

# Two-Kanji Compound Words in the Japanese Mental Lexicon

Terry Joyce, Ph.D.

Large-Scale Knowledge Resources - 21st Century COE Program

Tokyo Institute of Technology, JAPAN

[terry@valdes.titech.ac.jp](mailto:terry@valdes.titech.ac.jp)

## Abstract

Given the productivity of compounding as a word-formation process, the Japanese language is especially suitable for investigating the extent of morphological involvement in the organization of the mental lexicon. This presentation introduces some of the constituent-morpheme priming experiments that I have been conducting to examine the lexical representation and retrieval of two-kanji compound word in the Japanese mental lexicon from a morphological perspective (Joyce, 1999; 2002; 2003a; 2003b). Essentially, these experiments are comparing the patterns of facilitation on lexical decision responses to compound words due to the prior presentation of a constituent kanji (relative to a baseline condition), across a number of word-formation conditions such as modifier + modified, verb + complement, complement + verb, and synonymous pairs. The results from Joyce (1999; 2002) were that reaction times in both constituent conditions were significantly faster than the baseline condition and, in the majority of cases, at similar levels across the word-formation conditions. However, the finding of a significant difference between the constituents only in the verb + complement condition, where the first constituent condition was faster, would seem to suggest an effect of verbal semantics. Further support for this notion comes from experiments (Joyce, 2003a; 2003b) that manipulated the positional frequency of the verbal constituent in the verb + complement and the complement + verb compound words, and observed a reversed pattern of priming across the high positional frequency conditions. These results are discussed in terms of the Japanese lemma-unit model, as a connectionist model of the Japanese mental lexicon (Joyce, 1999; 2002; 2004).

## Introduction

Given the productivity of compounding as a process of Japanese word-formation, the Japanese language is particularly suitable for investigating the extent of morphological involvement in the organization of the mental lexicon. This paper introduces some visual word recognition research conducted to investigate the lexical retrieval and representation of two-kanji compound words in the Japanese mental lexicon from a morphological perspective (Joyce, 1999, 2001, 2002a, 2002b, 2003a, 2003b, 2004; Joyce & Masuda, 2004; Joyce & Ohta, 1999, 2002; Masuda & Joyce, 2005). The paper also outlines the Japanese lemma-unit model (Joyce, 1999, 2002a, 2004). Beyond providing an appealing way of thinking about the representational and processing issues relating to two-kanji compound words, the model has important advantages relating to the complexities of the Japanese writing system, which arise from its multi-script nature (the mixture of kanji and two sets of kana, as well as the alphabet) and from the system of

nature (the mixture of kanji and two sets of kana, as well as the alphabet) and from the system of dual-readings for kanji (with both on- and kun-readings). Finally, the paper mentions the construction of a large-scale database of Japanese word associations as a resource for cognitive science research, highlighting two direct applications of the word association data to mental lexicon research (Joyce, 2005a, 2005b, 2005c).

### **The linguistic classification of kanji as a writing system**

Undoubtedly, the way that different writing systems are perceived has a direct impact on the kinds of questions that psychologists deliberate over when thinking about the mental lexicon. Based on a review of classifications of writing systems and the debate surrounding the various terms that have been applied at different times to kanji, such as ideographic and logographic, Joyce (2002b) has argued that, most precisely, kanji are a morphographic writing system—these orthographic units represent morphemes of the language. This claim is supported by examining the morphology of two-kanji compound words, because the word-formation principles underlying such polymorphemic words are primarily morphological in nature.

The principle psycholinguistic question motivating the research presented here is how is the morphographic nature of kanji reflected in the organization of the Japanese mental lexicon? That is, what role does morphological information play in the lexical retrieval and representation of Japanese polymorphemic words, such as two-kanji compound words? The involvement of morphological information in the mental lexicon has been one of the most debated topics within mental lexicon research over the last three decades, with the primary focus on contrasting alternative accounts of the lexical retrieval and representation of polymorphemic words.

### **Constituent-morpheme priming and the Japanese lemma-unit model**

The diversity in the morphological structure of two-kanji compound words is a matter of special concern for models of the Japanese mental lexicon. While a number of important models have been proposed for the Japanese mental lexicon, the discussion here focuses on two proposals that specifically address the issues of lexical retrieval and representation for two-kanji compound words; namely, Hirose's (1992) hypotheses and the Japanese lemma-unit model (Joyce, 1999, 2002, 2004).

Hirose (1992) has evoked search-like mechanisms within his hypotheses concerning the organization of two-kanji compound words. His hypotheses are based on results from a constituent-morpheme priming study that compared the facilitation on lexical decision responses to compound words due to prior presentation of a constituent kanji (relative to a baseline condition). Hirose interpreted the results of significantly faster reaction times in the first constituent prime condition compared to the second constituent prime condition as evidence of serial processing (from left to right) for compound words. He also suggested that the mental lexicon for compound words is structured so that words sharing the same first kanji are linked in clusters, with the first kanji serving as a retrieval cue, but that words sharing the same kanji as a second constituent are not.

In contrast, the Japanese lemma-unit model takes an alternative approach to modeling the lexical

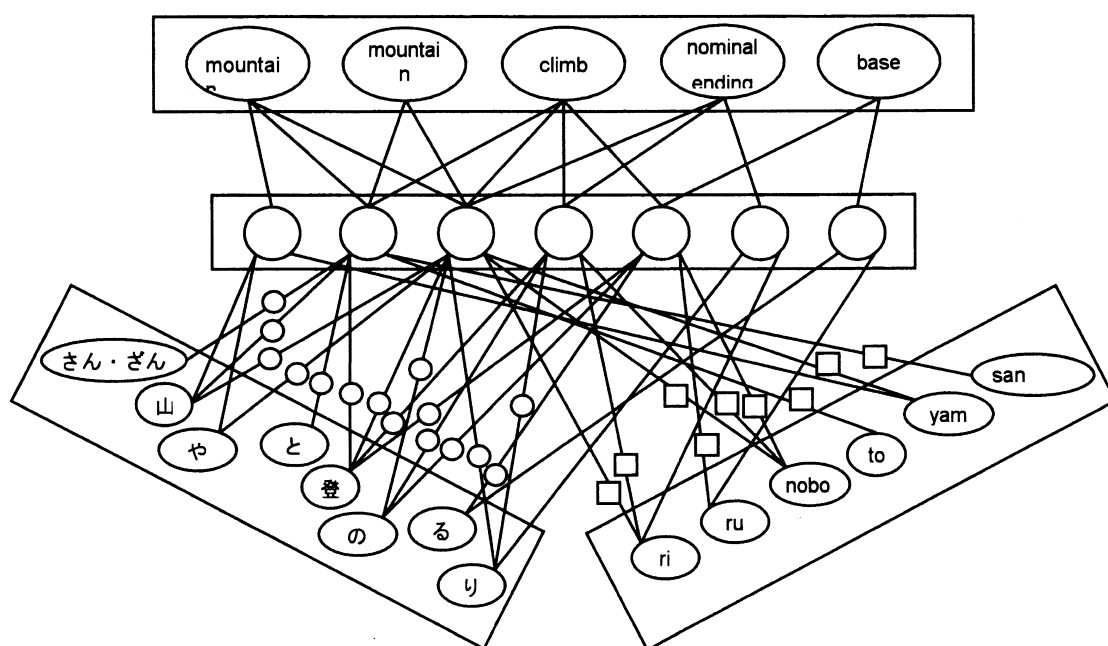
retrieval and representation of two-kanji compound words. This model is an adaptation for the Japanese mental lexicon of the version of the multilevel interactive-activation model for Chinese proposed by Taft, Liu and Zhu (1999). In order to solve problematic aspects with previous versions relating to representational redundancy, homographs, and the semantic transparency of compound words, Taft, Liu and Zhu made an important modification, which was to incorporate a layer of lemma unit representations, as abstract forms of a lexical entry, to mediate the connections between access representations (both orthographic and phonological) and semantic representations. Because these problems are also relevant in modeling the Japanese mental lexicon, the inclusion of lemma unit representations in a Japanese version is clearly well motivated.

As any model of the Japanese mental lexicon must capture in some way the morphological relations that exist between polymorphemic words, it is useful to compare the two proposals in terms of how they cope with the diversity in the morphological structure of two-kanji compound words. From a representational perspective, there appear to be problems for Hirose's (1992) hypothesis that compound words are linked in clusters based only on the first kanji. For example, for the reversed syntactic structures of verb + complement and complement + verb, there would be the rather strange situation that the words 登山 and 山登り, which are almost identical in meaning, are not be linked in the mental lexicon because of the reversed order of the kanji constituents. In the Japanese lemma unit model, on the other hand, the relations between such morphologically related words are modeled by the connections between representations. Turning to lexical retrieval, it is somewhat puzzling that Hirose's hypotheses are based on results from the constituent-morpheme priming paradigm, because using the same paradigm, Monsell (1985) found similar levels of facilitation from both constituents with both semantically transparent (e.g., tightrope) and opaque (e.g., butterfly) English compound words. Assuming activation as the mechanism of retrieval, the Japanese lemma-unit model would predict similar levels of priming from both constituents, but such a finding would not be compatible with the search mechanisms Hirose has proposed given the reliance on the first kanji as a retrieval cue.

As these two proposals make different predictions concerning constituent-morpheme priming, Joyce (1999, 2002) conducted two experiments to examine the pattern of facilitation once compound-word morphology was controlled for by including five word-formation principles as experimental conditions. The five conditions were (1) modifier + modified, (2) verb + complement, (3) complement + verb, (4) associative pair, and (5) synonymous pair compound words. Similar to the procedure used by Hirose (1992), Experiment 1 employed an SOA of 3,000 ms. The results indicated that across all five word-formation conditions, the reaction times in both constituent conditions were significantly faster than in the unrelated condition. However, apart from the verb + complement condition where reaction times in the first constituent condition were faster than in the second condition, no significant differences were observed between the constituents in most of the word-formation conditions. Hirose's (1992) experiment has, however, been criticized because of its extremely long SOA, with the suggestion being that at such a long SOA participants might adopt some response strategy when making lexical decisions. As Experiment 1 also used this SOA in replicating Hirose's procedure, it was necessary to investigate whether

participants did indeed adopt some response strategy that influenced the results from Experiment 1. According Experiment 2 employed a short SOA of 250 ms, as priming effects at that level are generally regarded to be relevantly automatic. The results of Experiment 2 closely matched those observed in Experiment 1. Again, reaction times in both constituent conditions were faster than the baseline condition in all five word-formation principle conditions, and the only significant difference between the first and second constituent conditions was in the verb + complement condition. Replicating the same pattern of results from Experiment 1, and at very similar reactions times, suggests that the participants in Experiment 1 did not adopt response strategies in making lexical decisions. Overall, these results are more consistent with the prediction from the Japanese lemma-unit model, providing support for the validity of this model as a model of the Japanese mental lexicon.

As shown in the figure of the model overleaf, the special feature of this model is the incorporation of lemma unit representations to mediate the connections between access representations and semantic representations, providing a very appealing way of thinking about the connections between morphological, semantic and syntactic information that underlie the processes of word-formation in Japanese. The incorporation of the lemma unit representations also makes it possible to model the complex relationships that exist within the Japanese writing system. One feature of the Japanese lemma unit model, which is not addressed in other Japanese models concerned with kanji processing, is the fact that it integrates both kanji and kana processing within a single model. Another advantage is the ability to account for the dual reading system of on- and kun-readings. Those aspects of the model have been investigated in a series of two cross-script and two cross-modality versions of the constituent-morpheme priming paradigm (Joyce, 2004).



Lemma-unit-mediated connections in the Japanese lemma unit model (Joyce, 1999, 2002)

### **Positional effects for verbal constituents**

One aspect of the results from Joyce (1999, 2002) requiring further consideration is the finding that lexical decisions were significantly faster for the first constituent compared to the second constituent only for verb + complement compound words. While Taft, Zhu and Peng (1999) suggest that character representations are not positionally-sensitive for Chinese compound words, the possible influence of word-formation principle at the compound word level cannot be ruled out, because the positional frequency of a kanji is likely to be closely related to its word class. To investigate this possibility, Joyce (2003a, 2003b) conducted a set of three experiments that controlled for the positional frequency of the verbal constituents in verb + complements and the reversed order of complement and verb compound words, based on the constituent morpheme frequency data provided by Joyce and Ohta (2002).

In the results from Joyce (2003a, 2003b), in the high positional ratio (HPR) verb conditions, reversed patterns of facilitation were observed across the reversed syntactic patterns of verb + complement and complement + verb compound words. That is, reaction times for the verbal constituent as the first constituent for HPR-verb + complement compounds were faster compared to the complement as the second constituent, while reaction times for the verbal constituent as the second constituent for complement + HPR-verb compound words were faster compared to the complement as the first constituent. Thus, these results suggest a positional frequency effect for verbal constituents, indicating that at least some kanji representations include information about their positional frequencies.

Additional evidence for the notion of verb morphology effects has also come from an experiment conducted by Joyce and Masuda (2004) with three short SOA conditions of 60 ms, 150 ms, and 250 ms in order to investigate the time courses of orthographic, phonological, morphological and semantic activation for two-kanji compound words. Interestingly, a reversed pattern of priming was observed between the verb + complement and the complement + verb compound words across the two shortest SOA conditions. Specifically, the priming effect for the verbal constituent as the first constituent condition of verb + complement compound words was greater than for the complement as the second constituent condition under the shortest SOA condition of 60 ms. However, under the 150 ms SOA condition, the priming effect for the verbal constituent as the second constituent of complement + verb compound words was greater than for the complement as the first constituent. In order to further investigate the time courses of activation for two-kanji compound words, a new SOA condition of 90ms has subsequently been added to this experimental design and the results for that condition will soon be reported (Joyce & Masuda, forthcoming). Taken together, these results suggest that models of the Japanese mental lexicon, like the Japanese lemma unit model, need to be capable of accounting for positional frequency effects at the morpheme level in the lexical retrieval of polymorphemic words.

### **Database of two-kanji compound words featuring morphological family, morphological structure and semantic category data (Masuda & Joyce, 2005)**

Masuda and Joyce (2005) have constructed a database of two-kanji compound words featuring morphological family, morphological structure, and semantic category data, as a valuable

resource for research into the morphological aspects of two-kanji compound words.

The central component of the database is a list of approximately 78,400 two-kanji compound word that are headwords in the Kōjien dictionary and of which both constituent kanji belong to the Japanese Industrial Standard level 1 (JIS1) kanji list. A major component of the database is the morphological family data, where newly calculated counts for the Kōjien list of compound words are combined with the newspaper usage-based constituent morpheme frequency counts compiled by Joyce and Ohta (2002). Another important component of the database is the data relating to the morphological structure of two-kanji compound words. In addition to the word-formation classification evaluations for a corpus of approximately 1,500 compounds collected by Joyce (Joyce & Ohta, 1999), this component also presents data from 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, currently covering approximately 11,000 compound words. Finally, the database includes semantic category data for approximately 24,500 compound words, based on the National Institute for Japanese Language's (2004) word list according to semantic principle.

### **Construction of a large-scale database of Japanese word associations (Joyce, 2005c)**

In another line of research to investigating the nature of lexical knowledge within the Japanese mental lexicon, Joyce (2005a, 2005b, 2005c) is currently conducting a research project to map lexical knowledge for basic Japanese vocabulary. Specifically, the project is (1) constructing a large-scale database of word association norms for basic Japanese vocabulary, (2) utilizing the word association norm data to create and develop lexical association network maps, which capture important properties of words and their connectivity, and (3) exploring various applications of the word association norms and the maps in the area of cognitive science, as well as Japanese lexicography and language instruction. Based on the results from two large-scale questionnaire surveys, the present version of the database (version 1.0) provides word association responses from up to 50 respondents for a corpus of 2,100 items randomly sampled from the larger project corpus of 5,000 basic kanji and words. In order to efficiently obtain large-scale quantities of word association responses in constructing the database, a web-based version of the survey is currently under development and will be online in the near future.

While it is beyond the scope of the present paper to outline all the applications of the word association database and the maps being considered, it is worth singling out two applications that are particularly relevant to the present focus on the Japanese mental lexicon. While the series of constituent-morpheme priming experiments introduced above provides strong evidence indicating a central role for morphological information in the organization of the Japanese mental lexicon, clearly other factors need to be carefully controlled for before alternative explanations can be ruled out completely. One factor that seems pertinent to these experiments is the theoretical possibility that the results are reflecting association effects between the single kanji primes and the two-kanji compound word targets. Although the patterns of priming have

been fairly robust across a number of experiments employing different sets of stimuli, still this is clearly something that warrants further investigation. Accordingly, one of the central motivations for the mapping lexical knowledge project has been to create the kind of comprehensive database of Japanese word associations required to be able to systematically manipulate the associative strengths between primes and targets in future versions of the constituent-morpheme priming experiments. Another interesting application in the future will be to further refine the representation of lexical and semantic information within the Japanese lemma-unit model by incorporating the lexical association network maps being developed based on the word association data. The inclusion of the lexical association network maps within the Japanese lemma-unit is undoubtedly a promising approach to developing more realistic models of the Japanese mental lexicon.

In conclusion, this paper has discussed two-kanji compound words in the Japanese mental lexicon, and has introduced the Japanese lemma unit model, as a connectionist model that is capable of providing a satisfactory account of the lexical retrieval and representation of two-kanji compound words in the Japanese mental lexicon from a morphological perspective. While comparatively little attention to date has been given to specifying in detail how to implement the balance between morphological, semantic and syntactic information, as already noted, these are concerns that will be addressed more fully in future research. A key challenge for mental lexicon models seeking to account for the higher cognitive processes involved in comprehension is to implement more principled divisions of semantic and syntactic information than currently employed in many models. In such an endeavor, the lemma-unit representations will be extremely important, because the interfacing function of the lemma unit representations provides an elegant way of connecting the four main grouping of information constituting a lexical entry—with orthographic (kanji, kana, and alphabet), phonological (on- and kun-readings), semantic and syntactic representations all linked and mediated by the lemma unit representations at the center.

## Acknowledgements

Parts of the research discussed in this paper have been conducted under the “Mapping Word Knowledge for Basic Japanese Vocabulary” research project which 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.

## References

- Hirose, H. (1992). Jukugo no nichu katei ni kan suru kenkyū: Puraimingu hō ni yoru kentō [An investigation of the recognition process for jukugo by use of priming paradigms]. *Shinrigaku Kenkyū*, 63, 303-309.
- 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. (2001). Kanji niji jukugo no puraimingu: Kōzō yōso jōhō kurosu suririputo no teiji to kurosu

- modariti no teiji [Two-kanji compound word priming: Cross-script and cross-modality presentation of constituent information] *Proceedings of the 65th Meeting of the Japanese Psychological Association*, 7-9 November, University of Tsukuba, Tsukuba, Japan, p. 218.
- Joyce, T. (2002a). Constituent-morpheme priming: Implications from the morphology of two-kanji compound words. *Japanese Psychological Research*, 44, 79-90.
- Joyce, T. (2002b). *The Japanese mental Lexicon: The lexical retrieval and representation of two-kanji compound words from a morphological perspective*. Unpublished doctoral thesis. University of Tsukuba, Japan.
- 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 niiji 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. (2005a). Mapping word knowledge for basic Japanese vocabulary. *Symposium on Large-Scale Knowledge Resources (LKR2005)*, Tokyo Institute of Technology, Tokyo, Japan, p. 29-32.
- Joyce, T. (2005b). Daikibō rensōgo dētābēsu no kōchiku [Constructing a large-scale word association database]. *Proceedings of the 69th Meeting of the Japanese Psychological Association*, 10-12 September 2005, Keio University, Tokyo, Japan, 629.
- Joyce, T. (2005c). Constructing a large-scale database of Japanese word associations. (Special issue on kanji corpus research edited by K. Tamaoka). *Glottometrics*, 10.
- Joyce, T., & Masuda, H. (2004). Kōsei keitaiso toshite no kanji no tanjikan senkō teiji ga kanji niiji 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, 613.
- Joyce, T., & Masuda, H. (forthcoming). Brief-presentation constituent-morpheme priming effects on the processing of Japanese two-kanji compound words. Presentation to be made at the *11th International Conference on Processing Chinese and Other East Asian Languages (PCOEAL 2005)*, 9-11 December, 2005, Chinese University of Hong Kong, Hong Kong.
- 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.

Masuda, H., & Joyce, T. (2005). A database of two-kanji compound words featuring morphological family, morphological structure, and semantic category data. (Special issue on kanji corpus research edited by K. Tamaoka). *Glottometrics*, 10.

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.

Taft, M., Liu, Y., & Zhu, X. (1999). Morphemic processing in reading Chinese. In J. Wang, A. Inhoff, & H.-C. Chen, (Eds.), *Reading Chinese script: A cognitive analysis* (pp. 91-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Taft, M., Zhu, X., & Peng, D. (1999). Positional specificity of radicals in Chinese character recognition. *Journal of Memory and Language*, 40, 498-519.