MAPPING WORD KNOWLEDGE FOR BASIC JAPANESE VOCABULARY

Terry Joyce, Ph.D.

Large-Scale Knowledge Resources COE, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
Email: terry@valdes.titech.ac.jp

ABSTRACT

The aims of this project are to compile a large-scale database of word association norms for basic Japanese vocabulary based on surveys employing the discrete free word association task, and, utilizing this data, to create lexical association network maps that capture important properties of words and their interconnectivity. After briefly introducing some visual word recognition research that underscores the need for comprehensive association data for cognitive science experimentation, this paper outlines preparatory work for the surveys and database. Finally, the paper mentions two promising applications of the lexical association network maps: namely, in modeling the semantic representations of connectionist models, like the Japanese lemma-unit model [1] [2], and in the area of lexicography.

1. INTRODUCTION

For cognitive scientists seeking to understand the complexities of higher cognitive functions, mapping lexical knowledge—revealing the network of connections that exist between words—is an essential task. This paper reports on a project to compile a large-scale database of word association norms for basic Japanese vocabulary and to create lexical association network maps that capture important properties of words and their connectivity.

As the project is an extension of ongoing psycholinguistic research into the nature of the Japanese mental lexicon, the paper begins by introducing some constituent-morpheme priming experiments and the Japanese lemma-unit model [1] [2]. The paper then outlines the compilation and coding of an initial survey corpus and data-collection preparations with a traditional questionnaire as well as computer and web-based formats.

Finally, the paper touches on some applications of the database and the lexical association network maps. The association maps represent a promising approach to modeling semantic representations for connectionist models, such as the lemma unit model, and this is a tangible example of how the maps can help to deepen our understanding of word knowledge. Another application is in the area of Japanese lexicography, where the database could not only enhance the variety of lexical information provided, but also support user-friendly search functions for electronic dictionaries.

2. VISUAL WORD RECOGNITION

2.1. Constituent-morpheme priming experiments

Compounding is an extremely productive word-formation process in Japanese, which makes the language particularly interesting for investigations into the extent of morphological involvement in the organization of the mental lexicon. In that context, Joyce has examined the lexical representation and retrieval of two-kanji compound words [1] [2]. Specifically, he has conducted a series of constituent-morpheme priming experiments—comparing the facilitation from component kanji in the lexical decision task—with two-kanji compound words representing various word-formation principles, such as modifier + modified, verb + complement, the reversed order of complement + verb, and semantic combinations of associative pairs and synonymous pairs.

The results from two early experiments were that, compared to an unrelated prime condition, reaction times in the two constituent prime conditions were significantly faster, and that that was the case for the five kinds of compound words. These results suggest that morphology is very important for the organization of the mental lexicon.

The results also indicated a possible effect of verbal semantics, because the only type of compound word with significant differences between the constituent prime conditions was the verb + complement combination, where the first constituent was faster than the second constituent. Albeit with qualification, further support for this effect of verbal semantics comes from a set of three experiments which manipulated the positional frequency of the verbal constituents in verb + complement and complement + verb types of two-kanji compound words. In the low positional frequency conditions, the constituents tended to facilitate lexical decisions at similar levels.

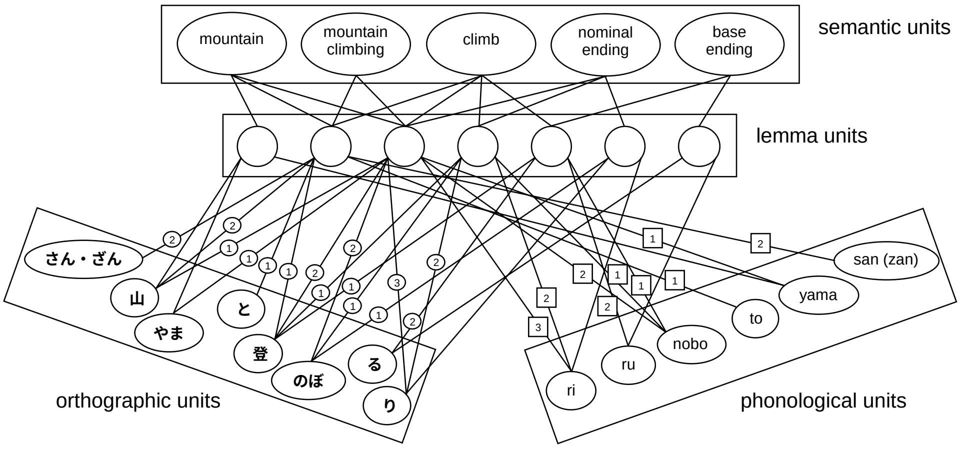

Figure 1. The Japanese Lemma-Unit Model (based on Joyce (2002, 2004))

In contrast, however, when the positional frequency was high, reaction times in the verbal constituent conditions were significantly faster than in the respective complement conditions. That is, for the verb + complement compound words, priming from the verb in the first constituent condition was greater than the priming from the complement in the second constituent condition, while for compound words with the reversed order of complement + verb, the pattern of priming was reversed. Additional evidence in the notion of verbal semantic effect has also been provided by a recent experiment that included very short stimulus onset asynchronicity (SOA) conditions to examine the time courses of orthographic, phonological, morphological and semantic activation [3]. Again, a reversed pattern of priming was observed between the verb + complement and complement + verb compound words across the two shortest SOA conditions of 60 ms and 150 ms.

2.2. The Japanese Lemma Unit Model

In order to account for the lexical representation and retrieval of two-kanji compound words, Joyce has proposed adapting for the Japanese mental lexicon a modified Chinese version of the multilevel interactive-activation framework [1] [2]. As the basic mechanism of the connectionist model is activation, it is able to explain the patterns of priming observed in the constituent-morpheme priming experiments.

As shown in Figure 1, a special feature of the model is the incorporation of lemma-unit representations, which may be regarded as abstract forms of the lexical entries. These mediate the connections between access representations (both orthographic and phonological form representations) and semantic representations.

The inclusion of the lemma unit representations is particularly appealing for handling the complex nature of the Japanese writing system. One important aspect of the lemma unit model, which is not addressed in other Japanese models concerned solely with kanji processing, is 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, which are differentiated by the links to the lemma-units.

Summarizing this visual word recognition research, the results from a series of constituent-morpheme priming experiments suggests a central role for morphological information in the organization of the Japanese mental lexicon. However, while this claim seems highly plausible, being consistent with related research in other languages, there are factors that need to be carefully controlled for before alternative explanations can be ruled out completely. One such factor that seems pertinent to the constituent-morpheme priming experiments is the possibility of association effects. The counter notion would be that these results are reflecting, or are being distorted by, differences in the associative strengths between members of a morphological-related family of compound words; the basic claim being that the results are due to the stimulus items employed.

Although the pattern of priming has been quite robust across the experiments, clearly this is something that warrants further investigation. Unfortunately, the kind of comprehensive word association data for Japanese that is needed to examine these issues is simply not available. Accordingly, a central objective of the mapping word knowledge project is to construct a comprehensive database of word association norms by conducting large-scale surveys for basic Japanese vocabulary.

3. WORD ASSOCIATION NORMS FOR JAPANESE

3.1. Existing databases of word association norms

This section briefly outlines work in compiling and coding an initial survey corpus and preparing for data collection with a traditional questionnaire as well as computer and web-based formats. As frames of reference regarding the scale of the project, this section starts by introducing some existing databases.

While Moss and Older have collected between 40-50 responses for some 2,400 words of British English [4], Nelson and his colleagues have compiled the largest database of American English covering some 5,000 words with approximately 150 responses per item [5]. It should be noted, however, that both these databases of association norms are the products of combining several smaller surveys conducted over quite a number of years in response to more immediate experimental interests rather than more systematic attempts to construct comprehensive databases.

Turning to Japanese language surveys, although the early survey by Umemoto is quite well-known, and gathered responses from 1,000 university students, the word corpus is very small with only 210 words [6]. More recently, Ishizaki has collected word associations for use in building an associative concept dictionary [7]. Ishizaki's data covers 1,656 nouns with 10 responses each in the version currently available. A major drawback with this data, however, is the fact that response category was specified, so it tells us little about free associations.

3.2. Initial survey corpus

Three primary reference sources have been used in compiling an initial corpus of basic Japanese vocabulary for the word association survey. The first is the survey of basic vocabulary for Japanese language education conducted by the National Language Research Institute, which includes approximately 6,800 words including a core set of about 2,200 words [8]. The second source is a similar, recent list of about 4,000 words prepared by Tamamura [9].

Because of its influence on Japanese language education, an important standard to consider in looking at what constitutes basic Japanese vocabulary is the sanctioned list of Jōyō kanji. Accordingly, the third reference source has been a handbook of Japanese orthography, listing all 1,945 Jōyō kanji with official readings, together with some 13,000 word examples (tokens) [10].

Once these lists were input, they were compared in order to identify common words. While the task was made more difficult by the fact that Tamamura's list has more items transcribed in hiragana, reflecting shifts in orthographic conventions over the last 20 years or so, this highlights the need to include orthographic variants within the survey. As a result of this work, an initial survey corpus of 5,000 kanji and words has been compiled. As data collection proceeds, this corpus will be expanded through the inclusion of associates that are not already part of the corpus.

3.3. Data-collection preparations

The discrete free word association task itself is relatively straightforward—the respondent is simply asked to provide the first word that comes to mind when presented with a stimulus word. However, a major concern for the project, in seeking to efficiently construct a large-scale database, 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.

To this aim, much of the preparatory efforts have been devoted to coding the kanji and words with information relating to the following criteria.

(1) Pronunciation: One interesting question arising from the complex nature of the Japanese writing system that the project will address is whether the associates of a word vary according to its orthographic form (i.e., kanji vs. kana orthography). Accordingly, it is necessary to ensure that each pronunciation only appears once in a survey list to control for orthographic variants and the high incidence of homophones in Japanese.

(2) Orthographic type: To reduce the possibilities of respondent strategies, the survey lists contain a mixture of orthographic types (i.e., single kanji, multi-kanji, and mixed kanji-kana words), achieved through type codes and limits based on the distribution within the corpus.

(3) Component kanji codes: In order to ensure that a given kanji only appears once in a survey list, all words containing kanji elements were marked with the kuten codes for the component kanji.

(4) Semantic category codes: It is also important to ensure that the survey lists do not contain multiple words from a particular semantic category, such as not having two colour words like 黒 'black' and 白 'white' together. In order to control for this, the corpus items were marked with the semantic category codes used in the NTT thesaurus [11].

(5) Unique ID code: All items have an ID code, and as data collection proceeds, this will serve as an additional measure in eliminating intra-list associations.

In order to obtain the large-scale quantities of responses required for the database, the project is developing computer-based and web-based survey formats. However, data collection is already underway employing a traditional paper questionnaire. A random sample of 1,000 items was taken from the corpus, and 50,000 responses have been collected from 500 university student respondents (100 items per questionnaire). Although this format involves data inputting burdens, this data is essential to address reliability issues related to data collection from a web-based survey. As this initial block of data is being input, the project is also focusing on the design and structure of the database, so that it is capable of supporting flexible searching for both respondent and association norm data.

4. ASSOCIATION DATA APPLICATIONS

The large-scale database of free word association norms will be a valuable resource for cognitive science research, such as memory experiments and the visual word recognition research outlined earlier. The project will also utilize the association database to create lexical association network maps, as a means of representing associate sets and their connectivity.

Figure 2 is an example of an association map for the English word 'planet' [12], which illustrates the basic concept behind the lexical association network maps that this project will develop. In addition to representing associate set size, the maps can highlight differences in association strengths, both forwards and backwards (which are independent), as well as properties such as the association density of the associate set.

The lexical association maps will complement other approaches to capturing aspects of lexical knowledge, while avoiding some of their problems. For instance, unlike the lexicographer's expertise required to define synonym sets WordNet [13], this maps will be based on the free responses from respondents. In this respect, the structure in the network will emerge from tapping free associations, rather than being theory-driven or constrained by specific relationships, which are possible concerns for associative concept dictionaries and ontologies where the hierarchical structure is pre-defined. The use of free associations is also a way of avoiding domain biases that can be an issue for collocation data.

In capturing the networks of association that exist between words, the maps open up a promising approach to modeling the semantic representations within connectionist models, like the Japanese lemma unit model. This approach is to incorporate maps, like Figure 2, into the upper portion of the model, as in Figure 1.

Another application that the project will explore is in the area of Japanese lexicography. The inclusion of word association data within the lexical entry will not only greatly enrich the variety of lexical information, but the association network maps could be used as a means of exploiting and enhancing electronic dictionaries by supporting user-friendly search functions. And, given the importance of contextual relevance for second language learning, the lexical associative networks could also aid in developing effective learning strategies for Japanese language instruction. In pursuing these applications, the project will create a comprehensive kanji database and integrated kanji instruction system. The key idea behind these is an electronic study notebook that can build into a personalized dictionary by drawing on the database reference source through learning assignments.

Accordingly, this project is endeavoring to contribute to the Large-Scale Knowledge Resources COE program in a number of ways, ranging from: (a) the construction of the word association database, and through (b) the development of lexical association network maps, to (c) the exploration of Japanese lexicographical and language learning applications.

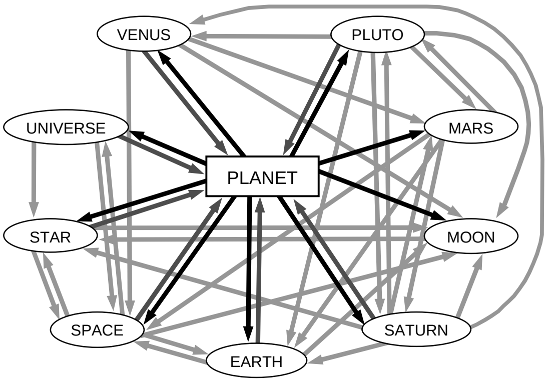

Figure 2. Basic concept of the lexical association network maps that the project will develop (based on Nelson & McEvoy, 2003)

5. REFERENCES

[1] T. Joyce, "Constituent-morpheme priming: Implications from the morphology of two-kanji compound words," *Japanese Psychological Research*, Blackwell, Japan, pp. 79-90, 2002.

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

[3] T. Joyce, and H. Masuda, "Priming effects from brief presentations of constituent kanji on lexical decisions for two-kanji compound words," (in Japanese), *Japanese Psychological Association Proceedings*, p. 613, 2004.

[4] H. Moss, and L. Older, *Birkbeck word association norms*, Psychological Press, Hove, 1996.

[5] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber, *The University of South Florida word association, rhyme, and word fragment norms*, http://www.usf.edu/FreeAssociation, 1998.

[6] T. Umemoto, *Rensō kijunhyō: Daigakusei 1000 nin no jiyū rensō ni yoru*, Tokyo Daigaku Shuppankai, Tokyo, 1969.

[7] S. Ishizaki, "*Rensō gainen jisho Verison 1.0*" CD, 2004.

[8] National Language Research Institute, *Nihongo kyōiku no tame no kihon goi chōsa*, Shuei Shuppan, Tokyo, 1984.

[9]. F. Tamamura, "Chūkyūyō goi: Kihon 4000 go," *Nihongo Kyōiku*, Tokyo, pp. 5-28, 2003.

[10] Sanseidō Henshūjo, *Atarashii kokugo hyōki handobukku (Dai yonhan)*, Sanseidō, Tokyo, 1991.

[11] S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ōyama, and Y. Hayashi, *Nihongo goi taikei [Goi-taikei-A Japanese lexicon] (CD-Rom)*, Iwanami Shoten, Tokyo, 1999.

[12] D. Nelson, and C. McEvoy, *Implicitly activated memories*: *The missing links of remembering*. TIC4, Tsukuba, Japan, 2003.

[13] C. Fellbaum (Ed.), *WordNet: An electronic lexical database*, MIT Press, Cambridge, MA, 1998.