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

Compound	Average Frequency	Morphological structure (%)					Average appropriateness rating				Pronunciation familiarity
		M+M	V+C	C+V	SP	Others	M+M	V+C	C+V	SP	
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 re-

ported 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.