Mapping word knowledge in Japanese: Constructing and utilizing a large-scale database of Japanese word associations

Terry Joyce, Ph.D.

Large-Scale Knowledge Resources COE,
Tokyo Institute of Technology, Tokyo, Japan
terry@valdes.titech.ac.jp

Abstract

This project is investigating lexical knowledge by mapping out the associative structures that exist for Japanese words. More specifically, the project is (1) constructing a large-scale database of Japanese word associations, by conducting free word association surveys with native speakers, (2) utilizing the association database to create lexical association network maps as a means of capturing patterns of association, and (3) exploring applications of the Japanese word association database and the lexical association network maps in the areas of cognitive science, lexicography and Japanese language instruction.

1. Introduction

Echoed in Firth's remarks about a word's company [1], in Church and Hank's idea of mutual information [2], and in Cantos and Sánchez's concept of lexical constellations [3], as well as in Hirst's comparison of lexicons and practical ontologies [4] is the simple yet extremely significant notion that the rich networks of associations that exist between words mirror in important ways the structured relations that exist between concepts, because association is a basic mechanism of human cognition [5][6]. In a similar vein, a number of recent studies provide some instructive examples of utilizing word association normative data to explore the structures embedded within lexical knowledge and to gain valuable insights into human cognition. These studies draw on the largest database of word associations for American English [7]. For example, Nelson and McEvoy have shown that the associative structures of known words effect memory performance [8]. Similarly, Steyvers, Shiffrin, and Nelson found that measures derived from a semantic space based on word associations were more predictive of episodic memory performance than LSA-based measures [9]. While Steyvers and Tenenbaum's analyses of semantic networks based on word associations, WordNet and Roget's thesaurus showed that they all shared common characteristics [10].

This paper briefly reports on the mapping word knowledge for basic Japanese vocabulary project, which is investigating the nature of lexical knowledge in Japanese [11][12][13][14][15][16][17][18]. After outlining the on-going construction of a large-scale database of Japanese word associations, through two conducted questionnaire surveys and the development of a web-based survey in Section 2, Section 3 touches on the development of lexical association network maps and highlights some interesting applications of the database and the maps in the areas of cognitive science, particularly mental lexicon research, as well as lexicography and Japanese language instruction.

2. Constructing the word association database

The mapping Japanese word knowledge project is seeking to create a database of Japanese word associations which will be large-scale both in terms of the number of words that are surveyed and the number of word association responses collected for each corpus word in the database. This section notes the compilation of an initial survey corpus of basic Japanese vocabulary, describes the first collections of word association responses through two conducted questionnaire surveys, and outlines plans for the future development of the database.

2.1. Initial survey corpus of basic Japanese vocabulary

The first task in constructing the large-scale database of Japanese word associations was to compile an initial corpus of basic Japanese vocabulary for which word association would be elicited in free word association surveys. To that aim, three reference sources [19][20][21] of basic vocabulary for Japanese language instruction were consulted and through a process of identifying common items, an initial survey corpus of 5,000 kanji and words was created.

In order to be able to automatically generate respondent lists in the future, while minimizing intra-list associations as far as practically possible, the corpus was also coded with various kinds of information. The kinds of information included pronunciation transcriptions in hiragana, orthographic-form codes (i.e., single kanji, multi-kanji, and mixed kanji-kana words), and component kanji codes (kuten codes), as well as semantic category codes, based on the National Institute for Japanese Language's recently revised semantic classification [22]. As a further measure, ID codes for collected word responses will also be added as feedback data.

2.2. Questionnaire surveys

The first collection of word association data was carried out through two large traditional pen-and-paper questionnaire surveys, which differed in terms of scale and coverage of the survey corpus. The aim of the first survey was to collect up to 50 word association responses for a random sample of 2,000 items, while the aim of the second survey was to collect ten responses for the remaining 3,000 items in the corpus. The responses from the more focused first survey will later be used to examine issues of the consistency and reliability when collecting data using different formats, while the response from both surveys will be extremely important as an additional feedback control on the automatic generation of respondent lists in the future.

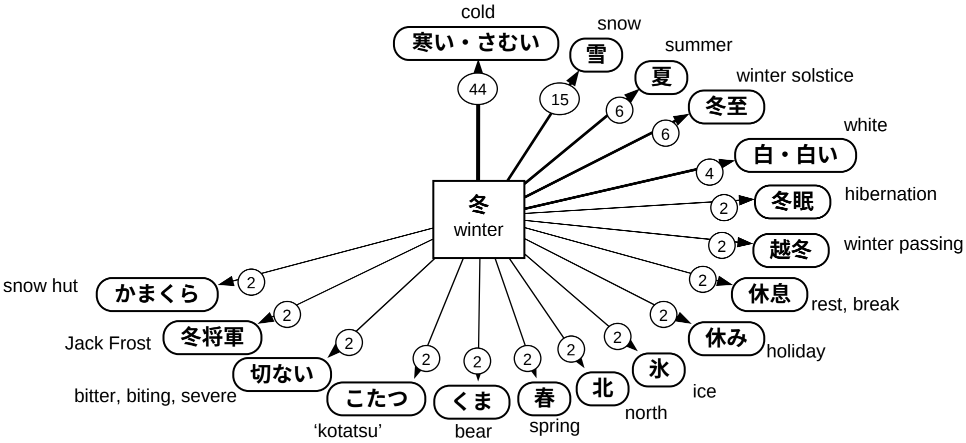

Figure 1: The associate set for 冬 'winter'.

2.2.1. Method

Participants: Native Japanese university students (N = 1,486; 934 males and 552 females; average age 19.03, SD = 0.97) participated in the two surveys on a volunteer basis.

Questionnaire sheets: For both surveys, the target items were divided into lists of 100 items. Based on the coded information, care was taken to ensure that no pairs of homophone words appeared in the same list, and that no kanji appeared more than once in a list (either as free morpheme, stem, or compound constituent). The lists were also examined by native Japanese graduate students to eliminate intra-list associations. In the case of the first survey, 20 lists were prepared in this way, while a total of 36 lists were similarly created for the second survey.

In addition to a cover sheet of instructions, a survey questionnaire consisted of 10 pages with 10 items printed per page as a centered column of words with underlined blank spaces for the association responses (e.g., 本 __________). The instructions asked the participants to look at each printed item and to write down in the blank space the first semantically-related Japanese word that comes to mind. The instructions also included some directions relating to aspects of the Japanese writing system.

2.2.2. Results and data coding

In two traditional paper questionnaire surveys, approximately 148,600 word association responses for a corpus of 5,000 basic Japanese kanji and words were collected from 1,486 native Japanese speakers.

The word association responses obtained through the surveys have been entered into a database and coded for error responses. Blanks that were clearly due to a respondent skipping a page or failing to complete a questionnaire are treated as non-presented items, but other cases are being recorded. Illegible and nonword responses are also treated as non-presented items, but minor writing errors have been corrected.

Through two questionnaire surveys, 2,100 items drawn at random from the initial survey corpus were presented to up to 50 respondents for word association responses. The responses to these items are being processed to form the first version of the Japanese word association database, in order to make this data publicly available. A list of the 2,100 items, with respondent counts and associate set sizes, is available at http://www.valdes.titech.ac.jp/~terry/jwad.html.

2.3. Future development of the database

While the data collected through the two surveys represent a substantial initial stage in constructing the large-scale database, the traditional paper questionnaire format involves considerable burdens in terms of preparation and data inputting. Accordingly, the project is also developing a computer-based version of the word association survey, with the aim of conducting the survey over the Internet to collect large-scale quantities of responses for the database.

The preparation for computer-based and web-based versions of the word association has involved the creation of two programs. The first program is a survey list generation program that will automatically create a survey list of N items (normally, 100 items) from the survey corpus, while running checks on the coded information to eliminate intra-list associations. The main bottleneck in developing this was not the program itself, but the preparation of the coded information for the entire corpus, particularly the semantic category codes and the feedback data. The second program is a presentation program to present the items on the computer screen one at a time and to save the item and the response as a paired set of data. Data collection with the web-based version of the survey is expected to start early in the 2006 academic year.

3. Applications of the database

After noting how the database can be used to create lexical association network maps, this section briefly outlines some interesting application in the areas of mental lexicon research, as well as Japanese lexicography and language instruction.

3.1. Lexical association network maps

A central objective of the mapping lexical knowledge project is to utilize the database of Japanese word associations to develop lexical association network maps, as a means of capturing and highlighting the patterns of associations that exist between Japanese words.

The basic component of the maps is the set of associates given in response to a target word and the strengths of those associates indicated in terms of response frequency.

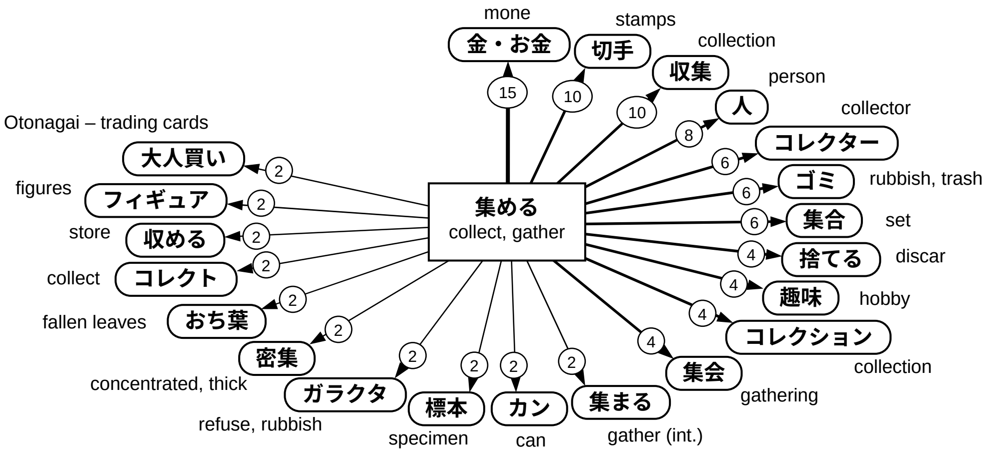

Figure 2: The associate set for 集める 'collect, gather'.

Figs. 1 and 2 show the associate sets for the words 冬 'winter' and 集める 'collect, gather', respectively. As Nation points out [23], one crucial aspect of lexical knowledge is knowing about the words a particular word is associated with. Fig. 1 shows that the primary associate of 冬 'winter' is 寒い・さむい 'cold', which accounts for 44% of the responses, while the association strengths of other associates are comparatively much lower, such as 雪 'snow' (15%), 夏 'summer' (6%) and 冬至 'winter solstice' (6%). In contrast, to the strong association between the adjective 寒い・さむい 'cold' and the noun 冬 'winter', many of the associates of the verb 集める 'gather, collect', as shown in Fig. 2, have the relationship of direct object, such as お金・金 'money' (15%), 切手 'stamps' (10%), 人 'people' (8%) and ゴミ 'rubbish, trash' (6%). However, with the larger set of core associates (that is, responses given by more than one respondent) for 集める 'gather', the frequency of its primary associate (お金・金 'money') is much lower than the frequency of the primary associate for 冬 'winter'.

The basic associate set is defined by the forward association relationship between a target word and its associates. The lexical association network maps will also feature backward association both in terms of numbers and strengths. The backward association data is very important because association is clearly not a symmetrical feature. The third vital aspect of the lexical association network maps is the representation of associate density, that is, the numbers and strengths of associations between all the words within a particular association set.

While the lexical association network maps are primarily at the single-word level, they can obviously be combined to build the kind of global semantic network that Steyvers and Tenenbaum [10] created based on word association data and compared with a network based on WordNet. As Steyvers and Tenenbaum speculate, the similarities that they observed between these networks would seem to be due to pervasive and deep features of semantic knowledge.

3.2. Mental lexicon research

In a different approach to investigating the organization of lexical information within the Japanese mental lexicon, the author is also conducting visual word recognition research [24][25][17][18]. A series of constituent-morpheme priming experiments has provided results that clearly indicate that morphology plays an important role in the organization of the mental lexicon. For instance, in experiments that contrasted the positional frequency of the verbal constituents in verb + complement and complement + verb compound words, a reversed pattern of priming was observed in the high positional frequency conditions, where the priming from verbal constituents was significantly greater than from the complement conditions. The different patterns of priming for various word-formation principle conditions across very brief prime presentation conditions of 60, 90, 150, and 250 in a recent study [17] also suggest that the verbal constituents of two-kanji compound words may be more effective in activating a morphological family. However, this line of research also needs to address the issue of association effects between the prime and target stimuli, and will be able to use the Japanese word association database to control for word association strengths in further examining the role of morphology in the Japanese mental lexicon.

The lexical association network maps also represent an extremely promising approach to developing the semantic aspects of the Japanese lemma-unit model [24][25]—a connectionist model of the Japanese mental lexicon adopted to account for the results from the constituent-morpheme priming experiments—by incorporating the lexical association network maps within the model.

3.3. Japanese lexicography and language instruction

There are also direct Japanese lexicographical and instruction applications of the database and the maps [13], which the project plans to address in more detail during the 2006 academic year.

There are two ways in which the database and the maps could be applied to Japanese lexicography. Firstly, the inclusion of core associates, together with phrase patterns where appropriate, would enrich the variety of information provided in dictionaries. Similar to the kind of collocation data that Terashima and Moriguchi [26] argue is so important for learners, incorporating word association would be especially useful feature for Japanese learner dictionaries. Secondly, the database and the maps could be used to enhance electronic dictionaries in supporting user-friendly look-up functions [27]. The basic notion is that, although conventional form-based entry searching is not possible in the fairly common situation of the tip-of-the-tongue phenomenon, a user would be able to use available related information to search along association connections to locate the target word if the lexical association network maps were incorporated within the dictionary.

Turning to the Japanese language instruction applications, although memory research has long demonstrated that the categorization and semantic organization of stimulus materials dramatically influences retrieval performance [28], Tinkham [29] has argued that for foreign vocabulary learning, because interference effects can occur when simultaneously learning sets of L1-L2 word pairs that are semantically-related, thematic associations may be more effective. The effects of semantic clustering based on themes and associations have been demonstrated in learning Spanish as a second language [30], while Tokuhiro [31] has reported effects of using 'conceptual map' as a vocabulary learning technique for Japanese. These studies suggest that the lexical association network maps for basic Japanese vocabulary being developed within this project can be very helpful in creating effective vocabulary strategies for Japanese language instruction.

In summary, then, this paper has outlined the construction and utilization of a large-scale database of Japanese word associations.

4. Acknowledgements

The author would like to thank Prof. Furui, Prof. Tokosumi, Prof. Nishina and members of the LKR COE for their support to this research.

5. References

[1] Firth, J. R., *Selected papers of J. R. Firth 1952-1959.* (Edited by F. R. Palmer). Longman, London, 1957/1968.

[2] Church, K. W., and Hanks, P., "Word association norms, mutual information, and lexicography", *Computational Linguistics,* Vol. 16, 1990, pp. 22-29.

[3] Cantos, P., and Sánchez, A., "Lexical constellations: What collocates fail to tell", *Int. J. Corpus Linguistics,* Vol. 6, 2001, pp. 199-228.

[4] Hirst, G., "Ontology and the lexicon", In S. Staab, and R. Studer, (Eds.), *Handbook of ontologies,* Berlin, Heidelberg, and New York: Springer-Verlag, 2004.

[5] Deese, J., *The structure of associations in language and thought,* Baltimore, The John Hopkins Press, 1965.

[6] Cramer, P., *Word association,* New York and London, Academic Press, 1968.

[7] 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.

[8] Nelson, D. L., and McEvoy, C. L., "Implicitly activated memories: The missing links of remembering". In C. Izawa, and N. Ohta, (Eds.), *Human learning and memory: Advances in theory and application*, Mahwah, Lawrence Erlbaum Associates, 2005.

[9] Steyvers, M., Shiffrin, R. M., and Nelson, D. L., "Word association spaces for predicting semantic similarity effects in episodic memory". In A. F. Healy, (Ed.), *Experimental cognitive psychology and its applications,* Washington: American Psychological Association, 2004.

[10] Steyvers, M., & Tenenbaum, J. B., "The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth", *Cognitive Science*, Vol. 29, pp. 41-78, 2005.

[11] Joyce, T., "Mapping word knowledge for basic Japanese vocabulary", *Symposium on Large-Scale Knowledge Resources (LKR2005)*, Tokyo Institute of Technology, pp. 29-32, 2005.

[12] Joyce, T., "Building a word association database for basic Japanese vocabulary" (in Japanese), *Proceedings of the 3rd Annual Meeting of the Japanese Society for Cognitive Psychology,* Kanazawa University, Kanazawa, Japan, p. 70, 2005.

[13] Joyce, T., "Lexical association network maps for basic Japanese vocabulary", In Ooi, V. B. Y., Pakir, A., Talib, I., Tan, L., Tan, P. K. W., and Tan, Y. Y., (Eds.). *Words in Asia cultural contexts.* Singapore: National University of Singapore. pp. 114-120, 2005.

[14] Joyce, T., "Constructing a large-scale database of word associations", (in Japanese), *Proceedings of the 69th Meeting of the Japanese Psychological Association*, Keio University, Tokyo, Japan, p. 629, 2005.

[15] Joyce, T. "Constructing a large-scale database of Japanese word associations", (Special issue on kanji corpora research edited by Katsuo Tamaoka), *Glottometrics*, Vol. 10, (in press).

[16] Joyce, T., "Two-kanji compound words in the Japanese mental lexicon", Invited presentation *6th International Forum on Language, Brain, and Cognition (Cognitive Psychology of East Asian Languages: Cognitive Studies and their Application to Second Language Acquisition),* Tohoku University, Sendai, Japan, 3-4 December 2005.

[17] Joyce, T., and Masuda, H., "Brief-presentation constituent-morpheme priming effects on the processing of Japanese two-kanji compound words," *The 11th International Conference on Processing Chinese and Other East Asian Languages (PCOEAL 2005)*, Chinese University of Hong Kong, Hong Kong, 2005.

[18] Masuda, H., and Joyce, T., "A database of two-kanji compound words featuring morphological family, morphological structure, and semantic category data," (Special issue on kanji corpora research edited by Katsuo Tamaoka), *Glottometrics*, Vol. 10, in press.

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

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

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

[22] National Institute for Japanese Language. *Word list by semantic principles,* (Revised ed., in Japanese), Tokyo: Dainihon Tosho, 2004.

[23] Nation, I. S. P. *Teaching and learning vocabulary,* New York: Newbury House, 1990.

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

[25] 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.

[26] Terashima, K., and Moriguchi, M., "The first collocation dictionary for the Japanese language", In Ooi, V. B. Y., Pakir, A., Talib, I., Tan, L., Tan, P. K. W., and Tan, Y. Y., (Eds.). *Words in Asia cultural contexts.* Singapore: National University of Singapore. pp. 303-307, 2005.

[27] Zock, M., and Bilac, S. "Word lookup on the basis of associations: From an idea to a roadmap." *COLING2004 Workshop on Enhancing and using electronic dictionaries*, August, Geneva, 2004.

[28] Bower, G. H., Clark, M. C., Winzenz, D., and Lesgold, A., "Hierarchical retrieval schemes in recall of categorized word lists", *J. of Verbal Learning and Verbal Behavior*, Vol. 8, 1969, pp. 323-343.

[29] Tinkham, T., "The effects of semantic and thematic clustering on the learning of second language vocabulary", Second Language Res., Vol. 13, 1997, pp. 138–163.

[30] Morin, R., & Goebel, J., Jr. "Basic vocabulary instruction: Teaching strategies or teaching words?" *Foreign Language Annals*, Vol. 34, 2001, pp. 8-17.

[31] Tokuhiro, Y. Kanji ninchi shori kara mita kōkateki kanji shūtokuhō no kenkyū: Sōgoketsugōgata gainen chizu sakusei no kokoromi, Waseda Daigaku Nihongo Kyōiku Kenkyū, Vol. 2, 2003, pp. 151-176.