suffixes. For example, the Japanese prefix 不 meaning 'un-' is added to 自然

[1] The present study was supported by the Research Institute of Science and Technology for Society (RISTEX) at the Japan Science and Technology Corporation (JST), Kawaguchi, Japan.

[2] Address correspondence to: Yayoi Miyaoka, Hiroshima University of Economics, 37-1, 5-chome, Gion, Asaminami-ku, Hiroshima, Japan 731-0192. E-mail: y.miya8411@hue.ac.jp.

[3] Katsuo Tamaoka, International Student Center, Hiroshima University 1-1, 1-chome, Higashihiroshima, Japan 739-8524. E-mail: ktamaoka@hiroshima-u.ac.jp.

meaning 'natural', creating the morphologically complex word 不自然 meaning 'unnatural'. Similarly, another prefix 全 meaning 'whole' is added to the base 世界 'world', producing the word 全世界 meaning 'whole world'. An example of a suffix is 的 which forms an adjective; it can be added to the end of the base 建設 'construction' to create a compound word 建設的 'constructive'. Similarly, the suffix 性 forms a noun, so that when it is added to the end of the base 安全 'safe' becomes the compound word 安全性 meaning 'safety'.

Nomura (1977) explained that the base is the semantic core element of a word, while the affix element adds meaning to the base and determines its grammatical category. Thus, the base can stand alone as a word, but the affix cannot be a word by itself. In previous studies (e.g., Kageyama, 1999; Nomura, 1977), it has been shown that prefixes add meaning to the base without changing the grammatical category, whereas suffixes not only add meaning but may also change the grammatical category of the base. For instance, 全世界 'whole world' is composed of the prefix 全 'whole' and the noun 世界 'world': the attachment of a prefix to the noun does not result in a change in grammatical category. On the other hand, the suffix 的 in 建設的 'constructive' changes the noun 建設 'construction' into an adjective. Thus, the hypothesis proposed by Williams (1981) that the right-hand side of a complex word determines the grammatical category of the word holds true not only for English but also for Japanese. This tendency is often referred to as the *right-hand head rule* (e.g., Kageyama, 1982; Kageyama, 1999; Namiki, 1982; Nishigauchi, 2004). This key difference between prefixes and suffixes may also affect the extent to which they co-occur with various nouns in printed-frequency. Prefixes do not change the grammatical category of the nouns to which they are attached, so they may be attached to a wide variety of nouns. By contrast, some suffixes change the grammatical category of a noun; this limits the range of nouns with which they may appear and results in the regular combination of a small group of certain suffixes and nouns.

In his seminal work, *A Mathematical Theory of Communication* (1948), American mathematician Claude Elwood Shannon (1916-2001) first developed the concepts of *entropy* and *redundancy* for information processing. *Entropy* is an index of the degree of disorder or chaos; *redundancy* refers to the degree of superfluousness. Since these two concepts can be applied to a wide range of corpus sizes, characteristics of prefix and suffix attachments to a variety of nouns can be directly compared (for details, see Hori, 1979 and Kaiho, 1989; for an example of an actual corpus study which applied these concepts, see Tamaoka, Miyaoka & Lim, 2003; Tamaoka, Lim & Sakai, 2004). Therefore, the present study utilizes the index of *entropy* to analyze co-occurring frequencies of affixes and nouns, hypothesizing that prefixes show higher entropy since they would be expected to be attached to a wider variety of nouns.

2. Selection of prefixes and suffixes

In the present study, 12 prefixes and 12 suffixes, all commonly used, were compared. The prefixes were 大 'big', 不 'un-', 無 'un-', 新 'new', 初 'first', 非 'un-', 全 'whole', 再 're-', 超 'super', 反 'anti', 未 'not yet', 毎 'every'. The suffixes were 的 '-tive/-like', 者 'person', 性 'nature', 学 'studies', 化 'characteristic', 論 'theory/-logy', 家 '-ist', 式 'manner/style', 界 'world', 風 'style', 状 'state', 用 'use'. Since these target affixes are the most commonly-seen items, they were sufficient to investigate the actual usage of Japanese affixes in a corpus. The simple printed-frequencies of the 24 kanji symbols used for the affixes are shown in Table 1.

3. Identifying prefix, suffix and base

There are some unresolved issues regarding the definitions of the prefix, suffix, and base in Japanese. The present study defines a base as an element that can be a single word by itself. Therefore, an element without an affix is presented in a single kanji. For instance the word 新世代 meaning ‘new generation’ is composed of the noun 世代 ‘generation’ attached to the prefix 新 ‘new’. In this case, 世代 is the base, as it can stand as a single word without the prefix. The base is easy to identify in this example, since 世代 is a two-kanji compound word. However, the base in an example such as 株式 ‘stock’ is more difficult to identify. Because the single kanji 株 can be found in the dictionary as a single word meaning ‘stump’ or ‘stock’, the present study interpreted the word 株式 in such a way that 式 is the suffix while 株 serves as the base. This is not to suggest that 式 can always be identified as a suffix, as by itself 式 also means ‘ceremony’. Accordingly, in the word 卒業式 meaning ‘graduation ceremony’, 式 is not considered to be a suffix. In the same way, the suffix ‘状’ has two meanings of ‘state’ and ‘letter’ when standing alone; only 状 meaning ‘state’ is considered a suffix in the present study.

Some single-kanji affixes have more than one pronunciation. For instance, 大 meaning ‘big’ is pronounced in two ways: /dai/ in On-reading (a Chinese-originated sound), and /oR/ (/R/ refers to a long vowel) in Kun-reading (a Japanese-originated sound). The pronunciation of 大 varies depending on how the prefix is attached to bases. The word 大混乱 meaning ‘big confusion’ is pronounced /dai+koNraN/ (/N/ refers to a nasal) while the word 大津波 meaning ‘big tsunami’ or ‘big seismic sea wave’ is pronounced /oR+tunami/. The meaning of the prefix 大 remains the same in both instances. So the present study regards these two different pronunciations as belonging to the same affix 大.

4. Selection of words incorporating the 24 selected affixes

The present study used a lexical corpus of the *Asahi Newspaper* printed from 1985 to 1998, produced by Amano and Kondo (2000). This corpus contains 341,771 types and 287,792,797 tokens. All words co-appearing with the selected 24 affixes were extracted from the *Asahi Newspaper* corpus using the software called EasySrch (Amano & Kondo, 2003). For example, 42 compound word types co-appearing with the prefix 再 meaning ‘re-’ were found, including 再検討 ‘reconsideration’ (3,859 tokens), 再確認 ‘reconfirm’ (3,109 tokens), 再開発 ‘redevelopment’ (2,704 tokens), 再構築 ‘reconstruction’ (2,012 tokens), 再評価 ‘reevaluation’ (1,069 tokens), and so on. This word selection process was applied to all 24 selected affixes. The base elements could be any type of Sino-Japanese word (*wa-go*), Chinese-originated word (*kan-go*), or loanword (commonly called *gairai-go* or *katakana-hyooki-go*). Nine corpus features were calculated for each affix using the lexical frequency index of the *Asashi Newspaper* corpus.

5. Calculating corpus features for the 24 affixes

Using the data collection process explained in Section 4, the present study calculated nine different corpus features, including a simple addition of *printed-frequency*, *productivity*, *accumulative productivity* and a more complex calculation of *entropy.* The calculation proces-

Table 1
Corpus features of prefixes and suffixes

Affix Type	Kanji	Printed-freq.	Productivity	Acc. Prod.	Affix-used ratio	Commonality	log TTR	Entropy	H-maximum
Prefixes	大	2,103,545	359	192,138	0.091	0.132	0.484	5.395	8.488
	不	446,439	158	85,223	0.191	0.093	0.446	5.272	7.304
	無	222,977	150	60,732	0.272	0.098	0.455	5.361	7.229
	新	767,768	132	49,477	0.064	0.368	0.452	3.804	7.044
	初	302,643	90	11,189	0.037	0.215	0.483	4.246	6.492
	非	108,303	58	31,364	0.290	0.193	0.392	3.624	5.858
	全	555,403	44	16,876	0.030	0.298	0.389	3.541	5.459
	再	190,673	42	20,971	0.110	0.184	0.376	4.013	5.392
	超	91,843	40	9,879	0.108	0.217	0.401	3.500	5.322
	反	295,758	30	6,957	0.024	0.329	0.384	2.924	4.907
	未	74,563	19	6,430	0.086	0.225	0.336	3.295	4.248
	毎	61,715	16	47,397	0.768	0.369	0.258	2.443	4.000
Suffixes	的	833,709	342	158,849	0.191	0.054	0.487	6.382	8.418
	者	1,080,160	310	415,725	0.385	0.129	0.443	5.461	8.276
	性	453,110	216	130,676	0.288	0.481	0.456	3.731	7.755
	学	781,148	178	36,747	0.047	0.125	0.493	5.533	7.476
	化	578,610	167	172,339	0.298	0.129	0.424	5.146	7.384
	論	265,332	141	5,298	0.020	0.184	0.577	5.177	7.140
	家	488,437	113	101,770	0.208	0.280	0.410	3.861	6.820
	式	229,697	40	23,460	0.102	0.922	0.367	0.669	5.322
	界	324,754	35	11,468	0.035	0.458	0.380	2.809	5.129
	風	102,317	22	956	0.009	0.270	0.450	3.443	4.459
	状	186,124	13	809	0.004	0.575	0.383	1.943	3.700
	用	495,611	6	2,721	0.005	0.680	0.227	1.337	2.585

Note 1 : 'Acc. Prod.' refers to accumulative productivity.

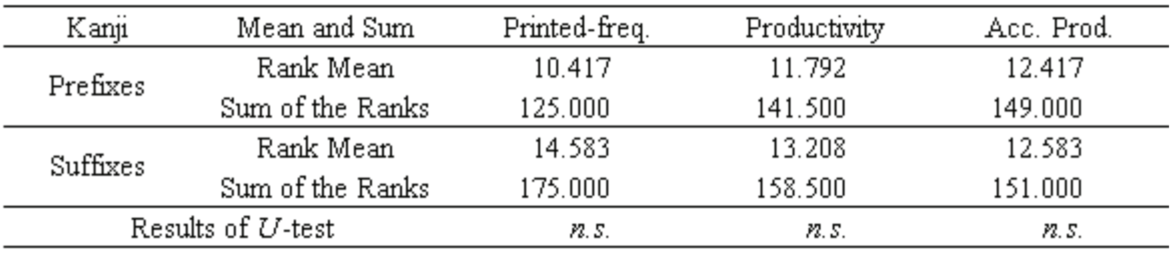
Note 2 : 'log TTR' refers to a type-token ratio calculated by legalism of type frequency divided by legalism of token frequency.

ses are explained in the following sub-sections. The figures for the 12 prefixes and 12 suffixes with regard to each of the nine features are presented in Table 1.

5.1 Printed-frequency, productivity and accumulative productivity

Since the three corpus features of *printed-frequency*, *productivity* and *accumulative productivity* are simple frequency counts, a nonparametric analysis of the Mann-Whitney *U*-test was used to compare the 12 prefixes and the 12 suffixes. The mean rank and the sum of the ranks are shown in Table 2.

Table 2
Comparisons between prefixes and suffixes by Mann-Whitney *U*-test



Kanji	Mean and Sum	Printed-freq.	Productivity	Acc. Prod.
Prefixes	Rank Mean	10.417	11.792	12.417
	Sum of the Ranks	125.000	141.500	149.000
Suffixes	Rank Mean	14.583	13.208	12.583
	Sum of the Ranks	175.000	158.500	151.000
Results of *U*-test		*n.s.*	*n.s.*	*n.s.*

Note: A sample size of prefixses and suffixes is 12 each.

The *printed-frequencies* for the 24 different kanji symbols used for the affixes considered in this study were derived from the lexical corpus of the *Asashi Newspaper* (Amano & Kondo, 2000). The printed-frequency of the 12 prefixes (rank mean = 10.417, sum of the ranks = 124.000) was not significantly larger than that of the 12 suffixes (rank mean = 14.583, sum of the ranks = 175.000) [$U = 47.000$, $p = .160$, *n.s.*]. This implies that simple frequencies of kanji symbols encoding affixes do not distinguish between prefixes and suffixes.

The *productivity* feature indicates in how many words each affix is combined with nouns (i.e., type frequency of words with the given affix). As shown in Table 1, the prefix 大 had the greatest productivity at 359 words, while the suffix 的 had the greatest productivity at 342 words. The prefix 毎 had the lowest productivity at 16 words, while the suffix 用 had the lowest productivity at only 6 words. The rank mean of productivity for the 12 prefixes (rank mean = 11.79, sum of the ranks = 141.50) was not significantly larger than that of the 12 suffixes (rank mean = 13.21, sum of the ranks = 158.50) [$U = 63.500$, $p = .630$, *n.s.*]. In other words, these 12 prefixes are attached to words as often as the 12 suffixes.

The *accumulative productivity* was calculated by summing the printed-frequencies of all the words with the given affix. In other words, *productivity* is type frequency whereas *accumulative productivity* is token frequency. Similarly to productivity, a U-test for accumulative productivity indicates no difference between the 12 prefixes (rank mean = 12.42, sum of the ranks = 149.00) and the 12 suffixes (rank mean = 12.58, sum of the ranks = 151.00) [$U = 71.00$, $p = .977$, *n.s.*].

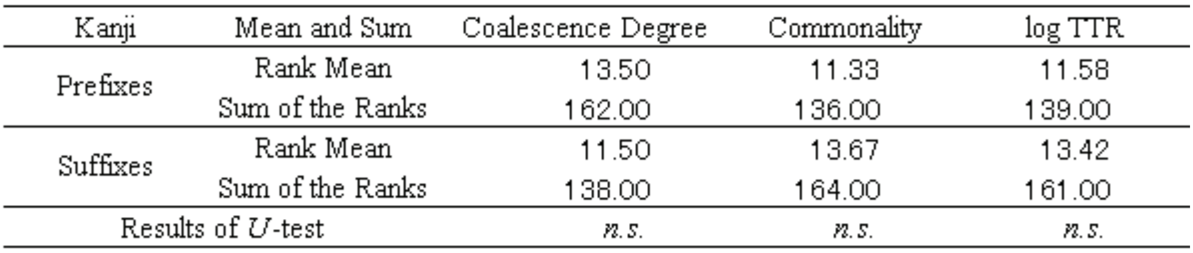
A nonparametric rank-order correlation coefficient of Spearman's rho was computed for all the 24 affixes together based on the three features mentioned above, since there were no differences between the 12 prefixes and the 12 suffixes. Correlation coefficients of all combinations of the three variables were significant; the correlation between printed-frequency and productivity [$r_s(24) = .704$, $p < .001$], the correlation between printed-frequency and accumulative productivity [$r_s(24) = .609$, $p < .01$], and the correlation between productivity and accumulative productivity [$r_s(24) = .813$, $p < .001$]. These frequency features have strong interrelations.

5.2 Coalescence degree, commonality and Herdan’s type-token ratio (log TTR)

Since simple type and token frequencies did not show any difference between prefixes and suffixes, somewhat more complex features of their behavior in the corpus were also calculated. The *coalescence degree* was calculated by dividing accumulative productivity by printed-frequency for a given affix. For example, the suffix 家 was printed 488,437 times (including proper nouns in the *Asashi Newspaper* [4]) and accumulative productivity 101,770 times. Thus, coalescence degree was 0.028 (101,770 divided by 488,437). The rank mean of the coalescence degree for the 12 affixes (rank mean=13.50, sum of the ranks=162.00) was not significantly larger than that of the 12 suffixes (rank mean=11.50, sum of the ranks=138.00) [*U*=60.00, *p*=.514, *n.s.*]. Thus, there was no difference between prefixes and suffixes in the number of times that the kanji were used for affixes.

Table 3

Comparisons between prefixes and suffixes by Mann-Whitney *U*-test



Kanji	Mean and Sum	Coalescence Degree	Commonality	log TTR
Prefixes	Rank Mean	13.50	11.33	11.58
	Sum of the Ranks	162.00	136.00	139.00
Suffixes	Rank Mean	11.50	13.67	13.42
	Sum of the Ranks	138.00	164.00	161.00
Results of *U*-test		*n.s.*	*n.s.*	*n.s.*

Note : '*n.s.* ' refers to not significant.

The *commonality* refers to how often the most frequently-used word with a target affix occupy the total accumulative frequencies of all the words with a target affix. In the case of the prefix 不, the most frequently-used compound noun was 不十分 consisting of the prefix 不 and the two-kanji compound word 十分 meaning ‘sufficient’. This word appeared 7,965 times in the corpus. Since the total accumulative productivity (i.e., the token frequency of all words with the prefix 不) was 85,223, the commonality was calculated by dividing the frequency of the most frequently-used word by the accumulative productivity. Thus, commonality for 不 was 0.093 (7,965 divided by 85,223). The rank mean of commonality for the 12 affixes (rank mean = 11.33, sum of the ranks = 136.00) was not significantly larger than that of the 12 suffixes (rank mean = 13.67, sum of the ranks = 164.00) [$U = 58.000$, $p = .443$, *n.s.*].

The *log type-token ratio* (log TTR) quantifies how many words with the target affix are in the corpus and how frequently they are used. However, the result of simple TTR is almost always equal to zero in a corpus. In the present study, as proposed by Wimmer and Altmann (1999) as one of the candidate calculations, a logarithmic function of TTR by Herdan (1960) was utilized to compare the 12 prefixes and the 12 suffixes. The calculation is simply:

[4] Tamaoka and Makioka (2004) calculated frequencies for the 1,945 basic Japanese kanji without proper nouns in order to avoid biases from specific popular incidents or events reported in the *Asashi Newspaper*.

$$TTR = \frac{\ln V}{\ln N}$$

where V is the number of words (i.e. productivity or type frequency) and N is the number of all accumulative frequencies of words (i.e. accumulative productivity or token frequency). For example, the suffix 界 was attached to 35 words (productivity) which appeared 11,468 times (accumulative productivity) in the corpus. Thus, the log TTR for 界 becomes 0.380 (log35 divided by log11,468). The rank mean of log TTR for the 12 affixes (rank mean = 11.58, sum of the ranks = 139.00) was not significantly larger than that of the 12 suffixes (rank mean = 13.42, sum of the ranks = 161.00) [U = 61.000, p = .551, *n.s.*][5].

Spearman's rho was computed for all the 24 affixes together. Rank-order correlation coefficients between coalescence degree and log TTR[$r_s(24) = -.839$, $p < .001$] was significant. However, the other two correlations between coalescence degree and commonality [$r_s(24) = -.383$, $p = .064$, *n.s.*] and between commonality and log TTR [$r_s(24) = .064$, $p = .765$, *n.s.*] were not significant. Coalescence degree and log TTR seem to indicate similar features while commonality differs.

Simple calculations based on type and token frequency manipulations did not indicate any differences between the 12 prefixes and the 12 suffixes. Thus, a more complex mathematical concept of *entropy* was applied to compare them.

5.3 Entropy

The feature *entropy* refers to how randomly a single affix is combined to various base words. It is calculated using the following formula.

$$H = -\sum_{j=1}^{J} p_j \log_2 p_j$$

[5] A test of difference between the 12 prefixes and the 12 suffixes using Herdan's TTR, the following asymptotic formula, should be used

$$z = \frac{\overline{TTR}_{prefix} - \overline{TTR}_{suffix}}{\sqrt{\frac{1}{12^2}Var(S_{prefix}) + \frac{1}{12^2}Var(S_{suffix})}}$$

which has a standard normal distribution. For this calculation, each affix variance should be calculated by the following:

$$Var(TTR) = \frac{N^2\sigma^2}{V^3\overline{\mu}_1^4 \ln^2 N}$$

The variance of the mean for each of the 12 prefixes and the 12 suffixes is computed by the following.

$$Var(S) = \frac{1}{12^2}\sum_{i=1}^{12} Var(TTR_i)$$

However, in the present study, we judged that Mann-Whitney U-test is good enough for computing Herdan's TTR for the 12 prefixes and the 12 suffixes.

In the present study, the entropy of affixes was calculated according to the base to which they were attached. For example, the prefix 超 'super' appeared to be attached to 40 different nouns in the newspaper corpus. The total number of times that the prefix 超 appeared with base nouns was 9,879 times, as seen with the highest frequency of 超伝導 'superconductivity' counted 2,140 times, the second highest frequency of 超党派 'nonpartisan' at 1,794 times, and the third highest frequency of 超大国 'super-power nation' at 1,736 times. The 'p' in the formula stands for the relative frequency of occurrence of a specific compound word among all compounds attached to affixes. In the case of the highest frequency of 超伝導, 'p' is 0.217, as calculated by dividing 2,140 by 9,879. The formula $\log_2 P_j$ for this word is evaluated as $\log_2 0.217 = -2.207$. Then, '$p_j \log_2 p_j$' for the 超伝導 is -0.479 (the result of 0.217 × -2.207). The values for the remaining 39 compound nouns were also calculated in the same manner. The entropy of 超 was finally determined as 3.500 by adding all the scores of $\log_2 P_j$ and multiplying by -1.

The variance of entropy is calculated by

$$\mathrm{V(H)} = \frac{1}{N}\left(\sum_j p_j \log_2^2 p_j - H^2\right) .$$

In this formula, N is the sum of all frequencies, $\log_2$ is the logarithm to base 2 and $\log^2$ is $(\log p)^2$. The standard deviation of means is given by

$$\sigma_{\bar{H}} = \sqrt{\frac{V(H)}{n}} .$$

The *t*-test for entropy is then calculated as

$$t = (\bar{H}_{pref} - \bar{H}_{suf}) / \sqrt{\sigma^2_{\bar{H}_{pref}} + \sigma^2_{\bar{H}_{suf}}} .$$

A *t*-value of the above formula was 2.096. The difference in entropy between the 12 prefixes and the 12 suffixes is significant [$t(22) = 2.096$, $p < .05$]. Thus, entropy, referring to the degree of affix attachment disorder, reveals a significant difference between the 12 prefixes and the 12 suffixes. The mean entropy of the 12 prefixes was 3.952 while the mean entropy of the 12 suffixes was 3.791. The result of the *t*-test suggested that the prefixes were more randomly or chaotically attached to nouns than the suffixes. Although this finding is limited to the selected 24 affixes, the result supported the *right-hand head rule* (Kageyama, 1982; Kageyama, 1999; Namiki, 1982; Nishigauchi, 2004; Williams, 1981).

6. Conclusion

The present corpus study assumed that Japanese affixes would generally follow the *right-hand head rule*, which predicts that prefixes would be attached to a wide variety of nouns while suffixes would be regularly attached to a smaller group of nouns. Twenty-four frequently-used affixes (12 prefixes and 12 suffixes) were compared with regard to seven corpus features. A series of Mann-Whitney *U*-tests calculated for the first six corpus features (printed-frequency, productivity, accumulative productivity, commonality, coalescence degree and log TTR) did not reveal any differences between the 12 prefixes and the 12 suffixes. Simple frequency counts and their ratios seem not to be able to distinguish between characteristics of the prefixes and suffixes attached to nouns. However, the *t*-test for entropy indicates a significant difference. This result suggested that the prefixes were more randomly or chaotically attached to nouns than the suffixes. Although the present findings are limited to the selected 24 affixes, this result supported the *right-hand head rule* proposed by various linguists (e.g., Kageyama, 1982; Kageyama, 1999; Namiki, 1982; Nishigauchi, 2004; Williams, 1981).

References[6]

Amano, N., & Kondo, K. (2000). *Nihongo-no goi tokusei – Dai-7-kan [Lexical properties of Japanese – Volume 7]*. Tokyo: Sanseido.

Amano, N., & Kondo, K. (2003). *Nihongo-no goi tokusei – Dai-2-ki CD-ROM-ban [Lexical properties of Japanese – The second volume of CD-ROM verison]*. Tokyo: Sanseido.

Herdan, G. (1960). *Type-token mathematics.* The Hague: Mouton.

Hori, J. (1979). *Entoropii towa nani ka [What is entropy?]* Tokyo: Koodansha.

Kageyama, T. (1982). Word formation in Japanese, *Lingua 57*, 215-258.

Kageyama, T. (1993). *Bunpoo to gokeesee [Grammar and word formation]* Tokyo: Hitsuji Shoboo.

Kageyama, T. (1999). *Keetairon to imi [Morphology and meaning]*. Tokyo: Kuroshio Shuppan.

Kaiho, H. (1989). Dai-1 koo: Joohoo o hakaru – entoropii, joohoo dentatsu-ryoo, joochoo-do [The first lecture: Measuring information – entropy, information trasnmitted quantity, redundancy]. H. Kaiho (Ed.), *Shinri/kyooiku deeta no kaisekihoo 10-koo – Ooyoo-hen [Ten lectures of psychological and educational data – Application]* (pp. 14-26), Tokyo: Fukumura Shuppan.

Namiki, T. (1982). The notion 'head of word' and core and periphery word formation, *Studies in English Linguistics 10*, 21-41.

Nomura, M. (1977). Zoogo-hoo [word-formation], *Iwanami kooza nihongo Vol. 9 - Goi to imi [Iwanami Japanese lecture series Vol. 9 - Vocabulary and meaning]*. Tokyo: Iwanami Shoten.

Nomura, M. (1978). Setsujisee-jion-goki no seekaku [The characteristics of the affix-natured phonological base], *Denshikeesanki-niyoru kokugo kenkyuu IX [The study of Japanese language by the computer], Kokuritsu kokugo kenkyuu-jo hookoku 61 [Report by the National Research Institute for the Japanese language]*, 103-138. Tokyo: Shuuee Shuppan.

Nishigauchi, T. (2004). Go no shikumi [The mechanism of words], *Kotoba-no kagaku hando-bukku [A handbook for the science of languages]* (pp. 1-36), Tokyo: Kenkyuusha.

Tamamura, F. (1985). Go no koosee to Zoogohoo [The structure of word and word-formation], *Goi-no kenkyuu to kyooiku (Ge) [The study and education for words (Vol. 2)] (Nihongokyooiku shidoosankoosho 12) [The reference book for Japanese language education 12]*. Tokyo: Kokuritsu kokugo kenkyuu-jo [National Insitute for Japanese Language].

[6] Japanese titles and journals are transliterated in the Hepburn Style of Romanization with the same vowel repeated for long vowels.

Tamaoka, K. & Makioka, S. (2004). New figures for a Web-accessible database of the 1,945 basic Japanese kanji, fourth edition. *Behavior Research Methods, Instruments & Computers 36(3)*, 548-558.

Tamaoka, K., Miyaoka, Y., & Lim, H. (2003). Entoropii to joochoo-do no tayoosei to kisokusei o arawasu kokoromi – kankokugo-kei nihongo gakushuusha no keigo hyoogen o rei ni [Indexes of entropy and redundancy for measuring variation and regularity of expressions: The example of polite expressions as used by Korean native speakers learning Japanese]. *Nihongo Kagaku [Japanese Linguistics] (National Institute for Japanese Language) 14*, 98-112.

Tamaoka, K. Lim, H. & Sakai, H. (2004). Entropy and redundancy of Japanese lexical and syntactic compound verbs. *Journal of Quantitative Linguistics, 11(3),* 233-250.

Williams, E. (1981). On the notions 'lexically related' and 'head of a word', *Linguistic Inquiry 12*, 245-274.

Wimmer, G., & Altmann, G. (1999). Review article – On vocabulary richness. *Journal of Quantitative Linguistics 6(1)*, 1-9.

Glottometrics 10, 2005, 55-72

Text genre and kanji frequency

Eric Long and Shōichi Yokoyama[1]
The National Institute for Japanese Language

Abstract: Various ways are explored in this study of using kanji frequency lists derived from multiple corpora to characterise kanji usage within the corpora. First we discuss the scope of, and issues in processing, four corpora derived from commercially available CD-ROMs: two encyclopedias, a database of newspaper articles, and a four-CD-ROM collection of the texts of mostly fictional paper back books. Next a summary of the kanji frequency data is given, and it is pointed out that the frequency distribution is noticeably different from a classic Zipf's law distribution. A comparison is made between the standard set of Jōyō kanji and high-frequency kanj in the corpora, and the degrees of similarity among the corpora are obtained with the Chi (χ^2) By Degrees of Freedom (CBDF) measure proposed by Kilgarriff (1997). Finally a simple method is tried and evaluated for identifying kanji that have a high frequency in a particular corpus compared to their cross-corpus frequency.

Keywords: corpus, kanji frequency, frequency distribution, chi-square measure of corpus similarity, characteristic kanji

1. Kanji surveys and frequency lists

The use of kanji, which express both sound and meaning, allows a great scope for variation in the Japanese writing system. A single word may be written either with kanji or in phonetic kana; the kanji themselves may be selected to express various semantic nuances; and a single basic kanji may appear in various alternate forms. Given the vast number of words that may be written in multiple forms, it might seem impossible to gain an overall picture of kanji usage. Nevertheless, simple techniques of analysis, including the examination of lists that record the frequency of kanji use in actual texts, can provide considerable useful information.

Kilgarriff (1997) demonstrates that simple frequency lists of English orthographic words provide a basis for statistical comparisons of text corpora that yield intuitively reasonable results. In this case "words" are defined quite simply as sequences of letters and numbers delimited by spaces or punctuation. While he acknowledges that with word frequency lists by far the greater part of the original information in the texts is lost, such as syntactic structures and the meaning selected of polysemous words, he points out that such lists have a great advantage in being very easy to make, so that results obtained by various researchers working at various times and places may be reliably compared. In contrast, any analysis involving preliminary grammatical parsing and interpretation of the data will depend to a great degree on the software, dictionaries and conceptual framework used.

[1] Address correspondence to: Eric Long, or Shōichi Yokoyama, Ph.D., The National Institute for Japanese Language, 3591-2 Midori-cho, Tachikawa-shi, Tokyo 190-8561, Japan. E-mail: kgd03011@nifty.com.

It is not immediately obvious what level of abstraction for Japanese texts would be the equivalent of orthographic words in English. Since Japanese texts make no consistent use of space characters, those are useless for parsing out linguistic units. Relying solely on punctuation marks would result in many units that are too long and too infrequent to allow any useful frequency-based analysis. Contiguous strings of kanji and kana may also easily be extracted and analyzed, but such a method results in too many units that are not contiguous with word boundaries.[2] Software, in particular the freely available ChaSen, is available to parse text into morphemes and tag them with parts of speech, but again the results would depend on the version of the software and dictionaries, as well as the extent of post-processing to clean up the misparsings.

Perhaps the method for Japanese text which is most similar to parsing for orthographic words in English is simply to count the frequencies of individual kanji and kana. This approach is, in fact, the one taken in the earliest application of computers to analyzing a large quantity of Japanese text, carried out at the National Institute for Japanese Language,[3] which was based on data randomly selected from the Asahi Shinbun [newspaper] in 1966 (NIJL 1976). Historically there has been a great demand for information on kanji usage, which has been used to establish and revise sets of standard kanji for use in printed materials (Tōyō kanji 1946 and Jōyō kanji 1981), to select kanji for inclusion in character encoding schemes, as well as to prioritize high-frequency for learners of Japanese as a foreign language.[4]

The parsing of Japanese text into individual characters does of course strip much of the essential information about their actual usage. Most importantly, greatly divergent usages of the same kanji are merged together, since most kanji used to write Japanese are read with different pronunciations depending on whether they are used to write native Japanese vocabulary (using *kun* readings) or Sino-Japanese vocabulary (using *on* readings). Further, simple lists of kanji frequency ignore the difference between kanji used in common vocabulary items and those in personal and place names, for which any semantic content to the kanji themselves is most likely irrelevant to the text itself.[5]

Even given these limitations, there is a significant utility in the compilation and analysis of kanji frequency lists. Previous surveys carried out at the National Institute for Japanese Language have focused on kanji used in newspapers and magazines. Some of these, such as NIJL (1976), were based on random sampling of the surface area of the printed pages, but the survey incorporating the greatest quantity of data was derived from a commercially available database of newspaper articles on CD-ROM. The present study seeks to move beyond the newspaper data by also examining and comparing data from electronic editions of Japanese-

[2] For one attempt at analysing kanji strings see Long & Yokoyama (1997).

[3] Kokuritsu Kokugo Kenkyūjo, formerly known in English as The National Language Research Institute.

[4] The newspaper data on which the present study is based has, for example, been incorporated into Halpern (1999), a kanji dictionary for learners.

[5] For example, the 70th most frequent kanji in our survey from the 1993 newspaper database was *kome/mai/bei* ‘rice’, but by far the most common usage therein is to refer to ‘America’, and a large portion of the remainder are used in various personal and place names. To avoid confusion, the word *kome* itself is frequently written in katakana.

language encyclopedias, and collections of paperback books (*bunkobon*—mostly works of fiction).

First we will give a brief description of the contents of the four data sources used, along with a summary of their sizes in terms of kanji types and tokens, and a comparison of the distribution properties of kanji frequency from the most to least frequent characters. Next we will consider the degrees of similarity observed among the corpora using the chi square similarity measurement proposed by Kilgarriff (1997). Finally we will consider a simple method for identifying kanji that seem to be characteristic of one or more of these corpora.

2. Data used

All four sources used for this study are commercially available CD-ROMs, all of which are computerized versions of printed material: the HIASK database of articles from the 1993 Asahi Shinbun [newspaper], the Shōgakukan *Sūpā Nipponika Hyakkajiten* [encyclopedia], the Heibonsha *Sekai Dai Hyakkajiten* [encyclopedia], and four CD-ROMs of pocket books published by Shinchōsha. The four sources will be referred to here as HIASK, SGK, Heibon and Shinchō, respectively. All of these CD-ROMs come provided with browsing software, but the first step in our processing of them into corpora for analysis was to transfer as much of the contents as practical to plain text files. The Shinchō data, in e-book format, was relatively easy to directly read off the CD-ROMs and convert into plain text, but the other three sources had to be copied or downloaded using the browsing software.

2.1. Character encoding

All of the data for the four sources had been for the most part stored on CD-ROM in the JIS 0208 encoding, but the original printed editions contain numerous kanji not available in this encoding (which are referred to in this context as 'external characters' or *gaiji*). Special processing was required to include these external characters in our corpora. For Heibon and Shinchō this was achieved by cataloguing characters included in the special fonts used for displaying these external characters, since the font information could be preserved during data conversion.

In the HIASK data external characters had been stripped out, either converted to kana marked with a Fullwidth Circumflex Accent[6] (for example the string 銭其シン外相＾ is the representation of the name and title of the then Foreign Minister of China, Qian Qichen 钱其琛, with シン *shin* being a Sino-Japanese reading for the external character 琛) or else simply replaced with the Geta Mark (for example Deng Xiaoping 邓小平, the retired leader of China was represented as 〓小平, which appeared in the printed newspaper as 鄧小平). For the survey reported in Yokoyama et al. (1998), all such external characters indicated in

[6] Character names used here are those specified in data from the Unicode Consortium: http://www.unicode.org/Public/UNIDATA/UnicodeData.txt.

HIASK were identified by examining the corresponding kanji in the 12-volume reduced-size print edition, and that data is included in this study as well.

In the SGK data, all external characters are automatically converted to the Geta Mark when copied and pasted, but they were identifiable as displayed in the browsing software. A comparison with the printed edition showed that there were some minor discrepancies between the electronic and printed editions in the exact variant form used for some kanji, but no attempt has been made to resolve or reflect the differences in the data used for this study.

2.2. Scope of the corpora

2.2.1. HIASK

The HIASK data includes the full text of newspaper articles from Asahi Shinbun in 1993, but there are many gaps in the coverage relative to the printed edition, which somewhat detracts from its value as a representative sample of written Japanese. It includes none of the advertising in the printed edition, and many features are omitted such as the serial novels, weather forecasts and listings of television programs, as well as the text of many articles for which copyright issues could not be resolved. Further, while the reduced-size print edition which we could consult reproduces the Tokyo edition, the CD-ROM version includes none of the special material for the Tokyo region, incorporating rather the Nagoya, Osaka and Western Japan editions.

In order to maximise the usefulness of the data, we compiled the data in two different fashions: One set simply comprises the raw data as given on the CD-ROM, and the other set includes all the external characters that we were able to identify, but excludes the regional material for which we were unable to identify them. For this smaller set of data we also checked the reduced-size print edition for pairs of kanji variants that were frequently switched for one another, and performed many other spot checks to get an overall picture of the differences between the electronic and printed versions. It turned out that the article headlines showed many differences so these were excluded from the data set as well, and we further attempted to eliminate as much as possible instances in which portions of longer articles were duplicated within the database as separate smaller articles. The data discussed in this study is this smaller data set, and is the same as that given in the frequency tables of Yokoyama et al. (1998) with some additional corrections.

One prominent characteristic of the HIASK data, both in the printed and electronic versions, is the tendency to use simplified forms for non-Jōyō kanji which are formed in a manner analogous to the simplifications adopted for the Jōyō kanji.[7] Many of these simplifications happen to be found in revised versions of JIS 0208 after the original 1978 standard, and these are of course reflected in the CD-ROM edition. Other simplified forms in

[7] These character shapes are often referred to as *Asahi moji* since they are so characteristic of the Asahi shimbun. For example the JIS character form 掴 is derived for 摑 *tsuka(mu)* ‘grasp’ by analogy with 国 which is the simplified Jōyō form for the traditional 國 *kuni* ‘country’. *Asahi moji* include many characters not encoded in JIS 0208, such as a simplified form of 瞼 *mabuta* ‘eyelid’ analogous to simplified 険 for 險 *kewa(shii)/ken* ‘steep/dangerous’.

the print edition, which are not found in JIS 0208, have been for the most part converted in the electronic edition into the closest corresponding JIS form. Sasahara et al. (2003) gives the results of our extended investigation into the use of these forms, but the data from this investigation has not been incorporated into the data used in the current study.

The newspaper print edition is for the most part in the vertical format, but the data as displayed in the electronic edition is in the horizontal format which predominates when Japanese is processed and displayed with computers. Some minor changes were apparently made to the text data to accomodate this change of format, but the data still retains many characteristics of the vertical format. For example a person's age may be expressed as 二十九歳 rather than as 29 歳, which would be the preferred orthography in horizontal writing.

2.2.2. SGK and Heibon

The two encyclopedias used for this study are similar in scale and scope, but show several noticeable differences. Most prominent is that while the printed version of SGK is composed in the vertical written format, Heibon is printed horizontally. Readings of non-Jōyō kanji, as well as non-standard readings of Jōyō kanji, are consistently displayed in the print version of SGK by inter-columnar *furigana*, while those in Heibon are given in parenthesised kana following the kanji words, and are provided much less consistently than in SGK. Numerical data in the print version of SGK are mostly expressed with numerical kanji, while Heibon makes heavy use of Western-style numerals. Nevertheless, SGK in the electronic edition has kanji readings and numerical data converted into a format essentially identical to that of Heibon, which leads to a significantly greater similarity between the two electronic editions than that found between the print editions.

The electronic editions of both SGK and Heibon include data in numerous charts which are difficult to access in a thorough fashion, so none of the chart data has been used in the present study.

One difficulty in using the Heibon electronic edition is that external characters are heavily used to display non-Jōyō kanji where the standard printed shape differs with the shapes given in the 1983 and 1990 versions of JIS 0208.[8] Since there is no relation between the codes in the external character font and JIS codes, these characters are replaced by an entirely different character if the data is copied and pasted into an editor that is incapable of retaining the font information. Thus a special font is used to display the standard printed form 餌 *esa* 'food (for animals)' rather than the current JIS form 餌, and if the font information is lost the character is corrupted to 品 *hin* 'goods'. SGK, on the other hand, only corrects the forms of a few characters that show a relatively major difference between the standard printed and JIS forms, such as 頰 versus 頬 for *hō* 'cheek', and these corrections are made at the original code points so even if the font information is lost an equivalent character is still displayed. We dealt with this issue by using Unicode encoding as the key for all frequency data. Thus in the case of 餌 both forms are covered by the same code point, and thus this particular

[8] See Lunde (1993:326-332) or Lunde (1999:919-925) for a useful table of these changes.

distinction is ignored, while for the two forms of 頰 there is a difference in Unicode and the distinction is reflected in the frequency data. This gives 頰 a frequency of zero in the HIASK data, and 頬 a frequency of zero in the other three data sources, even though the two forms are essentially identical in pronunciation, meaning and usage.

2.2.3. Shinchō

The data for the Shinchō pocket books comes from four separate CD-ROMs: *Shinchōsha hyakusatsu*, which includes 67 books originally written in Japanese, and 33 translations from other languages; *Meiji no bungō*, which includes 40 works by authors active during the Meiji period (1868-1912); *Taishō no bungō*, with 40 works by authors active during the Taishō period (1912-1926); and *Shinchōsha zeppan hyakusatsu*, which includes 68 native Japanese and 32 translated books which had become out of print.[9] The 135 native Japanese books in the two *hyakusatsu* collections represent the work of 117 authors, greatly vary in length, and date from a span of ninety years from 1895 to 1985, affording a wide variety of styles of writing. The 80 books in the *bungō* collections are the work of only 30 different authors, Natsume Soseki being the most heavily represented with 17 volumes. In addition, ten books in the original Shinchōsha *hyakusatsu* collection are duplicated in one or other of the *bungō* collections. For the purposes of this study, a collection of all 270 unique books was assembled, and three subcorpora were also prepared, one of all the translated works (65), one of the works dating from the Meiji and Taishō periods (96), and one of the works from the more recent Shōwa period (109). For the comprehensive corpus all parts of the texts were retained, while for the subcorpora any afterwords, timelines and explanatory notes were removed.

The browsing software for these Shinchōsha CD-ROMs displays the text in a vertical format very similar to the appearance of the actual printed books. In other ways as well, Shinchōsha is the most faithful representation of the actual text of the printed pocket books. Much of the data is, however, quite different from the original printed editions of the works. Any works that date from before the introduction of Tōyō kanji have been edited for modern editions, including the pocket book editions, to use the simplified kanji forms, and in many cases kanji orthography has been replaced by kana for words such as その ‘that’ (which had been commonly written with the kanji 其). The degree to which the texts have been brought into conformance with modern usage varies greatly from book to book.

3. Scale of the corpora

The size of the corpora are given in Table 1 below. They are not radically different in size, with the smallest, Shinchō, being slightly more than two thirds the size of the largest, Heibon. Table 2 shows how many kanji types are found in all four corpora, in three corpora, in two,

[9] A wealth of useful information about these CD-ROMs may be found at http://homepage1.nifty.com/mshibata/s100.htm.

and in only one of the corpora. While the encyclopedias show the greatest number of kanji types, Shinchō also shows a great variety of kanji, while HIASK, the newspaper corpus, shows by far the least number of types of any of the corpora. It turns out that nearly all Jōyō kanji are represented in all the corpora, with only , ‘imperial 1st person pronoun’ lacking in HIASK.

Table 1
Sizes of the corpora

corpus	types	Tokens
HIASK	4562	17,066,673
SGK	7103	23,572,690
Heibon	7420	25,467,739
Shinchō	6221	16,984,101
Total	9036	83,091,203

Table 2
Representation of types in multiple corpora

types in all four corpora	4022
types in three	1394
types in two	1416
types in one	2204

A breakdown for the Shinchō subcorpora is given in Table 3. It is apparent that the older works make the greatest contribution to the variety of kanji in the Shinchō corpus.

Table 3
Sizes of the Shinchō subcorpora

Shinchō subcorpus	Types	Tokens
Meiji/Taishō	5740	4,463,781
Shōwa	5042	7,768,539
translated	4198	3,921,356
total	6127	16,153,676

4. Frequency distributions

4.1. The relation between rank and frequency

There is, of course, an enormous difference between the frequencies of the most and least common kanji in a corpus. The highest raw frequency in any of the corpora used for this story is 271,047 for the kanji 年 'year' in SGK, and there are 1050 kanji which appear only a single time in the same corpus, 465 appearing twice, 268 appearing three times, and with fewer and fewer kanji for increasing frequencies.

This sort of frequency distribution calls to mind Zipf's Law, which is widely discussed, for example in Kornai (2002). To afford a rough test of the fit to a Zipf's Law distribution we give a graph of kanji frequency in the SGK corpus, plotted on a logarithmic scale against the rank when all kanji are sorted from high to low frequency. In a classic Zipf's Law distribution the line should be nearly straight with a slope of approximately -1. In fact the slope for high-frequency kanji is considerably gentler, and the drop-off for low-frequency kanji is quite precipitous. This suggests that while kanji form a set that tends to expand when larger and larger quantities of text are obtained, this set is considerably less open than a set of actual lexical items. When plotted in this fashion, the curves for the other three corpora are all similar to the one for SGK.[10]

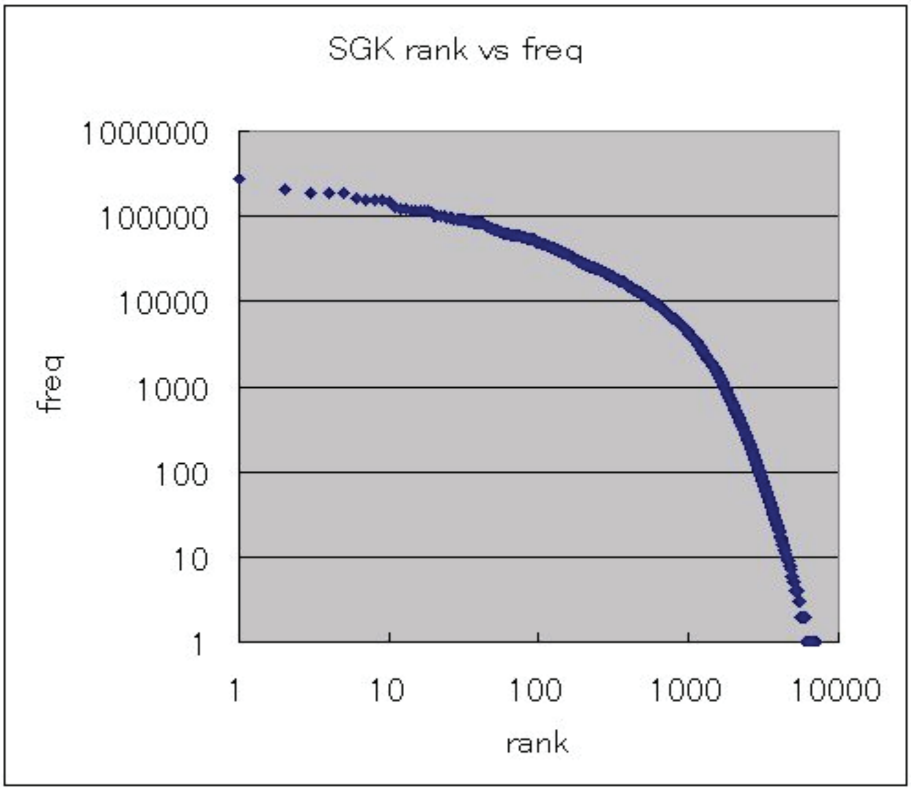


Figure 1. Frequency versus rank in SGK corpus

[10] An examination of the same data using the Altman Fitter 2.1 (information available at http://www.rst-gmbh.de/) also shows a poor fit with the Zipf's Law curve (Zeta distribution function), with a contingency coefficient C of .6010 (the software accepts values of less than .015 as an acceptable fit). The best fit obtained for SGK with the Fitter is with the Jain-Poisson function (C = .0032; parameters a = 1.75, b = .95), as well as for HIASK (C = .0072; a = 1.52, b = .94), while the Heibon data showed the best fit with a Mixed negative binomial (C = .0018). We were unable to obtain any fit for the Shinchō data.

One approach often taken to examining frequency distributions is to find the number of top-ranked types that will account for a given percentage of all tokens. Table 4 below shows the results for 50%, 90% and 99% coverage. HIASK shows the best coverage for a limited set of kanji, showing that half of all kanji in the newspaper articles can be read with a knowledge of just 151 kanji, and with 99% of all kanji accounted for by the top 1600, well under the size of the set of Jōyō kanji. The Meiji/Taisho Shinchō subcorpus requires by far the largest set of kanji types to cover 99% of all tokens.

Table 4

Number of kanji necessary to account for given proportions of total types

Coverage	HIASK	SGK	Heibon	Shinchō	Meiji/Taishō	Showa	Translated
50%	151	185	187	173	167	173	151
90%	762	964	956	1093	1099	1060	946
99%	1600	2188	2173	2517	2647	2382	2112

4.2. The relation between high-ranking kanji and Jōyō kanji

The figures in Table 4 are based on the top-ranked kanji for each individual corpus or sub-corpus, but it is also useful to get an idea of the coverage afforded by standard sets of kanji. The most convenient groups to use for this purpose are the set of Jōyō kanji, and the levels of Kyōiku kanji (Education kanji) which are earmarked to be taught over the six years of elementary school. The coverage for the main corpora in this study is shown in Table 5.

Table 5

Coverage by Kyōiku and Jōyō kanji (size of set shown in parentheses)

kanji set	HIASK	SGK	Heibon	Shinchō
1st year (80)	0.2126	0.1557	0.1477	0.2101
2nd year (160)	0.2178	0.2145	0.2125	0.2303
3rd year (200)	0.1773	0.1864	0.1849	0.1593
4th year (200)	0.1335	0.1421	0.1461	0.0954
5th year (185)	0.1051	0.1099	0.1146	0.0631
6th year (181)	0.0563	0.0594	0.0632	0.0525
total Kyōiku (1006)	0.9025	0.8681	0.8691	0.8106
other Jōyō (939)	0.0851	0.1057	0.1058	0.1337
total Jōyō (1945)	0.9876	0.9738	0.9750	0.9444
non-Jōyō (7091)	0.0124	0.0262	0.0250	0.0556

Again, HIASK shows the best coverage for limited sets of kanji, with the Kyōiku kanji accounting for over 90% and Jōyō kanji for nearly 99%.

Actually, since 99% of tokens in HIASK can be accounted for by a set of kanji 345 fewer in number than the set of Jōyō kanji, it is clear that there must be a substantial number of low-frequency Jōyō kanji, and and there are also many non-Jōyō kanji of relatively high frequency. To get some idea of the nature of this effect, first let us examine the ranks and frequencies of some of the higher-ranked non-Jōyō kanji and lower-ranked Jōyō kanji. From here on out we will use frequencies normalised to a corpus size of one million kanji, which will make it easier to keep a sense of the relative proportion of each kanji. First let us note that the kanji of rank 1945 in HIASK (which would be the least frequent Jōyō kanji if there were a perfect coincidence between Jōyō kanji and the 1945 top-ranked kanji) has a normalised frequency of 10.61 .

In total there are 157 non-Jōyō kanji among the most frequent 1945 in the HIASK corpus. Of these 90 were in the list of Jinmei-yō kanji (the set of additional kanji approved for use in given names) that was in force in 1993. Table 6 shows the ten highest-ranked non-Jōyō kanji, with the Jinmei-yō kanji indicated by an M, and the 493 additional kanji that were newly added to Jinmei-yō kanji in 2004 marked with an N.

Table 6
High-ranked non-Jōyō kanji in HIASK

Kanji	Rank	Normalised frequency	Typical uses
藤 M	253	1059.84	*fuji* ‘wisteria’; 佐藤 Satō (common surname)
岡 N	326	791.95	*oka* ‘hillock’; 岡田 Okada (common surname)
阪 N	444	557.23	大阪 Osaka
韓 N	547	435.82	韓国 Kankoku (South Korea)
伊 M	687	303.52	伊藤 Itō (common surname); 伊勢市 Ise-shi (city)
鹿 M	732	269.12	鹿児島県 Kagoshima-ken (prefecture)
奈 M	734	268.77	奈良市 Nara-shi (city)
狙	806	223.18	*Nera(u)* ‘aim for’; 狙撃する *sogeki suru* ‘snipe’
彦 M	854	204.20	龍彦 Tatsuhiko (boy’s given name)
之 M	908	180.06	雅之 Masayuki (boy’s given name)

Among the top 1945 HIASK kanji there are a number of kanji that are traditional forms of simplified Jōyō kanji, variant forms of Jōyō kanji, and a number of kanji that are rarely used in Japan but are necessary to write Korean and Chinese names. A selection of these are shown in Table 7.

Table 7
A selection of traditional and variant forms, and relatively obscure kanji in HIASK top 1945

Kanji	Rank	Normalised frequency	Typical uses
澤	1524	21.06	traditional for 沢 *sawa* 'mountain stream'; common in surnames
嶋 N	1125	105.12	variant of 島 *shima* 'island'; 長嶋茂雄 Nagashima Shigeo, then manager of Yomiuri Giants baseball team
舘	1923	11.60	variant of 館 *kan* 'public building'; 国士舘 Kokushikan, a university in Tokyo
鄧	1668	24.96	Deng Xiaoping, the retired leader of China (external character)
盧	1864	14.53	盧泰愚 Roh Tae-woo, president of South Korea until 1993
閧	1934	11.13	北勝閧 Kitakachidoki, sumo wrestler

In contrast, the lowest-ranked Jōyō kanji have extremely low frequencies, with the bottom ten ranging from zero to .41 (corresponding to a raw frequency of 7). Two of these can actually be used to write common words, but other kanji of similar meaning are usually used instead: 膨脹 => 膨張 *bōchō* 'swelling/ expansion'; and 遵守 => 順守 *junshu* 'compliance'. In fact, as noted in Takebe (1993) many of these low-frequency Jōyō kanji were never actually intended for ordinary use, but were included, rather, to have the complete set of kanji used in the Constitution of Japan, which was promulgated at nearly the same time as the original Tōyō kanji. Some of these low-frequency Jōyō kanji were also kanji judged by the Japan Newspaper Publishers and Editors Association[11] to be superfluous and not to be ordinarily used.[12]

Table 8
Low-ranked Jōyō kanji in HIASK

Kanji	Rank	Normalised frequency
虞 璽	3230	.41
遵 弐	3317	.35
勺 且	3752	.18
匁 斤 脹	4014	.12
朕	—	.00

The highest-ranked non-Jōyō kanji in the encyclopedias are for the most part the same as those in the newspaper articles. The numbers of non-Jōyō kanji in the top 1945 are 214 in SGK and 210 in Heibon, compared to 157 in HIASK, although the highest-ranked ones have

[11] Nihon Shinbun Kyōkai.

[12] In contrast, 狙 in Table 6 is from a small group of kanji judged by the Association to be common and useful enough to be used freely as if they were actually Jōyō kanji.

a much lower rank and frequency than in HIASK. For example 藤, although it is also the top-ranked non-Jōyō kanji in both the encyclopedias, is ranked 423 in SGK and 464 in Heibon, much lower than the 253 in HIASK. The most noticeable differences among the higher-ranked non-Jōyō kanji in the encyclopedias are shown in Table 9; the ones more highly ranked in the encyclopedias are particularly important in discussions of history and religion: 鎌 *kama* 'sickle/scythe' is used in 鎌倉 Kamakura, the political and military power center which gave its name to the Kamakura era of Japanese history; 宋 *sō*, for the Song dynasty (960-1279) of China; 阿 *a* is mostly used phonetically, most commonly in the encyclopedias in 阿弥陀 Amitābha, the Buddha of Limitless Light.

Table 9
High-ranking non-Jōyō kanji showing a large difference between newspapers and encyclopedias (rank among non-Jōyō kanji shown in parentheses)

kanji	HIASK freq	HIASK Rank	SGK Freq	SGK rank	Heibon freq	Heibon rank
韓 N	435.82	547 (4)	96.64	1316 (30)	92.51	1332 (27)
狙	223.18	806 (8)	3.86	3025 (1096)	8.17	2576 (681)
阿 M	82.38	1228 (26)	227.34	880 (5)	222.16	895 (6)
鎌 M	31.00	1590 (70)	214.32	916 (7)	264.73	798 (4)
宋 N	4.69	2159 (282)	119.71	1214 (18)	178.26	996 (9)

The extreme difference in the frequencies of 狙 *nera(u)* 'aim for' *nera(i)* 'purpose' seems mainly to be due to the fact that newspapers have no end of occasions on which to refer to the intentions and goals of politicians, corporations and sports figures, which is no doubt why their trade association decided to give it the same treatment as a Jōyō kanji. This kanji was one of a group of 88 kanji that were considered for adoption use in personal names as new Jinmei-yō kanji in 2004 but were rejected as unsuitable. Presumably the grounds for rejection were the unpleasant associations of words such as 狙撃者 *sogekisha* 'sniper', but ironically the vast majority of the kanji's uses carry no such connotation.

The Shinchō corpus includes several high-ranking non-Jōyō kanji quite different from those we have yet seen. Most of these are meant to be written out in kana when strictly complying with Jōyō kanji usage (云 誰 頃 俺 此), or have been replaced by a Jōyō kanji of similar meaning (坐 => 座; 廻 => 回). The bulk of the uses of 吾 are as an element of given names of characters; over two-fifths of the occurrences are mentions of 新吾 Shingō the eponymous character of a novel in Z*eppan hyakusatsu*.

Table 10
High-ranking non-Jōyō kanji in Shinchō

Kanji	Rank	Normalised frequency	Common usage
云 N	43	3216.01	*i(u)* 'say' (used after quotations)
誰 N	180	1118.87	*dare* 'who'
藤 M	233	915.62	

頃 N	234	913.62	*koro* ‘the time when …’
俺 N	438	517.72	*ore* (1st person singular pronoun)
之 M	447	512.01	
吾 M	505	453.84	*ware* (1st person singular pronoun)
坐 N	509	450.01	suwa(ru) ‘sit’
廻 N	514	446.18	mawa(ru) ‘turn’
此 N	534	429.75	*kore* ‘this’

5. Corpus similarity and corpus homogeneity

In the discussion above on high-ranking non-Jōyō kanji, we have seen particularly close similarities in the usage between the two encyclopedias, a general similarity between the encyclopedias and the newspapers, and a large difference between these three corpora and the pocket books. In this section let us see if these observations are backed up by a more general measurement of corpus similarity.

Kilgarriff (1997) makes a strong case for the utility of a measurement based on a χ^2 score derived from a table of the frequencies of the most common orthographical words in a pair of equally-sized samples of two corpora. This measurement is most commonly used to test for a significant statistical difference between sets of data, but as Kilgarriff points out, any large sets of frequency data derived from naturally occurring texts will inevitably show a significant difference. Kilgarriff demonstrates, however, that the χ^2 score also shows high values for greatly differing corpora, and relatively low values for similar corpora. In the method he describes, samples were taken from sub-corpora of the British National Corpus, selecting the first 200,000 words from each of 33 text sources. He prepared contingency tables of word frequencies for each pair of samples, taking the most frequent words in the union of the two samples, and calculated the value of χ^2 divided by the degree of freedom (CBDF: Chi By Degrees of Freedom; in this case the degree of freedom is one less than the number of words selected). Also by comparing χ^2 for subsets of the same corpus, a measure of its homogeneity may be obtained, that is, of the similarity of each part of the corpus to the other parts.

The method we adopt here is similar, only instead of using a sample of a fixed number of words from each corpus, we used the frequencies from the entire corpora normalised to a corpus size of one million kanji. We used the 1006 most frequent kanji in each pair of normalised corpora, the number of Kyōiku kanji. From these observed values a table was constructed of the frequencies that would be expected if both corpora had identical kanji distributions; the expected value of kanji *x* in corpus *a* would be:

$$\frac{\text{total number of } x \quad \times \quad \text{total kanji in } a}{\text{total kanji in table}}$$

Then the degree to which the observed value O differs from the expected value E is calculated as $(\mathrm{O} - \mathrm{E})^2 / \mathrm{E}$, and this value is summed over every cell in the table. Finally the sum is divided by the degrees of freedom for the table, which in this case is 1005. The results for comparison of the corpora are shown in Tables 11 to 13; the lower a score is the more similar the two corpora should be.

As expected, the results show that the two encyclopedias are highly similar, the newspaper article data is more distant from them, but still much closer than the paper back book data. The three Shinchō subcorpora all show great dissimilarity with the other three corpora, but while the more recent group (Shōwa) of Japanese books shows nearly the same degrees of similarity as Shinchō as a whole, the translated books are noticeably more dissimilar, and the older group of Japanese books (Meiji/Taishō) is even more so. Within the Shinchō subcorpora, however, the translated books stand out as more dissimilar with the other two subcorpora, expecially with Meiji/Taishō.

Table 11
CBDF similarity scores for the four corpora

	Heibon	HIASK	Shinchō
SGK	8.09	191.39	328.50
Heibon		198.18	326.98
HIASK			308.05

Table 12
CBDF similarity scores of three corpora with the Shinchō subcorpora

	Meiji/Taishō	Shōwa	Translated
SGK	393.81	324.97	362.52
Heibon	391.67	324.74	354.65
HIASK	390.42	295.13	348.08

Table 13
CBDF similarity scores of the three Shinchō subcorpora

	Shōwa	translated
Meiji/Taishō	65.85	133.14
Shōwa		93.32

Finally, to give some idea of the homogeneity of the corpora, we split the HIASK, SGK and Heibon corpora into chunks of 1, 2, 4 and 8 million kanji, as well as into half, according to the order in which the data was originally retrieved from the CD-ROMs (in the order of date for the newspapers, and by the JIS code order of first kana for the encyclopedia articles; we have not yet attempted this operation for the Shinchō corpus), and compared like-sized chunks. In each case we again used frequency values normalised to a total of one million kanji for each chunk. The greatest number of comparisons was for Heibon, which yielded 25 chunks of 1,000,000 kanji, resulting in 300 comparisons. The results are given in Table 14, listing the mean values for all comparisons where applicable.

All three corpora show considerable heterogeneity at a chunk size of 1,000,000 kanji, which gradually becomes smoothed out as the size increases, although HIASK is noticeably more homogenous than the others at the smaller chunk sizes. Notice that the heterogeneity observable in SGK and Heibon at chunk sizes up to 4,000,000 is greater than the dissimilarity observed between the total corpora, although the half-corpora show more similarity within one corpus than the whole corpora show with each other.

Table 14

Homogeneity values for three corpora: mean and standard deviation

chunk size	HIASK	SGK	Heibon
1,000,000	19.38 / 6.95	36.42 / 9.54	38.66 / 10.70
2,000,000	14.54 / 5.12	21.96 / 6.14	24.86 / 7.70
4,000,000	10.24 / 2.99	13.91 / 4.59	14.21 / 4.90
8,000,000	5.63	6.13	6.84 / 1.99
half total size	4.91	3.68	3.71

6. Characteristic kanji of the corpora

There are many approaches to determining the characteristic vocabulary of a text or corpus, of which Kilgarriff (1996) provides a useful survey. For our purposes, we would hope to identify information on characteristic kanji that would be useful for learners of Japanese in prioritising kanji for study purposes, based on the type of texts they would be most interested in reading. In this case three interrelated criteria suggest themselves: high frequency in the target corpus, higher-than-expected frequency in the target corpus, and relatively low frequency in general use. A kanji that has a strong tendency to appear in the target corpus, when it appears at all, but is still at a low frequency in the target corpus, would not necessarily be a good candidate for initial prioritisation. On the other hand, a high-frequency kanji that appears more often than expected in the target corpus might still have a high enough frequency in general usage that it would require no special prioritisation.

We are aware of no single metric that would sufficiently reflect all three criteria mentioned. One common approach is to use the value of χ^2 obtained from a two-by-two grid for kanji x:

freq of x in target corpus	freq of x in other corpora
total freq of all other kanji in target corpus	total freq of other kanji in other corpora

In this case the expected value is based on a table listing the frequencies for all four corpora. One can give a negative sign to the value whenever the observed value O was less than the expected value E, and then sort the whole list of kanji for the values obtained for each corpus. The trouble with this method is that high-frequency kanji are so favored that much of the result consists of kanji that are common in all the corpora. For example, for the top 21 kanji that sort out in this way from HIASK, 11 of them are numerical kanji, although others in this group, such as the kanji necessary to write 議会 *gikai* ‘parliament’ and 政党 *seitō* ‘political party’ are more obviously useful for reading newspaper articles in particular.

Another way to approach the criteria mentioned above is to set a cut-off point of the top N kanji in any particular corpus, and sort these by the ratio of their observed to expected values: O / E. This method has the advantage of heavily favoring kanji that have low frequencies in all but the target corpus. For this study we set the cut-off point at the 1006 top-ranked kanji for each corpus, and selected the 30 kanji that show the greatest ratio of observed to expected frequency (in this case the expected frequency is simply the mean of the normalised frequencies), and give the results in Table 15. With each kanji is listed a typical meaning, or the meaning of a compound typically written with that kanji, to give some idea of the fields in which they are used.

Once again it is seen that the two encyclopedias are quite similar to one another, though Heibon seems to favor slightly terms relating to history, while SGK favors terms used in describing plants and animals (SGK has many more articles describing individual species; it is in these articles that most of the uses of 褐色 *kasshoku* ‘brown’ appear. Also SGK seems to have an editorial policy of reproducing dates in Japanese history in both Western and Japanese reign-name-based dates, which results in the relatively heavy use of 昭和 Shōwa. Heibon shows a much higher frequency of 呼 *yo(bu)* ‘call’ because SGK seems to follow a policy of writing that verb in kana in the phrase …とよばれる ‘is called …’.

HIASK shows relative high frequencies of kanji used in writing numbers, which is due to the retention of such kanji mentioned in section 1.2.1. There also appear many kanji related to politics and administration. The high use of 塁 *rui* ‘base’ and 秒 ‘second’ is mostly due to sports articles. Shinchō, as we’ve seen before, has characteristic kanji that are now written in kana when using standard orthography, such as 貰 *mora(u)* ‘receive’, 筈 *hazu (da)* ‘is expected/supposed to …’, 或 *a(ru)* ‘a certain …’, 又 *mata* ‘again’, and others that we have already seen. The kanji 這 *ha(u)* was used in an old orthography for *hai(ru)* ‘enter’: 這入る instead of the now-standard 入る.

The fact that so many kanji associated with obsolete orthographies appear in this list shows that the method used here is liable to be influenced by highly skewed usage in a subset of a corpus, resulting in many ‘characteristic’ kanji that are actually *un*characteristic of large portions of a corpus. This problem is probably especially severe for the Shinchō corpus due to its extreme heterogeneity. It can be mitigated by first selecting from higher-frequency kanji, say the top 500 ranked, or by selecting from kanji with high frequency overall rather than high frequency in just one corpus. Another possibility would be to use some method of adjusting frequencies, as briefly mentioned in Kilgarriff (1996), that would favor kanji that are distributed evenly throughout the target corpus.

Table 15
Most characteristic kanji for each corpus selected from the top 1006 in each corpus, and listed by ratio of observed to expected frequency.

HIASK	SGK	Heibon	Shinchō
塁 ‘(first) base’	昭 Shōwa	宋 Song dynasty	貰 ‘receive’
票 ‘ballot/vote’	褐 ‘brown’	荘 ‘estate’	云 ‘say’
秒‘second (of time)’	雌‘female (animal)’	呼 ‘call’	俺 ‘I’
狙 ‘seek/snipe’	菌 ‘fungus/bacteria’	紀 ‘century’	訊 ‘ask/visit’
億 ‘100 million’	殻 ‘shell’	磁‘magnet/porcelain’	筈 ‘supposed to …’
午 ‘afternoon/p.m.’	茎 ‘plant stem’	暦 ‘calendar’	誰 ‘who?’
昨 ‘yesterday’	酸 ‘acid/oxygen’	栽 ‘cultivate/grow’	這 ‘crawl/enter’
党 ‘political party’	網 ‘(fishing) net’	唐 Tang dynasty	嘘 ‘lie/falsehood’
九 ‘nine’	脈 ‘vein/artery’	培 ‘cultivate/grow’	或 ‘a certain …’
埼 Saitama pref.	貝 ‘shellfish’	漢 ‘kanji/Han dyn’	僕 ‘I’
募 ‘recruit’	鉱 ‘mine/mineral’	銅 ‘copper’	噂 ‘rumor/story’
韓 South Korea	卵 ‘egg’	典 ‘classic/typical’	頃 ‘time when …’
百 hundred	膜 ‘membrane’	鎌 Kamakura	厭 ‘disliking’
選 ‘choose/election’	殖 breed/reproduce’	概‘concept/summary’	瓢 ‘gourd’
七 ‘seven’	培 ‘cultivate/grow’	晶 ‘crystal’	嬢 ‘daughter’
兆 ‘trillion’	栽 ‘cultivate/grow’	帝 ‘emperor’	此 ‘this’
委 ‘committee’	亜 ‘sub-‘	胞 ‘cell/spore’	燈 ‘lamp/light’
円 ‘yen’	径 ‘diameter’	徴 ‘characteristic’	逢 ‘chance to meet’
米 ‘rice/America’	蒸 ‘vapor/distill’	称 ‘call/name’	頬 ‘cheek’
副 ‘vice-‘	科 ‘silence’	溶 ‘dissolve’	且 ‘once … …’
挙 ‘election’	溶 ‘dissolve’	王 ‘king’	附 ‘be attached’
六 ‘six’	胞 ‘cell/spore’	項 ‘item’	坐 ‘sit’
捜 ‘search (warrant)’	糖 ‘sugar’	繊 ‘fiber’	眺 ‘gaze at’
十 ‘ten’	称 ‘call/name’	礎 ‘foundation’	又 ‘again’
援 ‘support’	湖 ‘lake’	膜 ‘membrane’	椅 ‘chair’
千 ‘thousand’	磁 ‘magnetism’	系 ‘lineage/system’	泣 ‘weep’
協 ‘cooperation’	属 ‘belong/pertain’	酸 ‘acid/oxygen’	其 ‘that’
庁 ‘gov’t office’	晶 ‘crystal’	彫 ‘sculpture’	笑 ‘laugh’
撤 ‘thorough’	液 ‘liquid’	雌 ‘female (animal)’	闇 ‘darkness’
税 ‘tax’	岳 ‘peak/Mt.’	液 ‘liquid’	婆 ‘old woman’

References

Halpern, Jack (1999). *The Kodansha Kanji Learner's Dictionary*. Tokyo, New York and London: Kodansha International.

Kilgarriff, A. (1996). Which words are particularly characteristic of a text? A survey of statistical approaches. *Proceedings AISB Workshop on Language Engineering for Document Analysis and Recognition*. Beijing and Hong Kong.

Kilgarriff, A. (1997). Using word frequency lists to measure corpus homogeneity and similarity between corpora. *Proceedings 5th ACL workshop on very large corpora*. Beijing and Hong Kong.

Kornai, András (2002). How many words are there? *Glottometrics 4, 61-86.*

Long, Eric & Yokoyama Shōichi (1997). An analysis of kanji strings in the CD-HIASK '93 Data Base. *Jimbun kagaku ni okeru sūryōteki bunseki (2) 15-20.* Tokyo: Tōkei sūri kenkyūjo.

Lunde, Ken (1993). *Understanding Japanese Information Processing*. Sebastopol, California: O'Reilly & Associates.

Lunde, Ken (1999). *CJKV Information Processing*. Sebastopol, California: O'Reilly & Associates.

National Institute for Japanese Language (1976). *A study of the use of Chinese characters in modern newspapers* (Report 56). Tokyo: Sanseidō.

Sasahara, H., Yokoyama, S., & Long, E. (2003). *Gendai Nihon no itaiji: kanji kankyōgaku josetsu* [Linguistic ecology of kanji variants in contemporary Japan: a preliminary study]. Tokyo: Sanseidō.

Takebe, Yoshiaki (1993). *Naru hodo Jōyō kanji* [Introduction to Jōyō kanji]. Tokyo: Nippon hyōronsha.

Yokoyama, S., Sasahara, H., Nozaki, H., & Long, E. (1998). *Shinbun denshi media no kanji* [A study of kanji in electronic newspaper media]. Tokyo: Sanseidō.

Data sources

Hitati Digital & Heibonsha (1998). *Sekai dai hyakka jiten* [The world encyclopedia]. Heibonsha. Tokyo.

Nichigai Associates & Asahi Shinbun (1994) *CD-HIASK '93: Asahi Shinbun kiji dētabēsu* [Asahi newspaper article database]. Kinokuniya. Tokyo.

Shinchōsha (1995). *CD-ROM ban Shinchōsha bunko no hyakusatsu* [CD-ROM edition Shinchōsha pocket books: 100 volumes]. Shinchōsha. Tokyo.

Shinchōsha (1997). *CD-ROM ban Shinchōsha bunko Meiji no bungō* [CD-ROM edition Shinchōsha pocket books: Meiji period literary masters]. Shinchōsha. Tokyo.

Shinchōsha (1997). *CD-ROM ban Shinchōsha bunko Taishō no bungō* [CD-ROM edition Shinchōsha pocket books: Taishō period literary masters]. Shinchōsha. Tokyo.

Shinchōsha (2000). *CD-ROM ban Shinchōsha bunko no zeppan hyakusatsu* [CD-ROM edition Shinchōsha pocket books: 100 volumes out of print]. Shinchōsha. Tokyo.

Shōgakukan (1998-2000). *Sūpā Nipponika Hyakkajiten* [Super Nipponika Encyclopedia]. Shōgakukan. Tokyo.

Glottometrics 10, 2005, 73-81

Predicting Attachment of the Light Verb *–suru* to Japanese Two-kanji Compound Words Using Four Aspects [1]

Katsuo Tamaoka [2]
Hiroshima University, Higashihiroshima, Japan
Chizuko Matsuoka [3]
Keimyung University, Daegu, *Korea*
Hiromu Sakai [4]
Hiroshima University, Higashihiroshima, Japan
Shogo Makioka [5]
Osaka Prefecture University, Sakai, Japan

Abstract. In the Japanese language, the light verb *–suru* can be attached to various two-kanji compound words containing a *verb-like* feature (or aspects) to allow them to be used as a verb. Using a large sample of the 2,000 two-kanji compound words, encompassing a little less than 80 percent of the total two-kanji compound words printed in 14 years of *Asahi Newspaper* issues, the present study investigates how much the light verb attachment is predicted by four aspects: *inchoative*, *durative*, *telic* and *stative*. A binary logistic regression analysis indicates that all four aspects are significant predictors. Among them, the *telic* aspect shows an overwhelmingly high predictive power. The quantitative theory type III analysis further demonstrates that, in contrast to the *stative* aspect, the *inchoative, durative* and *telic* aspects share a similar semantic feature of *time series*. Nevertheless, since the *telic* aspect overlaps not only the *time series* feature of the *inchoative* and *durative* aspects, but also the *stative* aspect, it is the most effective single predictor for light verb attachment, showing an extremely high prediction percentage of 93.64 with 1.05 percent error.

Keywords: light verb -suru, aspect, verb-likeness, two kanji-compound words, binary logistic regression analysis, Hayashi's Quantitative Theory Type III

1. Purpose and approach to the light verb attachment

In the Japanese language, many nouns can be used as a verb by simply adding the light verb –*suru* [6] as in 発表する 'to announce' created by a noun 発表 meaning 'announcement' plus

[1] The present research was supported by a grant-in-aid for scientific research (Category B) from the Japan Society for the Promotion of Science (#17320064).

[2] Address correspondence to: Katsuo Tamaoka, International Student Center, Hiroshima University 1-1, 1-chome, Higashihiroshima, Japan 739-8524. E-mail: ktamaoka@hiroshima-u.ac.jp.

[3] Chizuko Matsuoka, College of International Studies, Department of Japanese Studies, Keimyung University, 1000 Singdang-Dong, Dalseo-Gu, Daegu, 704-701, Korea. E-mail: chizuko@kmu.ac.kr.

[4] Hiromu Sakai, Graduate School of Education, Hiroshima University 1-1, 1-chome, Higashihiroshima, Japan 739-8524. E-mail: hsakai@hiroshima-u.ac.jp.

[5] Shogo Makioka, School of Humanities & Social Sciences, Osaka Prefecture University 1-1, Gakuen-cho, Sakai, Osaka 599-8531. E-mail:makioka@hs.osakafu-u.ac.jp.

[6] In the present study, using the theoretical framework of generative grammar, the term *light verb* is used for the

the light verb *–suru*. Because the light verb itself does not have a specific meaning, the meaning of the noun determines the meaning of the compounded verb, as in the verb 'to announce' from the noun 'announcement'. However, the light verb cannot be attached to all Japanese nouns. It is assumed that some specific meanings possessed by nouns must determine whether or not the light verb can be attached to them. In this regard, Iida (1987) and Ito and Sugioka (2002) propose that nouns can be classified according to their *verb-like* features. Furthermore, Grimshaw (1990) pointed out that the degree of *verb-liken* features can be measured in terms of the aspectual properties of nouns. Thus, the present study investigates the aspectual properties of nouns which decide the light verb attachment.

Linguistic testing using suffix attachment was applied nouns to determine which aspects related to *time* the nouns possessed (Iida, 1987; Shibatani & Kageyama, 1988). For example, a suffix 中 meaning 'during' can be added to the end of a word 建設 meaning 'construction', to produce 建設中 'during construction'. Since this suffix-based derivation is possible, this word is judged to have the *durative* aspect. Likewise, a suffix 後 meaning 'after' can be added to the same word to form 建設後 'after construction', implying that this word contains the *telic* aspect. Using this type of linguistic testing, it is possible to judge whether nouns contain *verb-like* features. In addition to the *durative* and *telic* aspects, Kageyama (1996) proposes two more aspects, *inchoative* and *stative*. Matsuoka (2004, 2005) elaborates on these two additional aspects: *inchoative* (e.g., 販売 'sale', 輸入 'importation') and *stative* (存在 'existence', 占領 'occupation'). The present study approaches this investigation by asking to what degree the light-verb attachment to two-kanji compound words[7] can be predicted by these four aspects: *inchoative*, *durative*, *telic* and *stative*.

2. Selection of two-kanji compound words for investigation of light verb attachment

Using the *Asahi Newspaper* printed from 1985 to 1998, Amano and Kondo (2000) created a large lexical corpus (or database) of word printed-frequencies in the seventh volume of their series. This corpus contains 341,771 words for type frequency and 287,792,797 words for token frequency. In this corpus, the light verb *–suru* is recorded separately from two-kanji compound words. For example, 散歩する meaning 'take a walk' is composed of two parts. A two-kanji compound word such as 散歩 meaning 'a walk' cannot be a verb by itself. However, with the attachment of the light verb *–suru* (written in hiragana as する), this word can function as a verb in a sentence like 公園を散歩した meaning '[I] took a walk in the park' (the subject 'I' is an empty subject). In Amano and Kondo's corpus, the printed-frequency of 散歩 is 2,537. The light verb is classified as a *verb ending* (動詞語尾) which naturally appears very frequently, 1,374,420 times, including every word that incorparates a light verb. Thus, all two-kanji compound words (selected by taking the category of 'general nouns' and

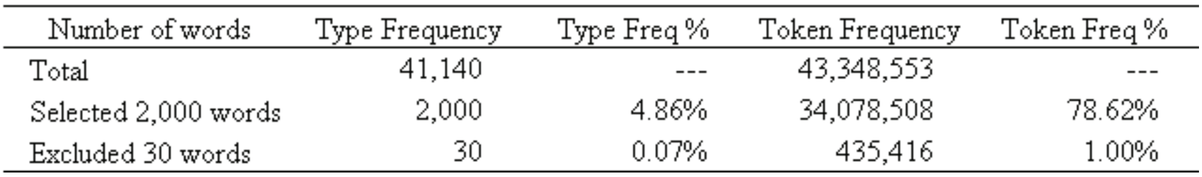
–suru attachment to a noun, contrasting with other verbs called *heavy verbs*.

[7] Although these aspects are often found among Sino-Japanese two-kanji compound words (e.g.運転する meaning 'to drive', derived from a noun 'drive' plus the light verb *–suru*), it extends to originally Japanese words (such as おしゃべりする meaning 'to chat', derived from a noun 'chatting' plus *–suru*) and even alphabetic-loan words (such as デザインする meaning 'to design', derived from a noun 'design' plus *–suru*) (Matsuoka, 2004, 2005). However, the present study only focuses on the most commonly suffixed two-kanji compound words.

excluding the category of ‘proper nouns’) were easily chosen from the corpus, for a total of 41,140 types and 43,348,553 tokens.

To have a manageable sample of two-kanji compound words, 2,000 items were selected from the top-ranking items in token frequency. However, 30 items incoroproating kanji numerals, such as 第一 meaning ‘the number one’, 十月 meaning ‘October’, were excluded from the present study. For this reason, 30 items were added from a ranking of the 2001st to 2030th in order to obtain 2,000 two-kanji compound words. As shown in Table 1, the 30 excluded items represent about 0.07% of the total type frequency of two-kanji compound words and 1.00% (435,416 instances) of the total token frequency. As a result of this sampling procedure, the desired number of 2,000 compound words was reached. These compound words represent only 4.86% of the total 41,140 two-kanji compound general noun types found in 14 years of the *Asahi Newspaper*. However, the same 2,000 compounds represent 78.62% (frequency 34,078,508) of a total of 43,348,553 tokens, as shown in Table 1. Since the 2,000 compounds encompass a little less than 80 percent of the total two-kanji compound words printed in the *Asahi Newspaper*, this sample of 2,000 compounds meets the requirements of meaningful investigation into light verb attachment.

Table 1
Percentages of selected 2,000 two-kanji compound words and 30 excluded words
in proportion to total type and token frequencies



Number of words	Type Frequency	Type Freq %	Token Frequency	Token Freq %
Total	41,140	---	43,348,553	---
Selected 2,000 words	2,000	4.86%	34,078,508	78.62%
Excluded 30 words	30	0.07%	435,416	1.00%

3. Coding of aspects (X_1 to X_4) and light verb attachment (Y)

Each two-kanji compound word was coded as 1 if it is possible to attach a light verb to it, and coded as 0 if this is not possible. For example, a light verb is not attached to a compound word 社会 meaning ‘society’, so this word was coded as 0. In other words, a verb 社会する, a compound of ‘society’ plus *–suru*, does not exist in the Japanese language. Another compound word 計画 meaning ‘a plan’ can have a light verb attached (a compound of ‘a plan’ plus *–suru* and becomes ‘to plan’), so this word was coded as 1. In this coding procedure, 802 out of the selected 2,000 two-kanji compound words (40.10%) were coded as 1.

The same process was used to code the existence of each of the four aspects in the 2,000 compounds. For instance, a compound word 交渉 meaning ‘negotiation’, to which the light verb can be added to form 交渉する meaning ‘to negotiate’, contains the *inchoative* aspect in its meaning. It can be easily understood that a combination of two compound words 交渉開始 (‘negotiation’ plus ‘beginning’) can be created. Thus, this word is coded 1 for the *inchoative* aspect. Likewise, this word also includes the *durative* aspect since a suffix of 中 can be attached to 交渉, as in 交渉中 meaning ‘during negotiation’. This word is also coded 1 for the *durative* aspect. Similarly, a compound word 交渉終了 meaning ‘completion of negotiation’ can be created by compounding together ‘negotiation’ and ‘completion’. Furthermore, the suffix ‘after’ 後 can be attached to this word as 交渉後 ‘after negotiation’. Thus, it is judged that the compound word ‘negotiation’ contains the *telic* aspect in its meaning and receives a 1 code for this aspect. However, the word ‘negotiation’ does not have the *stative* aspect as does, for example, the word 健康, ‘health’. Thus, ‘negotiation’ is given 0 for this aspect. In sum, the word 交渉 is coded 1 for the *inchoative* aspect, 1 for the *durative* aspect, 1 for the *telic* aspect and 0 for the *stative* aspect. Values consist of only 1 if nouns contain an aspect or 0 when they do not contain an aspect. Variables to which 0 or 1 values are assigned are commonly referred to as *dummy variables*. In the present study, the four aspects were deemed X_1 for *telic*, X_2 for *durative*, X_3 for *stative*, and X_4 for *inchoative*.

4. Percentages of light verb attachments predicted by the four aspects

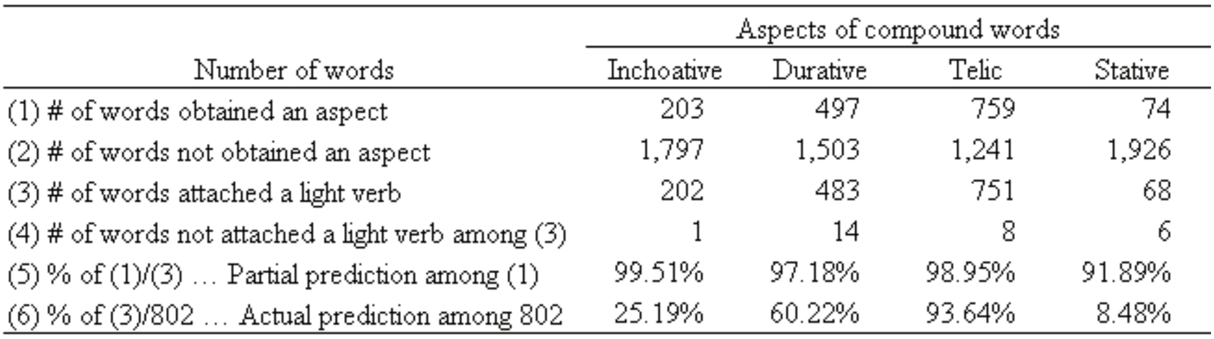
Once the presence or absence of the four aspectual properties of a two-kanji compound word has been calculated, it is possible to see to what extent these features predict whether or not the light verb is attached; that is, the simple predicted percentage of light verb attachment may be estimated.

As shown in Table 2, 203 two-kanji compound words were judged to have the *inchoative* aspect, of which the light verb could be attached to 202 words, or 99.51%. Only one noun possessing inchoative aspect, 戦争 ‘war’, does not combine with the light verb, i.e. the form 戦争する does not occur[8]. It is noteworthy, however, that it is possible to add *–suru* if this word has the accusative case marker *–o*, as in 戦争をする [$_{\text{vp}}$ NP(‘war’)-*o* V]. The light verb can be attached to 802 types out of 2,000; of these 802 types, the *inchoative* aspect only occurs with 202 types, or 25.19 percent.

Similarly, 497 two-kanji compound words were judged to have the *durative* aspect. Out of these 497, the light verb can be attached to 483 words, or 97.18%. This is 60.22% of the total of 802 types to which the light verb can be attached.

The highest prediction was provided by the *telic* aspect. A large group of 759 words contain this aspect. The light verb can be attached to 751 words, or 98.95%. This is 93.64% of the 802 words. Only 8 words do not have the light verb attachment.

Table 2
Predicted percentages of the light verb attachment by four aspects of 2,000 two-kanji compound words



	Aspects of compound words			
Number of words	Inchoative	Durative	Telic	Stative
(1) # of words obtained an aspect	203	497	759	74
(2) # of words not obtained an aspect	1,797	1,503	1,241	1,926
(3) # of words attached a light verb	202	483	751	68
(4) # of words not attached a light verb among (3)	1	14	8	6
(5) % of (1)/(3) … Partial prediction among (1)	99.51%	97.18%	98.95%	91.89%
(6) % of (3)/802 … Actual prediction among 802	25.19%	60.22%	93.64%	8.48%

Note : The light verb is attached to 802 two-kanji compound words out of 2,000 (40.10%).

[8] A combined form of the compound word ‘war’ plus the light verb as 戦争する may be used in colloquial speech, however it is not often accepted in writing. Thus, in this study, this word is judged not to have the light verb attachment.

The last aspect, *stative*, is seen with only 74 words. The light verb can be attached to 68 of these (i.e. 91.89% prediction). However, this accounts for just 8.48% of the 802 words to which the light verb can be attached.

In sum, as far as a simple cross-tabulation of the four aspects and light verb attachment shows, the *telic* aspect seems to be the best predictor, at more than 90% of the 2,000 selected compound words. Since the cross-tabulation in Table 2 only indicates a simple prediction of each aspect, a further analysis was conducted to estimate the predictive value of all four aspects together for attachment of the light verb.

5. Data analyses

Two different analyses were conducted for binomial data of the four aspects and light verb attachment. First, a binary logistic regression analysis was used for the selected 2,000 two-kanji compound words. Second, Hayashi's Quantification Theory Type III was used to estimate the similarity of the four aspects.

5.1 Binary logistic regression analysis

Using SPSS 11.0J for Windows - Regression Models with the same version of Base System, a binary logistic regression analysis was conducted to predict the light verb attachment (Y) from the four aspects (X_1 = *telic*, X_2 = *durative*, X_3 = *stative*, X_4 = *inchoative*) among the selected 2,000 two-kanji compound words. The results, reported in Table 3, give the following regression equation:

$$\mathrm{Log}_2(Y) = -3.47 + 6.93X_1 + 2.90\,X_2 + 2.81X_3 + 3.74X_4\,.$$

The determination coefficient (or variable explained by this equation) using Nagelkerke R^2 is high at 0.913, indicating that the four aspects have a high predictive power. All four aspects are significant predictors as shown in Table 3.

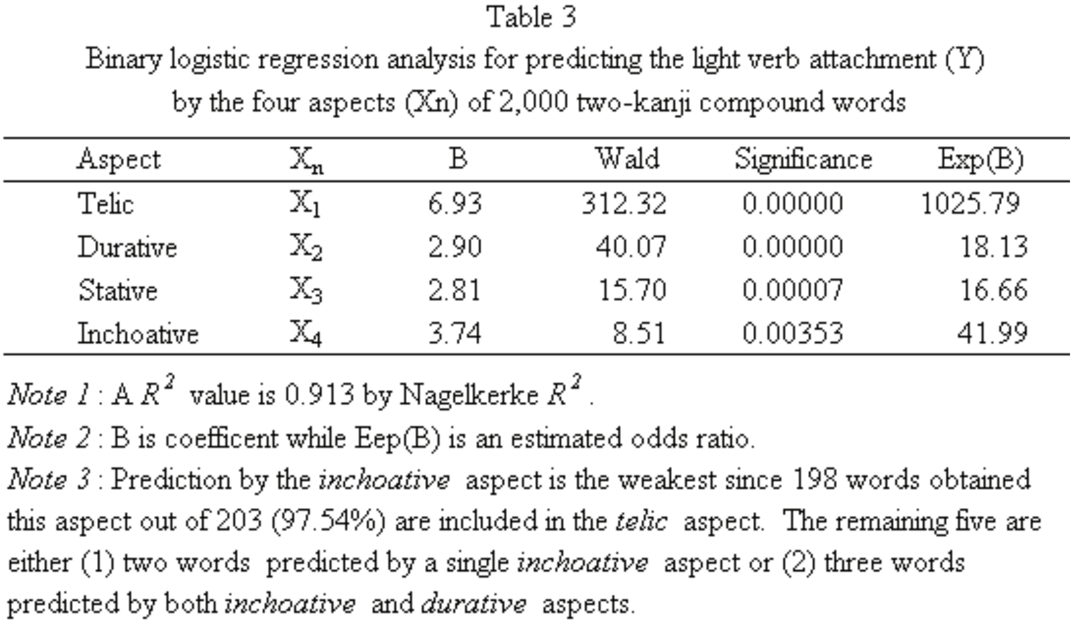


Table 3

Binary logistic regression analysis for predicting the light verb attachment (Y) by the four aspects (Xn) of 2,000 two-kanji compound words

Aspect	X_n	B	Wald	Significance	Exp(B)
Telic	X_1	6.93	312.32	0.00000	1025.79
Durative	X_2	2.90	40.07	0.00000	18.13
Stative	X_3	2.81	15.70	0.00007	16.66
Inchoative	X_4	3.74	8.51	0.00353	41.99

Note 1 : A R^2 value is 0.913 by Nagelkerke R^2.

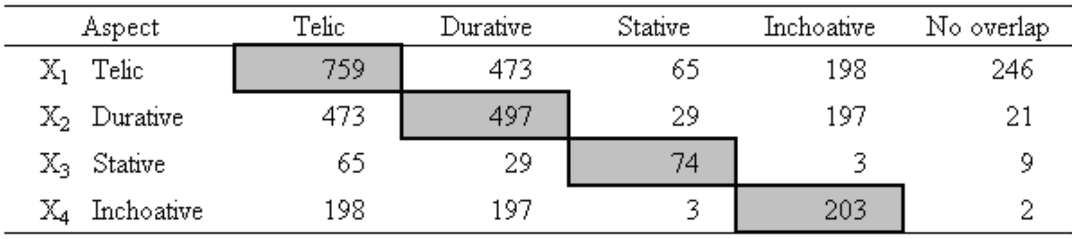
Note 2 : B is coefficent while Eep(B) is an estimated odds ratio.

Note 3 : Prediction by the *inchoative* aspect is the weakest since 198 words obtained this aspect out of 203 (97.54%) are included in the *telic* aspect. The remaining five are either (1) two words predicted by a single *inchoative* aspect or (2) three words predicted by both *inchoative* and *durative* aspects.

As expected from the cross-tabulation in Table 2, the aspect with the highest predictive value is the *telic* aspect, having a Wald value of 312.32 ($p < .00001$). The *durative* aspect is the second strongest predictor, having a Wald value of 40.07 ($p < .00001$). The third strongest is the *stative* aspect with a Wald value of 15.07 ($p < .0001$). The least predictive variable is the *inchoative* aspect with a Wald value of 8.51 ($p < .01$). 770 words out of the 802 nouns to which the light verb can be added contain at least one of the four aspects, while 32 words do not have any of these aspects. This represents a 96.01% prediction for light verb attachment. The *telic* aspect can predict the light verb attachment in 93.64% of cases (751 words out of 802 light-verb attached words), and the difference of prediction between this single aspect and all four aspects together is only 2.37% (32 words). This very small difference is the result of the overwhelming predictive power of the *telic* aspect, as shown by its very large Wald value of 312.32 and a very high estimated odds ratio of 1025.79.

The question which then arises is, how does the single variable of *telic* aspect predict light verb attachment so powerfully? The clue is in the overlap of these four aspects. As previously discussed, 759 two-kanji compound words had the *telic* aspect in their meanings. Among these, the light verb was attached to the 751 words. Only eight words containing *telic* aspect (e.g., 葬儀 'funeral', 犯行 'crime', 軍縮 'disarmament'[9]) cannot be directly combined with the light verb. Table 4 shows that, surprisingly, the 759 words with *telic* aspect include 198 out of 203 words with *inchoative* aspect (97.54%), 473 out of 497 words with *durative* aspect (95.17%), and even 65 words out of 74 with *stative* aspect (87.84%). Since the 751 words with *telic* aspect to which the light verb can be attached constitute 93.64% of the 802 words to which the light verb can be attached, this specific aspectual property by itself covers a great majority of the words which allow light verb attachment.

Table 4
Number of Overlaps as Classified by the Four Aspects



Aspect		Telic	Durative	Stative	Inchoative	No overlap
X_1	Telic	759	473	65	198	246
X_2	Durative	473	497	29	197	21
X_3	Stative	65	29	74	3	9
X_4	Inchoative	198	197	3	203	2

Note : The total of two-kanji compound words with the four aspects is 794 out of 2,000.

5.2 Hayashi's Quantification Theory Type III

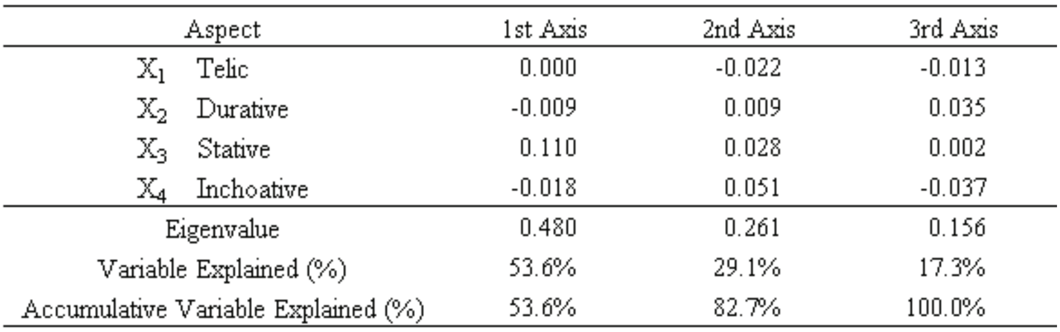
Hayashi's Quantification Theory Type III, of which model was created by Tokio Hayashi, is applied for investigating similarities of variables X_1 to X_4 with 0 and 1 binary data. The present study used the package created by a Japanese company of called *Esumi* produced software for the use of Hayashi's Quantification Theory Type I to Type III. Before making a strong prediction of the light verb attachment by the single aspect *telic*, Hayashi's Quantification Theory Type III was applied to estimate the similarity of the four aspects, by binary data of the four aspects using the 794 words which have at least one of the four aspects. The result

[9] Combined forms of these compound words may be used in colloquial speech, so that some cases of the light verb attachments may be acceptable. However, since the present study utilized the linguistic testing and intuition based mostly on written forms, a conservative position was adopted when determining whether or not the light verb could be attached.

is shown in Table 5. Three meaningful axes were found. The first axis has a high Eigenvalue of 0.480 with a high 53.6 percent of variability explained. The second axis has a reasonably high Eigenvalue of 0.261 with 29.1 percent of variability explained. The third axis has a rather low Eigenvalue of 0.156 with 17.3 percent of variability explained. The first and second axes together achieve an extremely high 83.6 percent of variability explained. Figure 1 depicts these two axes.

Table 5

Hayashi's Quantification Theory Type III analysis of the four aspects of two-kanji compound



Aspect	1st Axis	2nd Axis	3rd Axis
X_1 Telic	0.000	-0.022	-0.013
X_2 Durative	-0.009	0.009	0.035
X_3 Stative	0.110	0.028	0.002
X_4 Inchoative	-0.018	0.051	-0.037
Eigenvalue	0.480	0.261	0.156
Variable Explained (%)	53.6%	29.1%	17.3%
Accumulative Variable Explained (%)	53.6%	82.7%	100.0%

Note : This analysis was conducted using 794 words exhibiting at lease one of the four aspects.

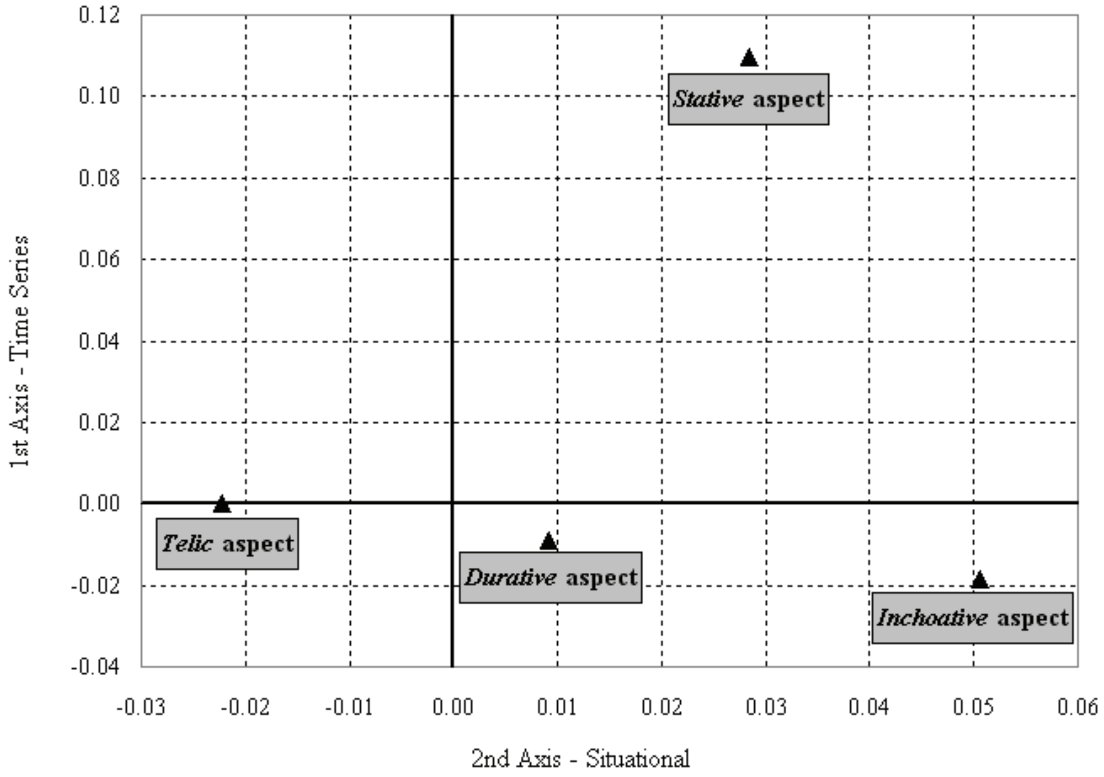


Figure 1 The four aspects of two-kanji compound words analyzed by Hayashi's Quantification Theory Type III

Three aspects – *telic*, *durative* and *inchoative* – seem to lie roughly within the small range between 0.00 and -0.02 of the first axis (the Y-axis in Figure 1). The locations of these three aspects close to the Y-axis may suggest that they, unlike the *stative* aspect, share a single property which will be referred to here as *time series* since these aspects refer to a sequential series of initiation, duration and ending. A word like 放送 meaning ‘a broadcast’ has an initiation point to start the ‘broadcast’, and continues during the ‘broadcast’ until the end of the ‘broadcast’. Thus, a single noun ‘broadcast’ can contain all the time series of the *inchoative*, *durative* and *telic* aspects. As such, the *time series* of the three aspects on the Y-axis, which provide nouns with a *verb-like* feature, can co-exist or exist independently in a single lexical item. Nevertheless, as indicated by the binary logistic regression analysis, of the three it is the *telic* aspect that is seen to be the dominant semantic feature in terms of making possible the attachment of a light verb. In addition, the *stative* aspect, which is plotted in a location apart from the other aspects, is also largely included in the *telic* aspect as shown in Table 4.

6. Conclusion – The most effective prediction for the light verb attachment

The present study investigated a unique linguistic characteristic of the light verb *–suru* which derives a verb form a noun when it is suffixed to that noun. However, the light verb cannot be attached to every noun, so the present study examined the four aspectual properties of nouns (Matsuoka, 2004, 2005) which may possibly determine light verb attachment. After investigating these four aspects in a large sample of 2,000 two-kanji compound words, a binary logistic regression analysis indicates that the dominant feature is the *telic* aspect. This single aspect can predict the light verb attachment of 93.64% of the 2,000 words. In contrast, there are only 8 items (out of 759 two-kanji compound words containing the *telic* aspect) to which the light verb cannot be attached. This implies an error rate of 1.05%. Again, it should be kept in mind that these 2,000 compound types cover a little less than 80% of the total two-kanji compound word tokens printed in the *Asahi Newspaper*. Therefore, the most effective approach to find the light verb attachment is simply as follows. If a word’s meaning includes the *telic* aspect, then there is an extremely high probability that the light verb can be attached to that word. This study has produced clear statistical evidence of the importance of the *telic* aspect for light verb attachment.

References

Amano, N., & Kondo, K. (2000). *Nihongo-no goi tokusei Vol. 7 [Lexical properties of Japanese – Volume 7]*. Tokyo: Sanseido.

Grimshaw, J. (1990) *Argument structure.* MIT Press.

Iida, M. (1987). Case-assignment by nominals in Japanese. In M. Iida, S. Wechsler and D. Zec (Eds.), *Working papers in grammatical theory and discourse structure: Interactions of morphology, syntax, and discourse* (pp. 93-138), Stanford, CA: CSLI.

Ito, T., Sugioka, Y. (2002). *Go no shikumi to go keisei [Functions of words and their formation]*. Tokyo: Kenkyuusha.

Kageyama, T. (1996). *Dooshi imi ron [Theory of verb semantics]*. Tokyo: Kuroshio Shuppan.

Matsuoka, C. (2004). Fukugoodooshi *–suru* o keiseisuru kango meishi ni tsuite [Sino-Japanese nouns used in *–suru* compounds]. *Nihongo Kyooiku [Journal of Japnaese Language Teaching]*, **120**, 13-22.

Matsuoka, C. (2005). *Ki’noodooshi –suru to meishi ga keisesuru bun no koozoo to imi: kangomeishi tono musubistuki o chuushin ni [Semantics and sentence structure constructed by the function verb –suru and nouns: Focusing upon combination with Sino-Japanese nouns]*. Doctorial dissertation submitted to Hiroshima University, Japan.

Shibatani, M., & Kageyama, T. (1988). Word formation in a modular theory of grammar: Postsyntactic compounds in Japanese. *Language*, **63**, 451-484.

Statistic packages used for the present study

SPSS 11.oJ for Windows – Base System
SPSS 11.oJ for Windows – Regression Model
Esumi Excel - Suuryooka Riron [Quantification Theory] Version 1.0

Glottometrics 10, 2005, 82-99

Constructing a Large-Scale Database of Japanese Word Associations[1]

Terry Joyce[2]

Tokyo Institute of Technology, Japan

Abstract. For cognitive scientists investigating the nature of lexical knowledge, one essential task is to map out the rich networks of associations that exist between words. This paper reports on a project to construct a large-scale database of word association norms for basic Japanese vocabulary and, utilizing the database, to develop lexical association network maps that tap into important aspects of words and their connectivity. The Japanese word association database will complement existing databases concerning the lexical features of Japanese vocabulary, such as familiarity ratings and frequency counts (Amano & Kondo, 1999; Yokoyama, Sasahara, Nozaki & Long, 1998), and the kanji corpus research highlighted in this special issue. Part 2 of this paper outlines the construction of the database, by detailing initial collections of word association responses from two major questionnaire surveys and the current state of the database. Part 3 introduces the lexical association network maps that will be developed based on the word association norm data and discuses some particularly promising applications of the database and the network maps in the areas of cognitive science and Japanese lexicography and language instruction.

Keywords: Japanese word association, large-scale database, lexical association network maps

1. Introduction

Given its crucial importance for many areas of cognitive science, such as psychology, artificial intelligence, computational linguistics and natural language processing, much research has, understandably, been devoted to investigating the nature of lexical knowledge and, in particular, to mapping out the rich networks of associations that exist between words. The interest in word associations is motivated by a number of converging perspectives and

[1] This research project is part of the 21st Century COE Program "Framework for Systematization and Application of Large-scale Knowledge Resources" (Program leader: Professor Sadaoki Furui) at the Tokyo Institute of Technology, Japan. The author would like to express his gratitude for help and support from Yokoyama Shōichi and Wada Yuriko at the National Institute for Japanese Language, Jim Elwood at the University of Tsukuba, as well as Tokosumi Akifumi, Matsumoto Naoko and Murai Hajime at Tokyo Institute of Technology.

[2] Address correspondence to: Terry Joyce, Ph.D., Tokyo Institute of Technology, W9-29, 2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552. Email: terry@valdes.titech.ac.jp

concerns within cognitive science. Central among these is the insight that, because association is a fundamental mechanism underlying human cognition, word associations mirror rather closely the structured patterns of relations that exist among concepts (Cramer, 1968; Deese, 1965). This notion is consistent with a number of influential assertions and inspirations within natural language processing research, such as Firth's (1957/1968) claim that a word's meaning resides in the company it keeps, Church and Hanks' (1990) notion of mutual information as a measure of the saliency of an association between two words, and Hirst's (2004) acknowledgement, notwithstanding certain caveats on the complex relationship between them, that a lexicon can often be a useful basis for developing a practical ontology. Lexical networks, whether represented with lexical nodes (i.e., in the tradition of Collins and Loftus, 1975) or as fully distributed features (e.g., Rumelhart, McClelland, and the PDP Research Group, 1986) are also at the heart of many connectionist models of human cognition, which, together with the key notion of spreading activation, derive much of their appeal from their neurological plausibility.

Recently, a number of studies highlight the valuable contributions that word association normative data can make to cognitive science (Nelson & McEvoy, 2005; Steyvers, Shiffrin, & Nelson, 2004; Steyvers & Tenenbaum, 2005). These studies all utilize *The University of South Florida word association, rhyme, and word fragment norms* (Nelson, McEvoy, & Schreiber, 1998), which is the largest database of word associations for American English, covering over 5,000 words with an average of 149 responses (SD = 15) per word collected from more than 6,000 participants. Demonstrating the influence of existing word associations on cognition, Nelson and McEvoy (2005) show that differences in the associative structures of known words—in terms of associate set size, resonance (or backward association), and the connectivity within an associate set—effect performance on the episodic memory task of extra-list cued recall. Steyvers, Shiffrin, and Nelson (2005) apply scaling techniques, such as singular value decomposition (SVD) and multidimensional scaling (MDS), to the word association norms database to create a word association space (WAS) representing the semantic (dis)similarity between pairs of words. Comparing three WAS-based measures with two latent semantic analysis (LSA) based measures (i.e., based on corpus collocations) in terms of predicting performance on three types of episodic memory tasks (recognition, free recall, and cued recall), they found that the WAS-based measures were better predictors of performance than the LSA-based measures for all three memory tasks. Steyvers and Tenenbaum (2005) also employ the database of word association norms as part of their analyses of the structures of large-scale semantic networks. Specifically, they used graph theory to analyze three semantic networks—one based on the word associations, one based on WordNet (Fellbaum, 1998), and another based on Roget's thesaurus. Steyvers and Tenenbaum found that all three semantic networks have statistical features in common, which they characterized as being small-world—having sparse connectivity, short average path lengths between words, and strong local clustering—and scale-free structures—most nodes have relatively few connections but are joined together via a small number of hubs with many connections.

This paper reports on a project to (1) construct a large-scale database of word association norms for basic Japanese vocabulary, to (2) utilize the word association norm data in creating and developing lexical association network maps, which capture important properties of words and their connectivity, and to (3) explore applications of the word association norms in the areas of cognitive science and Japanese lexicography and language instruction. Part 2 of the paper details the compilation of an initial survey corpus of basic Japanese vocabulary and initial collections of word association responses using a traditional questionnaire format. After outlining the current state of the database, Part 2 also notes plans for the future development of the database, including further collections of word association responses with computer-based and Internet-based versions of the survey, currently being developed, and for expanding the survey corpus. Part 3 of the paper starts by introducing the lexical association network maps that will be created from the database of word association norms. Part 3 also discusses some applications of the word association database and the network maps in the areas of cognitive science, such as in experimental control and as an approach to modeling the semantic representations of connectionist models, and of Japanese lexicography and language instruction, such as enriching the variety of lexical information within the lexical entry and providing user-friendly look-up functions.

2. Constructing the Database of Japanese Word Associations

2.1. Existing Word Association Norms

As it would be beyond the scope of this paper to attempt a review of word association norms (see Cramer, 1968; Deese, 1965; Moss & Older, 1996; Nelson, McEvoy, & Schreiber, 1998), the aim of this section is merely to briefly mention a couple of databases of word association norms for English and Japanese as frames of reference regarding the scale of the present project.

One large database of word association norms for British English has been created by Moss and Older (1996), which covers some 2,400 words with between 41-50 responses to each item. However, as already noted, the largest database of word association norms for American English is that constructed by Nelson, McEvoy, and Schreiber (1998), covering more that 5,000 words with an average of 149 responses for each item. It should be noted, however, that both these databases of association norms are the products of combining a number of surveys conducted over quite a number of years and that, rather than being systematic attempts to construct comprehensive databases, the inclusion of words in the surveys was usually in response to more immediate experimental interests at the time.

The first Japanese word association norms to mention are those collected in an early survey by Umemoto (1969). Although he gathered response from 1,000 university students, the word corpus is very small with only 210 words and thus of extremely limited value in controlling for the associative strength between stimulus items in experiments. More recently, Ishizaki (2004) has collected word associations as part of a project to build an associative concept dictionary (Okamoto & Ishizaki, 2001). Ishizaki's data covers 1,656 nouns with 10 responses for each item.[3] While arguably consistent with the aim of building an associative concept dictionary, a major drawback with this data, however, is the fact that response category was specified. Participants were asked to respond to a presented stimulus word according to one of seven randomly presented categories (hypernym, hyponym, part/material, attribute, synonym, action and environment), so the data tells us little about free associations.

[3] While this response count relates to version 1.0 made publicly available in March 2004, it seems that another version with 50 responses per item also exists.

2.2. Compiling a Survey Corpus of Basic Japanese Vocabulary

In order to compile an initial corpus of basic Japanese vocabulary for the word association survey, three reference sources were used. The first was the survey of basic vocabulary for Japanese language teaching conducted by the National Language Research Institute (1984). This list consists of approximately 6,800 words including a core set of about 2,200 words. The second reference source was Tamamura (2003), which is a recently prepared list of intermediate vocabulary of about 4,000 words. Because of its influence on Japanese language education, an important standard to look at when considering what constitutes basic Japanese vocabulary is the sanctioned list of Jōyō kanji. Accordingly, the third reference source was a handbook of Japanese orthography (Sanseidō Henshūjo, 1991), which lists all 1,945 Jōyō kanji with their official readings as well as a number of compound word examples (in total about 13,000 word tokens).

Once these lists were input, they were compared in order to identify common words, with priority on the overlap between the first two sources and particularly the core set of approximately 2,200 words within the National Language Research Institute's (1984) list. The task was made somewhat more difficult by the fact that Tamamura's (2003) intermediate vocabulary list has many words transcribed in hiragana that are transcribed in kanji in the National Language Research Institute's (1984) list. Reflecting the flexible nature of Japanese orthography and shifts in orthographic conventions over the last 20 years or so, the transcription differences highlight the merit of including orthographic variants within the survey. Related to that and the high incidence of homophones in Japanese, hiragana transcription words were frequently included for homophone sets within the corpus. For example, in the case of the homophone set of 合う 'to fit, suit, match', 会う 'to meet', and 遭う 'to meet, encounter (undesirable nuance)' sharing the pronunciation /au/, the hiragana transcription あう was also included. Also in an exploratory vein, a number of bound morpheme kanji were included. These include affixes, such as 不 /fu/ 'non-, un-', and verbal and adjectival stems, such as 書く /kaku/ 'to write' without the okurigana く /ku/, which are normally written with other kanji or okurigana endings and, in the strictest sense, are not words when written alone. Based on this work, an initial survey corpus of 5,000 kanji and words was created.

2.3. Questionnaire Surveys

In order to obtain the large-scale quantities of responses that will be required to complete the construction of the word association database, a computer-based version of the word association survey is being developed, so that the survey can also be conducted over the Internet. However, construction of the database is already progressing based on two surveys conducted using traditional pen-and-paper questionnaires. The first survey was conducted with the aim of obtaining up to 50 word association responses for a random sample of 2,000 items drawn from the survey corpus. Those responses will later be used to examine the consistency and reliability of word association responses to be collected with different formats of the survey, particularly those to be obtained from volunteer respondents participating in the survey via the Internet. The aim of the second survey was to obtain up to ten responses for the remaining 3,000 items in the corpus. Those responses will be used to control for intra-list association when respondent survey lists are generated automatically (as discussed in more detail in Section 1.4). Although the two surveys were conducted with different secondary aims, because the primary objective of obtaining word association responses in the construction of the database was common to both, and because the basic procedures were the same, they are outlined together.

2.3.1. Method

Participants. Native Japanese university students (N = 1,486; 934 males and 552 females; average age 19.03, SD = 0.97) participated in the surveys as volunteers.

Survey lists. For the first survey, 2,000 items were randomly drawn from the corpus and these were divided into 20 lists of 100 items. These items were divided so that each list consisted of a mixture of orthographic forms (i.e., single kanji, multi-kanji, and mixed kanji-kana words) in ratios closely matching the distribution within the overall corpus. Care was also taken to avoid intra-list associations, by ensuring that no two items within a list shared the same pronunciation and that no given kanji appeared more than once in a list either alone or as a constituent of a polymorphemic word. Finally, each list was examined by native Japanese graduate students so that all possible intra-list associations were eliminated. In order to obtain up to 50 word association responses for the 2,000 items, each survey list was presented to 50 respondents, but with the order of the items being randomized for each individual respondent. For the second survey, the remaining 3,000 items in the corpus were divided among 36 lists of 100 items. Care was again taken to control for the mixture of orthographic forms, incidences of homophones, and multiple inclusions of any given kanji. By the time the second survey was being prepared, the survey corpus had been coded with semantic category information (discussed further in Section 1.4 below). Thus, the lists for the second survey were created by also checking to ensure that no two items belonged to the same semantic category. Because of this extra control, it proved necessary to increase the number of survey list to 36 in order to cover some semantic categories with more member items. This also meant that some of the remaining 3,000 items appeared in two lists in the second survey. In order to obtain up to ten word association responses for the 3,000 items, each survey list was presented to at least 10 respondents, with the order of items within each respondent list randomized. In the event, more participants were available than minimally required for the secondary objective, so two of these lists were presented to 50 respondents, while another two were presented to 30 and 33 respondents respectively.

Each respondent list was printed with 10 items per page—the items were printed in 18pt Mincho beside an underlined blank space for the response (e.g., 本 __________) in a row centered on the page—forming a booklet of 10 pages plus a cover sheet with instructions. The instructions asked the participants to look at each printed item and to write down in the blank space the first semantically-related Japanese word that comes to mind. There were also instructions relating to aspects of the Japanese writing system. The first of these asked the participants to respond with what they considered to be the most natural orthographic representation of the associate response (i.e., whether they would normally write /manga/ in kanji as 漫画, in hiragana as まんが, or in katakana as マンガ). Another instruction asked participants not to change their response to another word if they found that they could not remember the correct strokes for the kanji of their first response, but to indicate that they were not confident of the correct strokes by providing the word's pronunciation in a hiragana gloss above the word.

Table 1
A Random Sample of 10 Items from the List of 2,100 Items in the Japanese Word Association Database (Version 1.0), with Respondent Counts, and Associate Set and Core Associate Set Sizes

Item	Respondents	Associate Set	Core Associate Set
遅らす	50	18	6
貝	50	23	5
耕す	50	14	5
引っ越す	50	34	6
公平	50	26	6
うらやましい	50	35	7
安心	50	32	8
ベンチ	50	18	6
触れる	50	26	4
最新	50	30	9

Note: Core associate set refers to the number of responses provided by two or more respondents.

2.3.2. Error Response Coding and Results

The word association responses collected with the paper questionnaires have been entered into a database by native Japanese graduate students. Blank spaces (no responses) were treated in two ways; in cases where a whole page had been skipped or where the participant failed to complete the questionnaire sheets, the items were regarded as having not been presented and accordingly are not reflected in the respondent counts, otherwise blank responses were recorded and, for the present, these are included as part of the set of word association responses for an item, as an indicator of words that are more difficult to make word association responses to. Items for which the response was illegible or involved a kanji selection error that resulted in an uninterpretable nonword were also treated as not presented items. When the response involved a minor writing mistake, such as incorrect kanji strokes or component element, but the intended response was clear from the presented word, the error was corrected. Responses based on phonological associations and transcription responses (i.e., where the respondent either provided the pronunciation in kana of a kanji orthography item or, more frequently, where the response to a kana orthography word was a kanji orthography word sharing that pronunciation) are currently being recorded and marked accordingly. Although the transcription responses could be indicating the need for more explicit instructions ruling out orthographic variants as invalid responses, it is also possible that the Japanese respondents regarded the orthographic variants as independent words. In cases of phrasal responses consisting of the presented item plus only one other word (excluding appropriate case markers), that word was taken as the response (i.e., when in response to 亡くなる ‘pass away’, one participant wrote おじいさんが亡くなった ‘my grandfather passed away’, おじいさん ‘grandfather’ was recorded as the response).

Table 2
Two Examples of Word Association Response Data in the Japanese Word Association Database (Version 1.0)

Item	Responses	Number	Item	Responses	Number
主語	述語	34	沸く	湯・お湯	35
	私・私は・わたし	4		沸騰	3
	動詞	2		水	3
	S (subject)	1		害虫	1
	が	1		歓声	1
	自分	1		敵	1
	修飾語	1		点	1
	誰か	1		電気	1
	使う	1		鍋	1
	名前	1		風呂	1
	話す	1		やかん	1
	人	1		ワールドカップ	1
	名詞	1			

Through two questionnaire surveys, 2,100 items randomly sampled from a survey corpus of 5,000 basic Japanese kanji and words were presented to up to 50 respondents. The responses to the 2,100 items have been processed to form the first version of the Japanese word association database, which is being made publicly available.[4] As illustrated with a random sample of 10 items in Table 1, a list of the 2,100 items together with respondent counts and the sizes of the associate set and core associate set (referring to the number of responses provided by two or more respondents) is available for download at http://www.valdes.titech.ac.jp/~terry/jwad.html. As the two examples presented in Table 2 show, the Japanese word association database lists all word association responses collected for the 2,100 items presented to up to 50 respondents. The associate set is ordered with the prime associate listed first. In the case of the two examples, the prime associate of 主語 ‘subject’ is 述語 ‘predicate’, given by 34 respondents, while the prime associate for 沸く ‘boil; get hot; get excited’ is 湯・お湯 ‘hot water’, given by 35 respondents. As more word association responses are collected for all items in the survey corpus, and the pattern of associations for each item becomes more stable, consistent with Nelson, McEvoy, and Schreiber (1998), the database will focus on the core associate sets, but responses provided by only one respondent are included in the present version of the database.

[4] Requests for the Japanese word association database (Version 1.0) may be directed via email to the author.

As Nation (1990) observes, in addition to knowing a word's spoken and written forms, its grammatical and collocation behavior, its frequency and stylistic register, as well as its conceptual meaning, one important aspect of lexical knowledge is knowing about the associations that a word has with other words. Figure 1 presents the associate set for the Japanese word 冬 'winter' based on the word association responses collected so far. The enclosed figures on the arrow connections represent the percentage of responses. As the figure shows, 冬 'winter' has a very strong primary associate with the word 寒い・さむい 'cold', which accounts for 44 percent of all responses. The second associate of 雪 'snow' represents only 15 percent of the responses, followed by 夏 'summer' and 冬至 'winter solstice', both at 6 percent, and 白・白い 'white' at 4 percent. Thus, 冬 'winter' has a relatively small set of core associates with one particularly strong associate.

In contrast, Figure 2 presents the associate set for the Japanese verb 集める 'gather, collect', which has a larger set of core associates, but, naturally, with weaker association strengths. The primary associate here is お金・金 'money' accounting for 15 percent of the responses. There are also two secondary responses at 10 percent; namely, 切手 'stamps' and 収集 'collection'. Some of the remaining core associates are 人 'people' (8%), 集合 'set' (6%), ゴミ 'rubbish, trash' (6%), and コレクター 'collector' (6%). Consistent with their respective word classes of noun and verb, these two words exhibit different kinds of syntagmatic responses. Compared to the very strong association between the adjective 寒い・さむい 'cold' and the noun 冬 'winter', more of the core responses for the verb 集める 'gather, collect' are nouns that could either occupy the direct object slot (i.e., お金・金 'money', 切手 'stamps', 人 'people', ゴミ 'rubbish, trash') or the subject slot (i.e., コレクター 'collector').

2.4 Future Development of the Database

In two traditional paper questionnaire surveys, approximately 148,600 word association responses for a corpus of 5,000 basic Japanese kanji and words were collected from 1,486 native Japanese speakers. These responses represent a substantial initial stage in the construction of a large-scale database of word association norms for basic Japanese vocabulary. However, given the preparation and inputting burdens involved in administering paper questionnaires, the present project is also developing a computer-based version of the word association survey, with a view to conducting the survey over the Internet in order to efficiently obtain the large-scale quantities of responses that will be required to complete the construction of the word association database. While the discrete free word association task is relatively straightforward—the respondent is simply asked to provide the first meaningfully-related word that comes to mind when presented with a stimulus word—the major issue in developing the computer-based survey has been to devise an automatic method of generating multiple individual respondent survey lists from the survey corpus, while minimizing as far as practically possible the effects of intra-list association.

Accordingly, much of the preparatory work on the project has been devoted to coding the survey corpus with information to use in eliminating intra-list associations. The first type of information added was phonological data in the form of hiragana transcriptions, to control for homophones and orthographic variants. The second type of information was a code relating to the orthographic form of the items (i.e., single kanji, multi-kanji, and mixed kanji-kana words, etc.) to ensure that respondent lists consist of a mixture of orthographic types, based on the distribution within the corpus, to reduce the possibility of form-related

Glottometrics 10, 2005, 99-108

Sizuo Mizutani (1926)
The Founder of Japanese Quantitative Linguistic

Naoko Maruyama [1]
Tokyo Woman's Christian University, Tokyo

1. Introduction

The Mathematical Linguistic Society of Japan was formed in December, 1956. It was the world's first academic society specifically focusing on mathematical aspects of linguistics. Sizuo Mizutani was involved as one of the driving members of its establishment. As the secretary-treasurer at the beginning, and as the president of the society in later years, he helped grow the society and created a new discipline – mathematical linguistics – for the Japanese language.

Mathematical Linguistics has several sub-fields: in addition to quantitative studies of Japanese such as quantitative lexicology and statistical stylometry, it also has studies on building mathematical (and algebraic) models of Japanese aiming at computational Japanese processing such as machine-translation.

Mizutani was a pioneer of quantitative and mathematical (algebraic) research in many fields (i.e. lexicology, grammar, semantics, and pragmatics). By proposing an original system in each of these fields, he established the basic principles and methodology in Japanese quantitative and mathematical (algebraic) linguistics in Japan.

Mizutani is regarded very highly by researchers of science and engineering, not only by Japanese language researchers.

Mizutani's life-long work is best represented by his book *Suuri Gengogaku (Mathematical Linguistics, 1982).* Also he was the editor of *Asakura Series on Japanese Linguistics,* consisting of six volumes (1983-87) which contains many papers related to Mizutani's researches.

2. Lexicology

In lexicology, Mizutani laid the foundation of quantitative lexicology through his research on word-count in several types of magazines and his analysis on this data by using mathematical methodology such as hypothesis testing, statistical estimation, and quantification theory while he worked at the National Language Research Institute. His representative work in this field can be found in the volume 2 of *Asakura Series on Japanese Linguistics, Lexicology (1983a)*. His main achievements in this field are as follows:

[1] Address correspondence to: Naoko Maruyama, College of Arts and Sciences, Tokyo Woman's Christian University, 2-6-1, Zempukuji, Suginami-ku, Tokyo 167-8585, Japan. E-mail: maruyama@lab.twcu.ac.jp.

(1) *Establishment of the field of "Quantitative Lexicology" through collecting data on word-count in magazines*

At the National Language Research Institute, Mizutani led several word-count projects for studying word usage. Through these activities, he established the methodology for quantitative lexicology. His methods are based on statistical sampling, and it was one of the first attempts of applying mathematical methods such as statistics into lexicology (Ito 2002). In his report on magazine word-count statistics (1957b), he showed the word-usage frequency of the top 1,000 Japanese words along with the estimated accuracy of the data. This was probably the world's first report on word usage statistics with estimated accuracy (Mizutani 1995c). Identification of word boundaries is one of the big problems in Japanese linguistics, but his data on the word-count of 90 magazines (1962, 1963) has very accurate word unit identification – it has by far the highest accuracy among similar word-count research to date (Ito 2002). Based on these contributions, combined with his work on creating the framework of word-count study at the National Language Research Institute, he is considered to be the father of Japanese lexicology. He introduced the term "quantitative lexicology" – Ito (2002) pointed it out as follows: "Establishment of quantitative lexicology provided substance to its parent field, lexicology. Lexicology research has been theories only. With the quantitative methods, lexicology now has proofs."

(2) *Distribution law of word classes*

About word usage distributions based on word classes, Ohno's Law had been well known. Mizutani expressed Ohno's Law as a mathematical formula, revised it, and implemented the revised version as a computer program written in BASIC (Mizutani 1981, Ito 2002).

Ohno's Law states that regarding 9 Japanese classic works, there are certain correlations between the style (genre of the literature) and the distribution of part-of-speech when counting number of different words (Ohno 1956). For example, Man'yo-shu (a collection of ancient Japanese poems) has the high proportion of nouns, followed by essays, journals, and fiction. On the other hand, verbs and adjectives are more common in fiction and journals. As shown in Figure 1, his discovery was represented graphically, which draw a lot of attention.

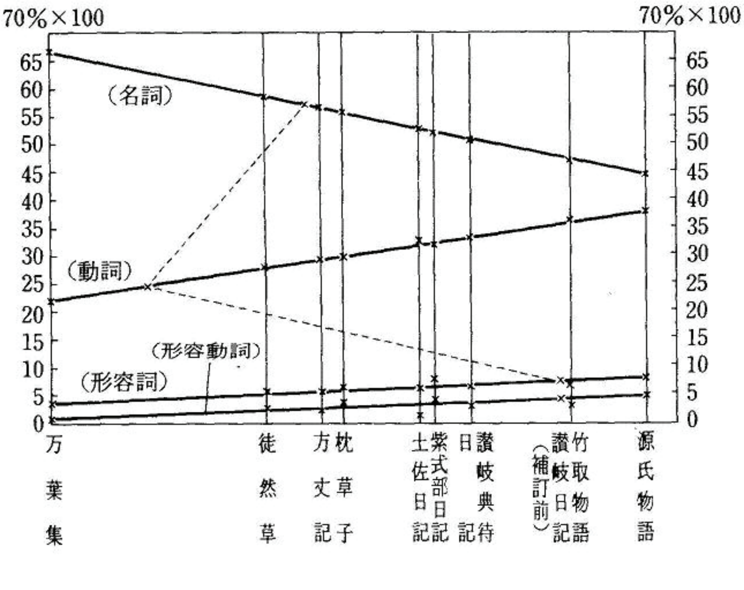


Fig. 1. Ohno's Original Chart (Reproduced)
名詞: noun, 動詞: verb, 形容詞: adjective, 形容動詞: adjective verb

Mizutani (1965) applied mathematical methods to Ohno's Law to give it a firm mathematical foundation. Instead of Figure 1 that has an unclear definition of X-axis, Mizutani proposed to use the proportion of nouns in X-axis and the proportion of other word classes in Y-axis. By applying the linear regression method, the proportion of each word class has a clear linear relationship to that of nouns.

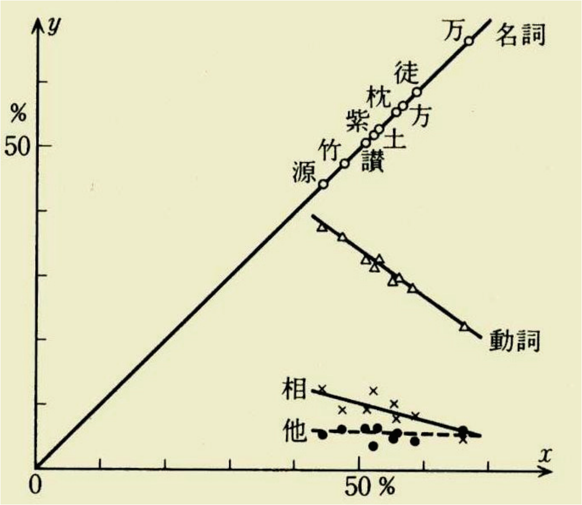


Fig. 2. Mizutani's Chart
名詞: Noun-class, 動詞: Verb-class, 相: Adjective-class, 他: Other-class

With Mizutani's corroboration, Ohno's law reveals favorable fitness for modern magazines as well as for classical literature.

(3) *Distribution law of relative word frequency*

Zipf's formula ($fr = C$, $kf^{2} = b$) and Mandelbrot's law ($p = A/(r+B)^{c}$) are well known as distribution laws of word frequency.

Mizutani proposed a formula as the form of distribution function based on word usage proportion p. He applied fitting against a projective-type distribution function (Mizutani 1953). Let α and β be structural parameters of the target set. $F(p)$ is the cumulative number of different words with frequency less than p, divided by the total size of the vocabulary. Then, for any p in the range of $0 < p \leq \max p$, $F(p)$ can be given by the following formula: $F(p) = p/(\alpha p + \beta)$

Mizutani's method has shown a higher degree of fitness than Zipf's law and Mandelbrot's law.

(4) *Estimation of amount of vocabulary*

The amount of vocabulary L is defined as the number of different words of some text h. The estimation problem of the amount of vocabulary is to estimate L given a text h. Mizutani proposed two methods (method A and method B) based on n-k relations – how k (the number of different words) increases as the size n of a random sample of h increases. See the following formula; K is the estimated number of different words in the sample of size n; a, b, and c are parameters.

A) $K = L(1 - \exp[- an - b])$

B) $K = l(1- \exp[- an]) + c$, (where $l = L- c$).

Given a set of (n, k) pairs from random sampling, we can estimate L along with the parameters a and b (in the case of Method A), or a and c (in the case of Method B).

(5) *Text similarity based on word similarity*

He also worked on quantifying text similarities based on the word usage in them. He proposed an index called Mizutani's D as a counter proposal to a previously known index called Miyajima's C. This comparison was reported in Chapter 3 of *Mathematical Linguistics (1982)*, in which he used the problem of estimating the origins of popular songs. Mizutani's D was shown to have a larger discriminatory power than Miyajima's index C.

In addition, he also applied Quantification Theory III and IV to word usage analysis of popular songs and classification of *waka* poems. In popular songs, his method could successfully distinguish such clusters as "hot-spring town elegies," "Lilu songs," "Flower girls," and "Shanghai songs." In *waka* poems, he could distinguish poems on plums from those on cherry blossoms.

3. Grammar (Syntax)

In the field of grammatical theory, Mizutani's goal was to create a formal grammatical description that can be executed on computers as well as specific enough to be refuted if a counter example is provided. His main works in this field are *Kokubunpo Sobyo (Sketch of Japanese Grammar, 1983b)* and *Kohon Koku-bunpo Daitai (Draft Framework of Japanese Grammar, 1991).*

His grammatical framework can find its root in the traditional Japanese theories of the Edo era (e.g., Motoori Norinaga) and Yamada's Grammar and Tokieda's Grammar, and refined them into a modern grammar. It distinguishes content words ("shi") and function words ("ji"), and the building rules of sentence structure can be expressed as a context free grammar.

Having started from the joint work with The Electrotechnical Laboratory (ETL) of Japanese Ministry of International Trade and Industry titled *Basic Japanese Grammar; Simple Sentence and Complex Sentence (1978),* he created his own theory of Japanese grammar through *Kokubunpo Sobyo (Sketch of Japanese Grammar, 1983b).* In later years he tried to describe a grammar in NBG logical formula in *Kohon Koku-bunpo Daitai (Draft Framework of Japanese Grammar, 1991).*

His originality in constructing Japanese grammar can be seen in many places, such as discussions on word-class, case-system, and modality. He proposed new ideas on how to treat keiyodoshi (adjective verb) as in *Keiyodoshi-Ben (Discussions on Adjective-Verb, 1951)* and *Keiyodoshi to Iu Mono (So-Called Keiyodoshi, 1952)*, and these papers drew a lot of attention. Later, he also tried to re-categorize nouns, adverbs, and adjective-verbs as in *New Refinement of Word Classes, From Nouns through Adverbs (1994)* and *Word Classification from the Point of Lattice-Theoretical View (2001).*

His works on case systems include the following: *Essay on the Case in Modern Japanese (1996), VALENZEN in a Broad Sense (as Case-combining Patterns) in Mathematical Literature (1997), Case-Combination Patterns in Japanese Novels 1948-1992 (1999).*

His analysis is always backed up by rich real data. In that sense, his research has both the theoretical side and the empirical side.

4. Semantics

In the field of semantics, Mizutani proposed the use of symbolic logic, which was an innovative approach that enables formal description of language semantics. His work in this field can be found in *Imi Kijutsu Taikei (Semantic Description Systems, 1995a).*

His started his research on semantics with *A Symbolic-Logic Based Method for Retrieving Facts from Legislation Sentences (1973).* He continued to focus on semantic description by means of symbolic logic, and compiled the work in *Imi Kijutsu Taikei (Semantic Description Systems, 1995a).* Logic-based semantics is given to the word level (such as kinship terms) then the semantics of sentences or even texts are constructed by logical formula. This logic-based approach allows a mathematical definition of language semantics and the capturing of objective relationships between expressions.

He also studied the semantics of compound words. *Coding Systems of Semantic Types of Compound Words (1989a, NTT Contracted Research), Distributions of Semantic Types in Compound Words Written in Chinese Characters (1988),* and *Word Construction Taken Account of High-Order Combinations (1989b)* are the representative works, of which there is no similar research.

5. Pragmatics (especially Honorifics)

In the pragmatics field, Mizutani proposed an original honorific system based on Tokieda's idea. We can learn about that system in *Taigu Hyogen-ron Teiyo (Outline of Honorific Expressions, 1995b).*

He defined numeric values representing levels of treatment, and created a model for which honorific expressions are used when the speaker, the listener, and the subject have specific treatment values. This model was implemented as a computer program, and used by several students to produce good experimental results..

The characteristic of Mizutani's honorific theory is to set three levels (the first order, the second order, and the third order) and to consider two kinds of honorific expression: those based on "shi" (content words) and those based on "ji" (function words). Mizutani's theory is based on Tokieda's idea, but Mizutani's contribution was to formalize Tokieda's idea and to make them computer implementable. Mizutani's theory can explain many honorific phenomena of Japanese including the concepts of affiliation and unification.

6. Other Achievements

Mizutani's other achievement are as follows:

(1) *Design of a Programming Language "Syusin (RedLip)"*
Although Mizutani is not an engineer, he designed a programming language called "Syusin (Redlip)". Syusin has SNOBOL-like character string handling. Its interpreter was implemented on top of LISP. Toshiaki Kurokawa of Toshiba helped Mizutani to design and implement Syusin (Kurokawa 1992). Its syntax is Japanese-like and easy to use, so many of Mizutani's students wrote programs for testing various theories on Japanese language phenomena by means of data processing. Many of these students are now working on computer programming in companies in Japan (or USA).

(2) *Editorship of Dictionary*
Mizutani has taken charge of co-editing *Iwanami Kokugo Jiten* since the second edition (1971). Now he is preparing the seventh edition. (The other two members have passed away, so since the fourth edition, Mizutani has been the supervising editor of the dictionary.)

(3) *Joint Research and Contracted Research*
Mizutani has been involved in numerous joint research and contracted research projects since early time. His major projects are as follows:

- *A Method for Retrieving Facts from Legislation Sentences Based on Symbolic-Logic (1973),* jointly with Toshiba
- *Basic Japanese Grammar; Simple Sentence and Complex Sentence (1978),* jointly with The Electrotechnical Laboratory (ETL)
- *Coding Systems of Semantic Types of Compound Words (1989a),* contracted with NTT

7. Closing

As we can see from his achievements above, undoubtedly Mizutani is one of the pioneers in Japanese mathematical linguistics. His main contributions to the fields are 1) bringing the ways of science and engineering into the field and 2) establishing the Mathematical Linguistics Society. His leadership enabled the growth of the society and the field at large – it is evidenced by the management of the society and the book series by Asakura Publishing. It would not be too much of an exaggeration to say that the most of the researchers in mathematical linguistics today have been coached or at least are influenced by Mizutani. They (including those in leading positions in the academia) respect Mizutani for his deep insights, broad knowledge, precise analysis, and his sharp criticisms. His contributions to the information engineering field should never be overlooked. Through serving as a committee member of a machine translation project of Japan Science and Technology Agency and developing programs himself, he played a key role in bridging the gap between linguistics and technology. His work is well known to researchers abroad: his grammatical theory was published in French, his lexicology was introduced in German. Also he is known as the editor of a famous dictionary (*Iwanami Kokugo Jiten)* to the general public – his influences and contributions to them are also very large.

Short Curriculum Vitae of Sizuo Mizutani

25 Mar. 1926	Sizuo Mizutani is born in Tokyo;
1940-1945	The First Higher School in Tokyo;
1945-1948	studies Japanese literature and linguistics at Tokyo University;

1948-1963	employment at National Language Research Institute;
	part-time lecturer of Kyoto University and Fukui University;

1956	driving member of establishment of Mathematical Linguistic Society of Japan; secretary-treasurer;
1964-1968	associate professor of Tokyo Woman's Christian University;
1966-1967	visiting scholar of Yenching Institute of Harvard University;
1967	casual research fellow of Computation Laboratory of Harvard University;

1968-1991 professor of Tokyo Woman's Christian University;
head of library, curriculum coordinator, and head of computer center in Tokyo Woman's Christian University;
part-time lecturer of Tokyo University, Tohoku University, and Tokyo University of Agriculture and Technology;
board member of The Institute of Behavioral Sciences;
chairperson of information retrieval committee of National Institute of Japanese Literature;
member of executive committee of the 8th International Conference on Computational Linguistics;
professional member of committee of Japan Society for the Promotion of Science;
councelor of the Society for Japanese Linguistics;
president of Mathematical Linguistics Society of Japan;
professional member of subcommittee of deliberative council for establishment of universities, etc.

1991 retired as professor emeritus of Tokyo Woman's Christian University

Main Books and Papers of S. Mizutani

1951 Keiyodoshi-Ben (Discussions on Adjective-Verb) , *Kokugo to Kokubungaku*, Vol. 28, No. 5, 31-46, revised No.6, 64.

1952 Keiyodoshi to Iu Mono (So-Called Keiyodoshi), *Kaishaku to Kansho*, Vol.17, No.12, 37-42.

1953 *Fujin Zasshi no Yoogo: Gendaigo no Goi Choosa (Research on Vocabulary in Women's Magazines)* National Language Research Institute (NLRI) Report 4, Tokyo, Shuei Shuppan. (joint work)

1957a Nobe-gosu to Kotonari-gosu tono Kankei (A Functional Relation between the Numbers of Different Words and Running Words). *Mathematical Linguistics*, No.3, 1-15, revised No.12, 4-6.

1957b,58 *Soogoo Zasshi no Yoogo [Zenpen/Koohen]: Gendaigo no Goi Choosa (Research on Vocabulary in Cultural Reviews)* NLRI Report 12, 13, Tokyo, Shuei Shuppan. (joint work)

1962,63 *Gendai Zasshi 90 shu no Yoogo Yooji, Vol.1, 3 (Vocabulary and Chinese Characters in Ninety Magazines of Today)* NLRI Report 21, 25, Tokyo, Shuei Shuppan. (joint work)

1965 Ohno no Goi-Hosoku ni tsuite (Notes on Ohno's Law of Vocabulary). *Mathematical Linguistics*, No.35, 1-13.

1970 *Gengo to Suugaku (Language and Mathematics)*, Tokyo, Morikita Shuppan.

1971 *Iwanami Kokugo Jiten (Iwanami Japanese Dictionary)* (the second edition), Tokyo, Iwanami Shoten. (joint work)

1973 *Kigoo Ronri-shiki ni yoru Hoorei-bun Kotogara Kensaku no Ichi-Hoohoo (A Symbolic-Logic Based Method for Retrieving Facts from Legislation Sentences)*, Tokyo Shibaura Denki Kabushiki-gaisha.

1974 *Kokugo-gaku Itsutsu no Hakken Sai-hakken (Japanese Linguistics; Five Discoveries and Rediscoveries)*, Tokyo, Tokyo Woman's Christian Univ., Sobun-sha.

1977 Keiryo-Goi-ron kara Mita Myozyo-ha to Negishi-ha (Myoozyoo School and Negisi School: A Statistical Analysis of *Waka*-Poem Vocabularies). *Mathematical Linguistics*, Vol.11, No.1, 30-37.

1978 *Nihongo Kihon Bunpo Tanbun-hen, Hukubun-hen (Basic Japanese Grammar; Simple Sentence and Complex Sentence)*, The Electrotechnical Laboratory (ETL) Report, 783, 784. (joint work)

1979 Yogo ni yoru Ume, Sakura no Uta no Benbetsu (Discrimination between *WAKA* Poems of Plum- and Cherry-Blossoms by their Word-Uses). *Mathematical Linguistics*, Vol.12, No.1, 1-13.

1980 Yogo-Ruiji-do ni yoru Kayokyoku Shiwake: "Yunomachi Elegy" "Shanhai-gaeri no Lilu" oyobi sono Shuhen (Classification of Popular Songs by Lexical Similarity: "Elegy of a Hot-Spring Town", "Lilu Returned from Shanghai" And Others). *Mathematical Linguistics*, Vol.12, No.4, 145-161.

1981 Kosei-hi no Senkei Kaiki Chosei Awasete Futatabi Ohno no Goi Hosoku (An Adjustment of Constituent Proportions by Linear Regression: Ohno's Law of Vocabulary, Again). *Mathematical Linguistics*, Vol.13, No.2, 92-97.

1982 *Suuri Gengogaku (Mathematical Linguistics)*, GendaiSuugaku Lectures D-3, Tokyo, Baifukan.

1983-87 *Asakura Nihongo Shin-kooza (Asakura Series on Japanese Linguistics)*, 6 Volumes, Tokyo, Asakura Shoten. (edit)

1983a *Asakura Nihongo Shin-kooza 2 Goi (Asakura Series on Japanese Linguistics, 2, Lexicology)*, Tokyo, Asakura Shoten.

1983b Kokubunpo Sobyo (Sketch of Japanese Grammar), *Asakura Nihongo Shin-kooza 3 Bunpo to Imi 1 (Asakura Series on Japanese Linguistics, 3, Syntax and Semantics 1),* the first chapter, Tokyo, Asakura Shoten.

1988 Kanji Hyoki Fukugogo ga Ninau Imi no Kata no Bunpu (Distributions of Semantic Types in Compound Words Written in Chinese Characters). *Mathematical Linguistics*, Vol.16, No.5, 205-211.

1989a *Fukugoo ga Ninau Imi no Kata no Code-Kei (Coding Systems of Semantic Types of Compound Words)*, NTT Contracted Research Report, a separate volume.

1989b Koji Ketsugo made Koryo-sita Go-Kosei Jookyoo (Word Construction Taken Account of High-Order Combinations). *Mathematical Linguistics*, Vol.17, No.2, 64-70.

1989c Japanese Quantitative Linguistics (*Quantitative Linguistics*, Vol.39 Bochum).

1990 *Class Ronri ni yoru Kokugo-gaku Seimitsu-ka (Japanese Linguistics with Class-logic)*, Tokyo Woman's Christian University, Department of Japanese Literature.

1991 *Kohon Koku-bunpo Daitai (Draft Framework of Japanese Grammer)*, Tokyo Woman's Christian University, Department of Japanese Literature.

1994 Meishi kara Fukushi made: Gorui no Atarashii Waku-duke (New Refinement of Word Classes, From Nouns through Adverbs). *Mathematical Linguistics*, Vol.19, No.7, 331-340.

1995a *Imi Kijutsu Taikei (Semantic Description Systems)*, Akiyama Shoten.

1995b *Taigu Hyogen-ron Teiyo (Outline of Honorific Expressions)*, The Institute of Behavioral Sciences.

1995c Suuriteki-kenkyuu (Mathematical Research). *Kokugogaku no Gojuunen (Fifty Years of Japanese Linguistics)*, 387-398, Tokyo, Musashino Shoin.

1996 Gendai-go no Kaku Shiron (Essay on the Case in Modern Japanese). *Mathematical Linguistics*, Vol.20, No.7, 283-303.

1997 Kogi Ketsugo-ka wo Suugaku-sho ni Miru (*VALENZEN* in a Broad Sense (as Case-combining Patterns) in Mathematical Literature). *Mathematical Linguistics*, Vol.20, No.8, 335-356.

1999 Sengo Shosetsu deno Kaku Ketsugo-gata (Case-Combination Patterns in Japanese Novels 1948-1992). *Mathematical Linguistics*, Vol.21, No.8, 345-360.

2001 Sokuron kara Mita Gorui-date (Word Classification from the Point of Lattice-Theoretical View). *Mathematical Linguistics*, Vol.23, No.3, 135-156.

References

ITO, Masamitsu (2002) *Keiryo Gengo-gaku Nyuumon (An Introduction to Quantitative Linguistics)*, Tokyo, Taishuukan Shoten.

OHNO, Susumu (1956) Kihon-goi ni kansuru Ni-san no Kenkyu (Studies on the Basic Vocabulary of Japanese: In the Japanese Classical Literature), *Kokugogaku*, No.24, 34-46.

KUROKAWA, Toshiaki (1992) *Software no Hanashi (Topics on Sortware)*, Iwanami Shin-sho, Tokyo, Iwanami Shoten.

Kokugo-gakkai (1980) *Kokugogaku Dai-jiten (Japanese Linguistic Dictionary)*, Tokyo, Tokyodo Shuppan. ("Keiryo Gengogaku" (Quantitative Linguistics), "Keiryo Goi-ron" (Quantitative Lexicology), "Suuri Gengogaku" (Mathematical Linguistics))