Constructing a Japanese Word Association Database

Terry Joyce (Tama University)

This paper reports on a project investigating lexical knowledge by mapping out the associative structures that exist for Japanese words. Specifically, the paper briefly outlines (1) the construction of the large-scale Japanese Word Association Database (JWAD), (2) the development of lexical association network maps, as a means of capturing association patterns, based on the JWAD, and (3) promising applications of the database and the maps. An example of a lexical association network map contrasting a small set of emotional words is presented to illustrate their potential in highlighting association structures and providing interesting insights into lexical knowledge.

1 Introduction

Association is a basic mechanism of human cognition. Inspired by that simple notion, a considerable amount of cognitive science research, particularly linguistic and psycholinguistic research, has sought to identify and understand the structured relations that exist between concepts by mapping out how concepts are represented in the rich networks of associations that exist between words (Cramer, 1968; Deese, 1965; Hirst, 2004; Moss & Older, 1996; Nelson & McEvoy, 2005; Okamoto & Ishizaki, 2001; Steyvers, Shiffrin, & Nelson, 2004; Steyvers & Tenenbaum, 2005; Umemoto, 1969).

This paper reports on a project seeking to elucidate fundamental aspects of lexical knowledge by mapping out the patterns of associative connections that exist for Japanese words. In particular, the paper describes (1) the construction of the large-scale Japanese Word Association Database (JWAD), (2) the use of the JWAD in developing lexical association network maps as a way of highlighting association patterns, and (3) some promising applications of the database and the maps.

2 Construction of JWAD

2.1 Existing word association databases

Although large word association databases exist for English (i.e., Moss & Older, (1996); Nelson, McEvoy, & Schreiber (1997)), databases of Japanese word associations have been comparatively scarce. Notable exceptions include the early, well-known survey conducted by Umemoto (1969), which gathered responses from 1,000 university students but only covered a very small set of 210 words, and, more recently, the association data for 1,656 nouns collected by Okamoto and Ishizaki (2001). However, a major drawback with the latter database, apart from only covering nouns, is the fact that response category was specified as part of the word association task, so it tells us little about free associations.

2.2 Version 1 of the JWAD

2.2.1 Questionnaire surveys

After compiling a survey corpus of 5,000 basic Japanese kanji and words, construction of the JWAD started with two large-scale questionnaire surveys. The first survey sought to collect up to 50 responses for a random sample of 2,000 items, while the second survey collected at least ten responses for the remaining 3,000 items.

2.2.2 Method

Participants: Native Japanese students attending the University of Tsukuba (N = 1,481; 929 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, target items were divided into lists of 100 items, and a page of the survey questionnaire consisted of 10 items as a centered column of words with underlined blank spaces for 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.

Results: In total, approximately 148,100 word association responses were collected. Through the two surveys, a random sample of 2,099 items was presented to up to 50 respondents for word association responses.

2.2.3 Coding of word association responses in JWAD-V1

The word association responses to the 2,099 items have been coded and processed together as version 1 of the JWAD (requests for JWAD-V1 may be directed to the author). Two levels of codes are applied to the database. The level 1 codes classify responses at a general level in terms of their appropriateness distinguishing between semantic associations (i.e., 耕す 'plow, cultivate' eliciting 畑 'field'), orthographic associations (i.e., 有様 'condition, state' eliciting 殿様 '(feudal) lord') and phonological associations (i.e., しまう /shimau/ 'to put away or finish' eliciting しまうま /shimauma/ 'zebra'). Another set of codes cover kinds of transcription responses, where the response word is essentially an orthographic variant of the item (i.e., 泣く 'weep, cry' for the homophone なく). Isolated blank responses are also recorded at this level as an index of words that do not easily elicit association responses. Level 2 codes attempt to provide additional information, such as marking foreign word responses (i.e., 謝る 'apologize' eliciting 'sorry'), verb conversion (i.e., 考慮 'consideration' eliciting 考慮する 'consider'), and proper nouns (i.e., 意識 'consciousness' eliciting フロイト 'Freud').

2.3 Web-based survey and future expansions to JWAD

In order to collect large-scale quantities of association responses, the project has also developed a web-based version of the word association survey (http://nerva.dp.hum.titech.ac.jp/terry/index.jsp). JWAD-V2 will be released once at least 50 association responses have been collected and coded for all 5,000 items in the present survey corpus. The survey corpus will shortly be expanded considerably, in order to further examine the asymmetrical nature of word associations.

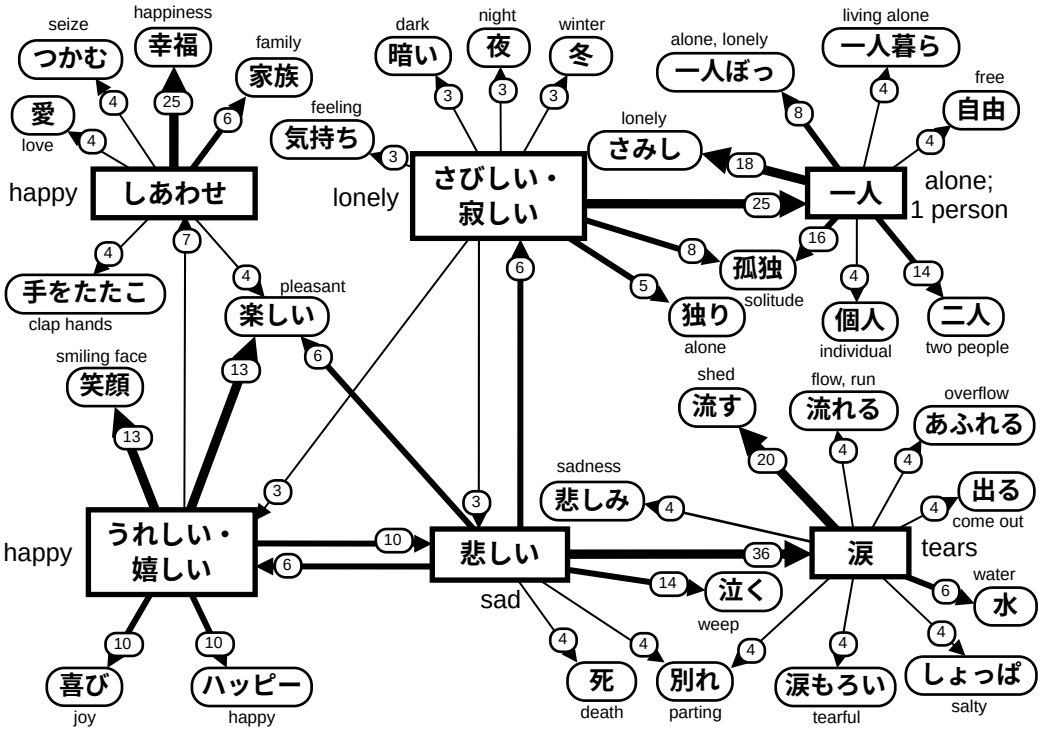

Figure 1. Example of lexical association network map building from and contrasting a set of emotion words. *Note*: The numbers on the arrows indicate response frequency as percentages for a particular association set.

3. Lexical association network maps

A central objective of the project is to utilize the JWAD in developing lexical association network maps as an approach to the visualization of lexical knowledge. The basic concept of the maps is to represent the set of forward associations evoked by an item (i.e., set size and response frequencies as index of association strength), together with backward associations from those associates to the item, as well as association connections among all set constituents. However, as Figure 1 illustrates, single-word level maps can also be combined to create semantic networks for various domains.

Even such a small map can clearly illustrate how related words can have different patterns of association. For while the positive synonymous words of しあわせ and うれしい・嬉しい, meaning 'happy', have rather strong associations to a small set of close synonyms, such as 幸福 'happiness' and ハッピー 'happy', interestingly, the negative emotion words of さびしい・寂しい 'lonely' and 悲しい 'sad' primarily elicit word association responses that can be regarded as having a causal or resultant relationship. For example, 一人 'alone; 1 person', 孤独 'solitude' and 独り 'alone' are strong associates of さびしい・寂しい, while 悲しい has a particularly strong prime association of 涙 'tears' (36%) followed by 泣く 'weep' (14%).

In a complementary approach to discerning the patterns of connectivity within the JWAD, Joyce and Miyake (2007) have applied graph clustering techniques to a semantic network representation of the JWAD. Graph theory analysis of the JWAD network indicates that it has scale-free characteristics.

Conceptually somewhat similar to combining related association maps, graph clustering techniques can be a very useful tool for automatically identifying wider groups of related words. For instance, applying Markov clustering to the JWAD network yields the word groups of {喜, 喜び, 喜ぶ, 喜寿, 歓喜, 大喜利, 喜怒哀楽, 悲しむ, 怒} for うれしい・嬉しい and {一人・1 人, 独り, 一人ぼっち, 孤独, 独身, 独身貴族, 未婚, さみしい, 二人} for さびしい. Such results underscore the potential of graph clustering techniques to automatically construct hierarchically-organized semantic spaces as an approach to the visualization of large-scale linguistic knowledge resources.

4. Applications of the JWAD and lexical association maps

Finally, the project is also exploring a number of applications of the JWAD and the lexical association network maps. In the area of lexicography, for instance, the incorporation of word association data into Japanese learner dictionaries in the form of core associates, together with phrase patterns where appropriate, would enrich the variety of information provided and be especially useful for Japanese language learners. The inclusion of associations and maps could also be used to enhance electronic dictionaries in supporting user-friendly look-up functions (Zock & Bilic, 2004).

Another application area is in Japanese language instruction, and Joyce, Takano, and Nishina (2006) have conducted a study to investigate the use of bilingual lexical maps as an instruction strategy for specialist vocabulary. Their results indicate that emphasizing semantic and thematic relationships within target L2 vocabulary through the spatial organization of concepts in the form of a bilingual lexical map can be useful in aiding the study of specialist vocabulary.

References

Cramer, P. (1968). *Word association.* New York and London: Academic Press.

Deese, J. (1965). *The structure of associations in language and thought.* Baltimore: The John Hopkins Press.

Hirst, G. (2004). Ontology and the lexicon. In S. Staab, & R. Studer, (Eds.), *Handbook of ontologies*. (pp. 209-229). Berlin, Heidelberg, & New York: Springer-Verlag.

Joyce, T. (2005) "Constructing a large-scale database of Japanese word associations", In Tamaoka, K. (Ed.). *Corpus Studies on Japanese Kanji*. (Glottometrics 10). pp. 82-98. Hituzi Syobo: Tokyo, Japan and RAM-Verlag: Lüdenschied, Germany.

Joyce, T., & Miyake, M. (2007). Gurafukurasutaringu ni yoru rensōgo no imi nettowāku no bunseki. *The 5th Annual Meeting of the Japanese Society for Cognitive Psychology*, Kyoto University, Japan, 76.

Joyce, T., Takano, T., & Nishina, K. (2006). "Senmongo no gakushū hōhō toshite no bairingaru goi map, *The 4th Annual Conference of the Japanese Society of Cognitive Psychology*, Chukyo University, Japan, 201.

Moss, H., & Older, L. (1996). *Birkbeck word association norms*, Hove, UK: Psychological Press.

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

Nelson, D L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. Retrieved May 31, 2007, from http://w3.usf.edu/FreeAssociation/.

Okamoto, J. & Ishizaki, S. (2001). Associative concept dictionary and its comparison electronic concept dictionaries, *PACLING2001*, 214-220.

Steyvers, M., Shiffrin, R. M., and Nelson, D. L. (2204). "Word association spaces for predicting semantic similarity effects in episodic memory". In A. F. Healy, (Ed.), *Experimental cognitive psychology and its applications,* (pp. 237-249. Washington: American Psychological Association.

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

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

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