Glottometrics 10, 2005, 82-98

Constructing a Large-Scale Database of Japanese Word Associations[1]

Terry Joyce[2]

Tokyo Institute of Technology, Japan

Abstract. For cognitive scientists investigating the nature of lexical knowledge, one essential task is to map out the rich networks of associations that exist between words. This paper reports on a project to construct a large-scale database of word association norms for basic Japanese vocabulary and, utilizing the database, to develop lexical association network maps that tap into important aspects of words and their connectivity. The Japanese word association database will complement existing databases concerning the lexical features of Japanese vocabulary, such as familiarity ratings and frequency counts (Amano & Kondo, 1999; Yokoyama, Sasahara, Nozaki & Long, 1998), and the kanji corpus research highlighted in this special issue. Part 2 of this paper outlines the construction of the database, by detailing initial collections of word association responses from two major questionnaire surveys and the current state of the database. Part 3 introduces the lexical association network maps that will be developed based on the word association norm data and discuses some particularly promising applications of the database and the network maps in the areas of cognitive science and Japanese lexicography and language instruction.

Keywords: Japanese word association, large-scale database, lexical association network maps

1. Introduction

Given its crucial importance for many areas of cognitive science, such as psychology, artificial intelligence, computational linguistics and natural language processing, much research has, understandably, been devoted to investigating the nature of lexical knowledge and, in particular, to mapping out the rich networks of associations that exist between words. The interest in word associations is motivated by a number of converging perspectives and

[1] This research project 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. The author would like to express his gratitude for help and support from Yokoyama Shōichi and Wada Yuriko at the National Institute for Japanese Language, Jim Elwood at the University of Tsukuba, as well as Tokosumi Akifumi, Matsumoto Naoko and Murai Hajime at Tokyo Institute of Technology.

[2] Address correspondence to: Terry Joyce, Ph.D., Tokyo Institute of Technology, W9-29, 2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552. Email: terry@valdes.titech.ac.jp

concerns within cognitive science. Central among these is the insight that, because association is a fundamental mechanism underlying human cognition, word associations mirror rather closely the structured patterns of relations that exist among concepts (Cramer, 1968; Deese, 1965). This notion is consistent with a number of influential assertions and inspirations within natural language processing research, such as Firth's (1957/1968) claim that a word's meaning resides in the company it keeps, Church and Hanks' (1990) notion of mutual information as a measure of the saliency of an association between two words, and Hirst's (2004) acknowledgement, notwithstanding certain caveats on the complex relationship between them, that a lexicon can often be a useful basis for developing a practical ontology. Lexical networks, whether represented with lexical nodes (i.e., in the tradition of Collins and Loftus, 1975) or as fully distributed features (e.g., Rumelhart, McClelland, and the PDP Research Group, 1986) are also at the heart of many connectionist models of human cognition, which, together with the key notion of spreading activation, derive much of their appeal from their neurological plausibility.

Recently, a number of studies highlight the valuable contributions that word association normative data can make to cognitive science (Nelson & McEvoy, 2005; Steyvers, Shiffrin, & Nelson, 2004; Steyvers & Tenenbaum, 2005). These studies all utilize *The University of South Florida word association, rhyme, and word fragment norms* (Nelson, McEvoy, & Schreiber, 1998), which is the largest database of word associations for American English, covering over 5,000 words with an average of 149 responses (SD = 15) per word collected from more than 6,000 participants. Demonstrating the influence of existing word associations on cognition, Nelson and McEvoy (2005) show that differences in the associative structures of known words—in terms of associate set size, resonance (or backward association), and the connectivity within an associate set—effect performance on the episodic memory task of extra-list cued recall. Steyvers, Shiffrin, and Nelson (2005) apply scaling techniques, such as singular value decomposition (SVD) and multidimensional scaling (MDS), to the word association norms database to create a word association space (WAS) representing the semantic (dis)similarity between pairs of words. Comparing three WAS-based measures with two latent semantic analysis (LSA) based measures (i.e., based on corpus collocations) in terms of predicting performance on three types of episodic memory tasks (recognition, free recall, and cued recall), they found that the WAS-based measures were better predictors of performance than the LSA-based measures for all three memory tasks. Steyvers and Tenenbaum (2005) also employ the database of word association norms as part of their analyses of the structures of large-scale semantic networks. Specifically, they used graph theory to analyze three semantic networks—one based on the word associations, one based on WordNet (Fellbaum, 1998), and another based on Roget's thesaurus. Steyvers and Tenenbaum found that all three semantic networks have statistical features in common, which they characterized as being small-world—having sparse connectivity, short average path lengths between words, and strong local clustering—and scale-free structures—most nodes have relatively few connections but are joined together via a small number of hubs with many connections.

This paper reports on a project to (1) construct a large-scale database of word association norms for basic Japanese vocabulary, to (2) utilize the word association norm data in creating and developing lexical association network maps, which capture important properties of words and their connectivity, and to (3) explore applications of the word association norms in the areas of cognitive science and Japanese lexicography and language instruction. Part 2 of the paper details the compilation of an initial survey corpus of basic Japanese vocabulary and initial collections of word association responses using a traditional questionnaire format. After outlining the current state of the database, Part 2 also notes plans for the future development of the database, including further collections of word association responses with computer-based and Internet-based versions of the survey, currently being developed, and for expanding the survey corpus. Part 3 of the paper starts by introducing the lexical association network maps that will be created from the database of word association norms. Part 3 also discusses some applications of the word association database and the network maps in the areas of cognitive science, such as in experimental control and as an approach to modeling the semantic representations of connectionist models, and of Japanese lexicography and language instruction, such as enriching the variety of lexical information within the lexical entry and providing user-friendly look-up functions.

2. Constructing the Database of Japanese Word Associations

2.1. Existing Word Association Norms

As it would be beyond the scope of this paper to attempt a review of word association norms (see Cramer, 1968; Deese, 1965; Moss & Older, 1996; Nelson, McEvoy, & Schreiber, 1998), the aim of this section is merely to briefly mention a couple of databases of word association norms for English and Japanese as frames of reference regarding the scale of the present project.

One large database of word association norms for British English has been created by Moss and Older (1996), which covers some 2,400 words with between 41-50 responses to each item. However, as already noted, the largest database of word association norms for American English is that constructed by Nelson, McEvoy, and Schreiber (1998), covering more that 5,000 words with an average of 149 responses for each item. It should be noted, however, that both these databases of association norms are the products of combining a number of surveys conducted over quite a number of years and that, rather than being systematic attempts to construct comprehensive databases, the inclusion of words in the surveys was usually in response to more immediate experimental interests at the time.

The first Japanese word association norms to mention are those collected in an early survey by Umemoto (1969). Although he gathered response from 1,000 university students, the word corpus is very small with only 210 words and thus of extremely limited value in controlling for the associative strength between stimulus items in experiments. More recently, Ishizaki (2004) has collected word associations as part of a project to build an associative concept dictionary (Okamoto & Ishizaki, 2001). Ishizaki's data covers 1,656 nouns with 10 responses for each item.[3] While arguably consistent with the aim of building an associative concept dictionary, a major drawback with this data, however, is the fact that response category was specified. Participants were asked to respond to a presented stimulus word according to one of seven randomly presented categories (hypernym, hyponym, part/material, attribute, synonym, action and environment), so the data tells us little about free associations.

[3] While this response count relates to version 1.0 made publicly available in March 2004, it seems that another version with 50 responses per item also exists.

2.2. Compiling a Survey Corpus of Basic Japanese Vocabulary

In order to compile an initial corpus of basic Japanese vocabulary for the word association survey, three reference sources were used. The first was the survey of basic vocabulary for Japanese language teaching conducted by the National Language Research Institute (1984). This list consists of approximately 6,800 words including a core set of about 2,200 words. The second reference source was Tamamura (2003), which is a recently prepared list of intermediate vocabulary of about 4,000 words. Because of its influence on Japanese language education, an important standard to look at when considering what constitutes basic Japanese vocabulary is the sanctioned list of Jōyō kanji. Accordingly, the third reference source was a handbook of Japanese orthography (Sanseidō Henshūjo, 1991), which lists all 1,945 Jōyō kanji with their official readings as well as a number of compound word examples (in total about 13,000 word tokens).

Once these lists were input, they were compared in order to identify common words, with priority on the overlap between the first two sources and particularly the core set of approximately 2,200 words within the National Language Research Institute's (1984) list. The task was made somewhat more difficult by the fact that Tamamura's (2003) intermediate vocabulary list has many words transcribed in hiragana that are transcribed in kanji in the National Language Research Institute's (1984) list. Reflecting the flexible nature of Japanese orthography and shifts in orthographic conventions over the last 20 years or so, the transcription differences highlight the merit of including orthographic variants within the survey. Related to that and the high incidence of homophones in Japanese, hiragana transcription words were frequently included for homophone sets within the corpus. For example, in the case of the homophone set of 合う 'to fit, suit, match', 会う 'to meet', and 遭う 'to meet, encounter (undesirable nuance)' sharing the pronunciation /au/, the hiragana transcription あう was also included. Also in an exploratory vein, a number of bound morpheme kanji were included. These include affixes, such as 不 /fu/ 'non-, un-', and verbal and adjectival stems, such as 書く /kaku/ 'to write' without the okurigana く /ku/, which are normally written with other kanji or okurigana endings and, in the strictest sense, are not words when written alone. Based on this work, an initial survey corpus of 5,000 kanji and words was created.

2.3. Questionnaire Surveys

In order to obtain the large-scale quantities of responses that will be required to complete the construction of the word association database, a computer-based version of the word association survey is being developed, so that the survey can also be conducted over the Internet. However, construction of the database is already progressing based on two surveys conducted using traditional pen-and-paper questionnaires. The first survey was conducted with the aim of obtaining up to 50 word association responses for a random sample of 2,000 items drawn from the survey corpus. Those responses will later be used to examine the consistency and reliability of word association responses to be collected with different formats of the survey, particularly those to be obtained from volunteer respondents participating in the survey via the Internet. The aim of the second survey was to obtain up to ten responses for the remaining 3,000 items in the corpus. Those responses will be used to control for intra-list association when respondent survey lists are generated automatically (as discussed in more detail in Section 1.4). Although the two surveys were conducted with different secondary aims, because the primary objective of obtaining word association responses in the construction of the database was common to both, and because the basic procedures were the same, they are outlined together.

2.3.1. Method

Participants. Native Japanese university students (N = 1,486; 934 males and 552 females; average age 19.03, SD = 0.97) participated in the surveys as volunteers.

Survey lists. For the first survey, 2,000 items were randomly drawn from the corpus and these were divided into 20 lists of 100 items. These items were divided so that each list consisted of a mixture of orthographic forms (i.e., single kanji, multi-kanji, and mixed kanji-kana words) in ratios closely matching the distribution within the overall corpus. Care was also taken to avoid intra-list associations, by ensuring that no two items within a list shared the same pronunciation and that no given kanji appeared more than once in a list either alone or as a constituent of a polymorphemic word. Finally, each list was examined by native Japanese graduate students so that all possible intra-list associations were eliminated. In order to obtain up to 50 word association responses for the 2,000 items, each survey list was presented to 50 respondents, but with the order of the items being randomized for each individual respondent. For the second survey, the remaining 3,000 items in the corpus were divided among 36 lists of 100 items. Care was again taken to control for the mixture of orthographic forms, incidences of homophones, and multiple inclusions of any given kanji. By the time the second survey was being prepared, the survey corpus had been coded with semantic category information (discussed further in Section 1.4 below). Thus, the lists for the second survey were created by also checking to ensure that no two items belonged to the same semantic category. Because of this extra control, it proved necessary to increase the number of survey list to 36 in order to cover some semantic categories with more member items. This also meant that some of the remaining 3,000 items appeared in two lists in the second survey. In order to obtain up to ten word association responses for the 3,000 items, each survey list was presented to at least 10 respondents, with the order of items within each respondent list randomized. In the event, more participants were available than minimally required for the secondary objective, so two of these lists were presented to 50 respondents, while another two were presented to 30 and 33 respondents respectively.

Each respondent list was printed with 10 items per page—the items were printed in 18pt Mincho beside an underlined blank space for the response (e.g., 本 __________) in a row centered on the page—forming a booklet of 10 pages plus a cover sheet with instructions. 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. There were also instructions relating to aspects of the Japanese writing system. The first of these asked the participants to respond with what they considered to be the most natural orthographic representation of the associate response (i.e., whether they would normally write /manga/ in kanji as 漫画, in hiragana as まんが, or in katakana as マンガ). Another instruction asked participants not to change their response to another word if they found that they could not remember the correct strokes for the kanji of their first response, but to indicate that they were not confident of the correct strokes by providing the word’s pronunciation in a hiragana gloss above the word.

Table 1
A Random Sample of 10 Items from the List of 2,100 Items in the Japanese Word Association Database (Version 1.0), with Respondent Counts, and Associate Set and Core Associate Set Sizes

Item	Respondents	Associate Set	Core Associate Set
遅らす	50	18	6
貝	50	23	5
耕す	50	14	5
引っ越す	50	34	6
公平	50	26	6
うらやましい	50	35	7
安心	50	32	8
ベンチ	50	18	6
触れる	50	26	4
最新	50	30	9

Note: Core associate set refers to the number of responses provided by two or more respondents.

2.3.2. Error Response Coding and Results

The word association responses collected with the paper questionnaires have been entered into a database by native Japanese graduate students. Blank spaces (no responses) were treated in two ways; in cases where a whole page had been skipped or where the participant failed to complete the questionnaire sheets, the items were regarded as having not been presented and accordingly are not reflected in the respondent counts, otherwise blank responses were recorded and, for the present, these are included as part of the set of word association responses for an item, as an indicator of words that are more difficult to make word association responses to. Items for which the response was illegible or involved a kanji selection error that resulted in an uninterpretable nonword were also treated as not presented items. When the response involved a minor writing mistake, such as incorrect kanji strokes or component element, but the intended response was clear from the presented word, the error was corrected. Responses based on phonological associations and transcription responses (i.e., where the respondent either provided the pronunciation in kana of a kanji orthography item or, more frequently, where the response to a kana orthography word was a kanji orthography word sharing that pronunciation) are currently being recorded and marked accordingly. Although the transcription responses could be indicating the need for more explicit instructions ruling out orthographic variants as invalid responses, it is also possible that the Japanese respondents regarded the orthographic variants as independent words. In cases of phrasal responses consisting of the presented item plus only one other word (excluding appropriate case markers), that word was taken as the response (i.e., when in response to 亡くなる 'pass away', one participant wrote おじいさんが亡くなった 'my grandfather passed away', おじいさん 'grandfather' was recorded as the response).

Table 2
Two Examples of Word Association Response Data in the Japanese Word Association Database (Version 1.0)

Item	Responses	Number	Item	Responses	Number
主語	述語	34	沸く	湯・お湯	35
	私・私は・わたし	4		沸騰	3
	動詞	2		水	3
	S (subject)	1		害虫	1
	が	1		歓声	1
	自分	1		敵	1
	修飾語	1		点	1
	誰か	1		電気	1
	使う	1		鍋	1
	名前	1		風呂	1
	話す	1		やかん	1
	人	1		ワールドカップ	1
	名詞	1			

Through two questionnaire surveys, 2,100 items randomly sampled from a survey corpus of 5,000 basic Japanese kanji and words were presented to up to 50 respondents. The responses to the 2,100 items have been processed to form the first version of the Japanese word association database, which is being made publicly available.[4] As illustrated with a random sample of 10 items in Table 1, a list of the 2,100 items together with respondent counts and the sizes of the associate set and core associate set (referring to the number of responses provided by two or more respondents) is available for download at http://www.valdes.titech.ac.jp/~terry/jwad.html. As the two examples presented in Table 2 show, the Japanese word association database lists all word association responses collected for the 2,100 items presented to up to 50 respondents. The associate set is ordered with the prime associate listed first. In the case of the two examples, the prime associate of 主語 'subject' is 述語 'predicate', given by 34 respondents, while the prime associate for 沸く 'boil; get hot; get excited' is 湯・お湯 'hot water', given by 35 respondents. As more word association responses are collected for all items in the survey corpus, and the pattern of associations for each item becomes more stable, consistent with Nelson, McEvoy, and Schreiber (1998), the database will focus on the core associate sets, but responses provided by only one respondent are included in the present version of the database.

As Nation (1990) observes, in addition to knowing a word's spoken and written forms, its grammatical and collocation behavior, its frequency and stylistic register, as well as its

[4] Requests for the Japanese word association database (Version 1.0) may be directed via email to the author.

conceptual meaning, one important aspect of lexical knowledge is knowing about the associations that a word has with other words. Figure 1 presents the associate set for the Japanese word 冬 'winter' based on the word association responses collected so far. The enclosed figures on the arrow connections represent the percentage of responses. As the figure shows, 冬 'winter' has a very strong primary associate with the word 寒い・さむい 'cold', which accounts for 44 percent of all responses. The second associate of 雪 'snow' represents only 15 percent of the responses, followed by 夏 'summer' and 冬至 'winter solstice', both at 6 percent, and 白・白い 'white' at 4 percent. Thus, 冬 'winter' has a relatively small set of core associates with one particularly strong associate.

In contrast, Figure 2 presents the associate set for the Japanese verb 集める 'gather, collect', which has a larger set of core associates, but, naturally, with weaker association strengths. The primary associate here is お金・金 'money' accounting for 15 percent of the responses. There are also two secondary responses at 10 percent; namely, 切手 'stamps' and 収集 'collection'. Some of the remaining core associates are 人 'people' (8%), 集合 'set' (6%), ゴミ 'rubbish, trash' (6%), and コレクター 'collector' (6%). Consistent with their respective word classes of noun and verb, these two words exhibit different kinds of syntagmatic responses. Compared to the very strong association between the adjective 寒い・さむい 'cold' and the noun 冬 'winter', more of the core responses for the verb 集める 'gather, collect' are nouns that could either occupy the direct object slot (i.e., お金・金 'money', 切手 'stamps', 人 'people', ゴミ 'rubbish, trash') or the subject slot (i.e., コレクター 'collector').

2.4 Future Development of the Database

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. These responses represent a substantial initial stage in the construction of a large-scale database of word association norms for basic Japanese vocabulary. However, given the preparation and inputting burdens involved in administering paper questionnaires, the present project is also developing a computer-based version of the word association survey, with a view to conducting the survey over the Internet in order to efficiently obtain the large-scale quantities of responses that will be required to complete the construction of the word association database. While the discrete free word association task is relatively straightforward—the respondent is simply asked to provide the first meaningfully-related word that comes to mind when presented with a stimulus word—the major issue in developing the computer-based survey 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.

Accordingly, much of the preparatory work on the project has been devoted to coding the survey corpus with information to use in eliminating intra-list associations. The first type of information added was phonological data in the form of hiragana transcriptions, to control for homophones and orthographic variants. The second type of information was a code relating to the orthographic form of the items (i.e., single kanji, multi-kanji, and mixed kanji-kana words, etc.) to ensure that respondent lists consist of a mixture of orthographic types, based on the distribution within the corpus, to reduce the possibility of form-related

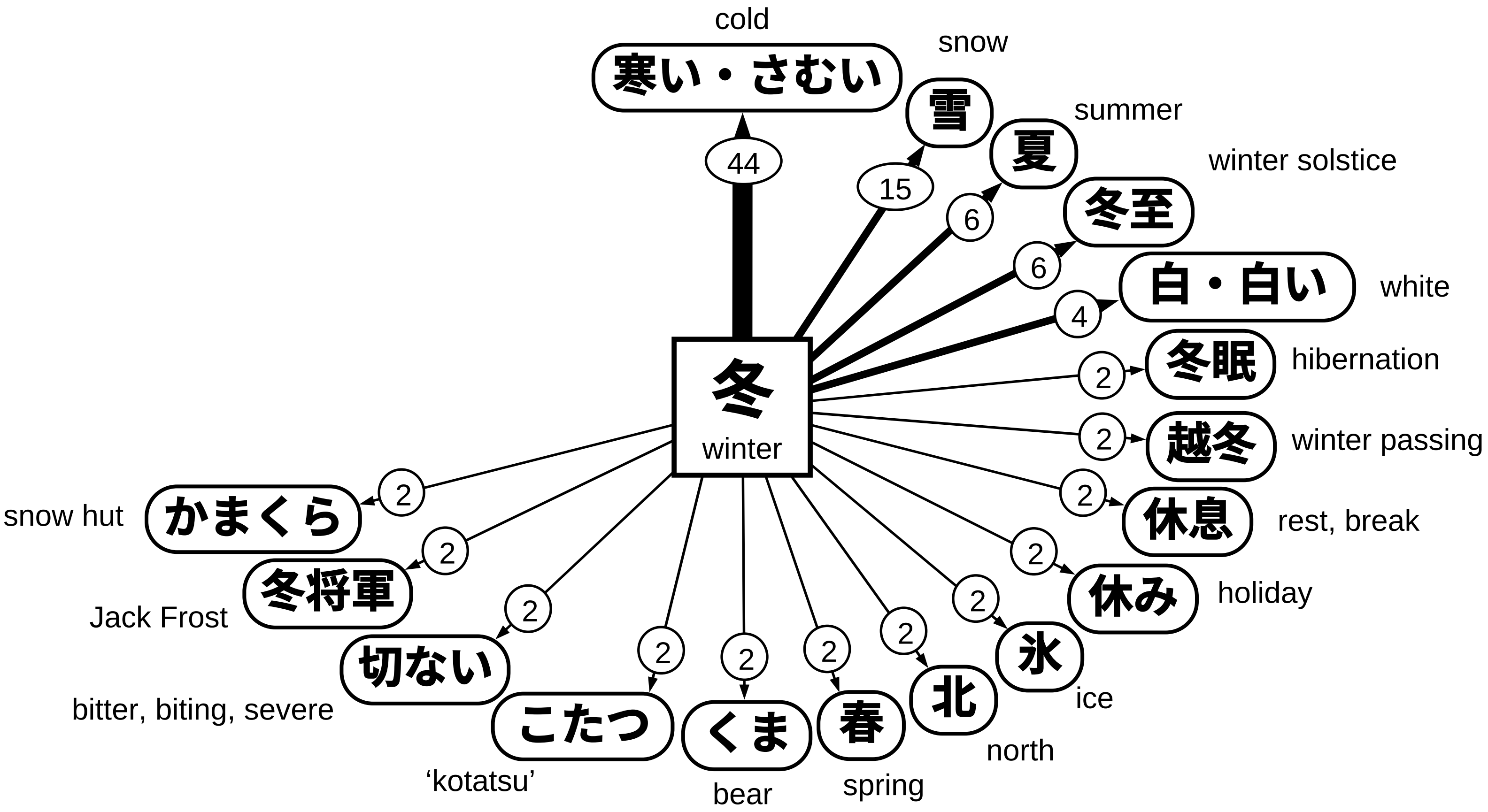

Figure 1. The associate set for 冬 ‘winter’, consisting of 17 associations and a set of five core associates given by two or more respondents.   Note the numbers on the connecting arrows indicate the percentage of respondents providing the response.

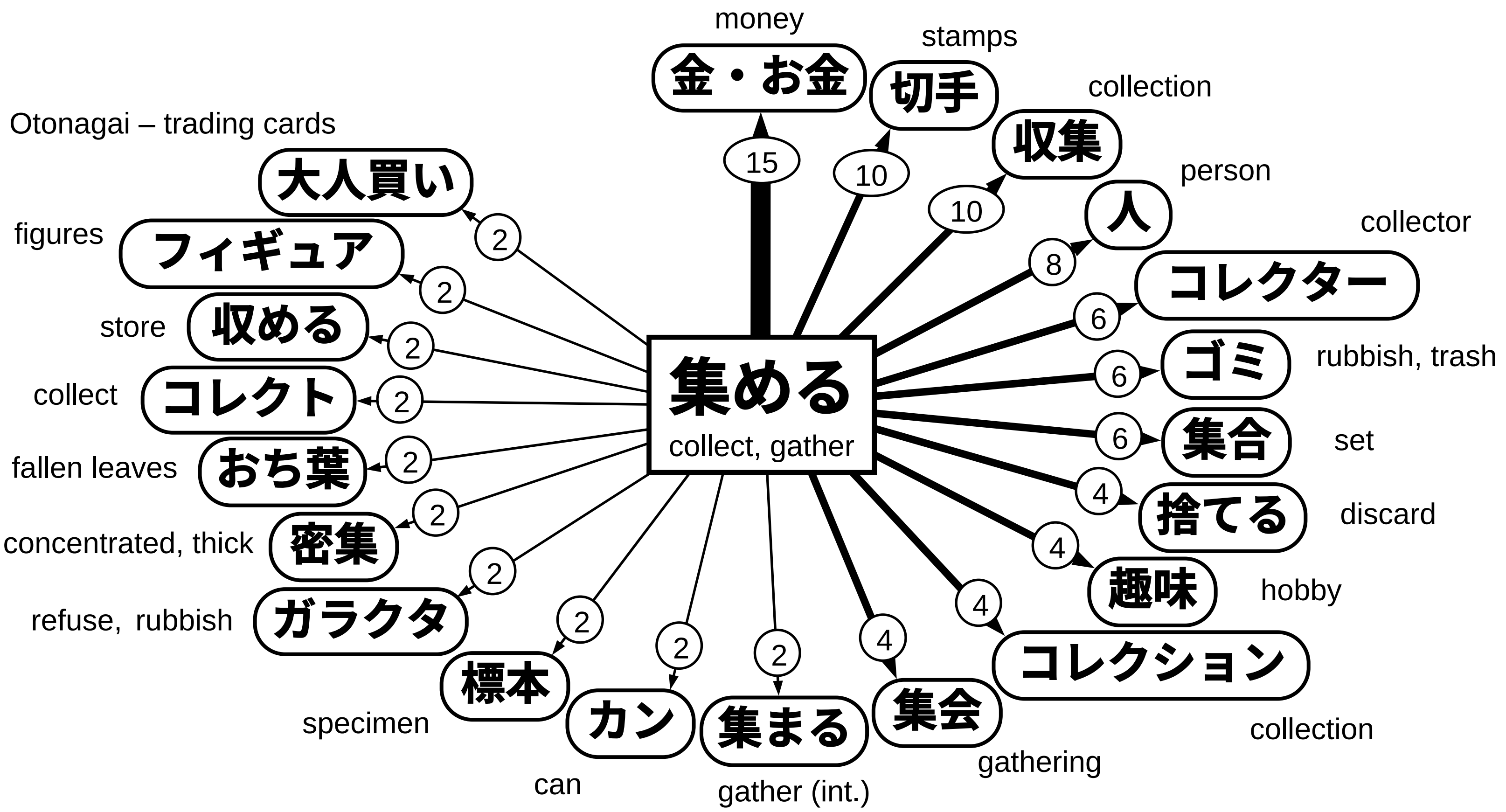

Figure 2. The associate set for 集める ‘collect, gather’, consisting of 21 associations and a set of 11 core associates given by two or more respondents. Note the numbers on the connecting arrows indicate the percentage of respondents providing the response.

response strategies. The third type of information was component kanji codes (kuten codes), so that a given kanji only appears once in a respondent list.

Beyond any possible phonological or morphological (i.e., constituent kanji morphemes) associations, it is also extremely important to ensure that a respondent list does not have multiple words from a particular semantic category, such as not having two colour words like 黒 'black' and 白 'white' within the same list. In order to assign appropriate semantic category information to the survey corpus items, the project initially consulted a well-known electronic Japanese thesaurus (Ikehara, Miyazaki, Shirai, Yokoo, Nakaiwa, Ogura, Ōyama, & Hayashi, 1999). However, while the thesaurus has a detailed classification for Japanese nouns involving 2,715 categories hierarchically organized into 12 levels, because it adopts a different approach to other word classes, it proved to be rather unsuitable for the project's purposes. Therefore, the semantic category codes assigned to the survey corpus are based on the National Institute for Japanese Language's (2004) recently revised semantic classification, which presents a more consistent approach to all word classes, albeit with fewer broader semantic categories. As noted earlier, the semantic category information was added to the survey corpus prior to the preparation of the second survey and was therefore also used in creating survey lists. This data proved to be extremely effective for very few intra-list associations were identified in the native speaker checks of the survey lists. As an additional measure in eliminating intra-list associations, the word association responses already collected will also be utilized within the program to generate respondent lists, based on unique item identification codes, so that all identified associates are also excluded.

Once the computer-based version of the survey is complete, the next phase of data collection will focus on obtaining up to 50 responses for all 5,000 kanji and words in the survey corpus. When the Japanese word association database reaches that stage, a new version of the database will also be made publicly available. Once 50 responses have been collected for all items, the core associate sets will be examined in order to identify all the associate words that are not already part of the survey corpus. For example, in the case of the associate set for 冬 'winter', although 寒い・さむい 'cold', 雪 'snow', 夏 'summer' and 白・白い 'white' are already included, 冬至 'winter solstice' is not. The survey corpus will then be expanded to include all core associates, required in order to complete lexical association network maps for the basic vocabulary items. The project also plans to start conducting the survey on the Internet by that stage to obtain the large quantities of word association responses necessary to complete the construction of the large-scale Japanese word association database.

3. Applications of the Japanese Word Association Database

Part 2 of the paper introduced the construction of a large-scale Japanese word association database, reporting on initial collections of word association responses through two large paper questionnaire surveys and noting plans to further develop the database. After briefly introducing the basic concept for the lexical association network maps to be developed based on the database, this part of the paper touches on some interesting applications in the areas of cognitive science and Japanese lexicography and language instruction.

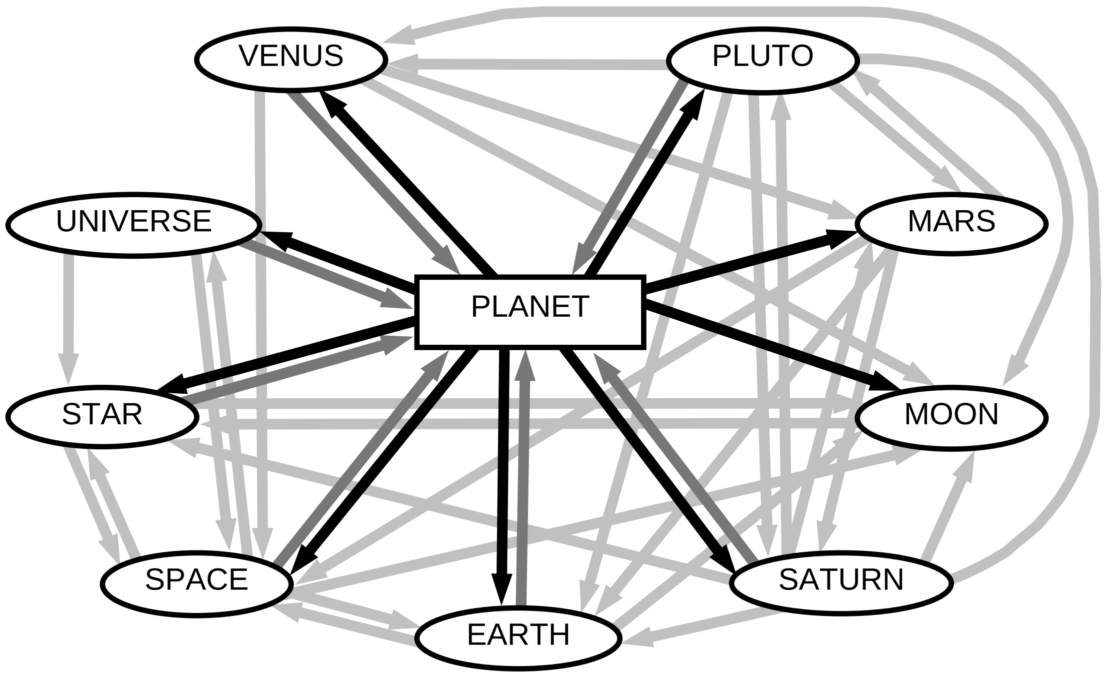

Figure 3. Basic concept of the lexical association network maps. In addition to forward associations, indicated by the black arrows from 'planet' to its nine core associates, the maps also include backward associations, indicated by the dark gray arrows, showing which associates also elicit the target word as a response and the associations between all member items of the associate set, indicated by the light gray arrows (based on Nelson & McEvoy, 2005)

3.1. Lexical Association Network Maps

A particularly promising application of the large-scale Japanese word association database is the creation of lexical association network maps as a means of capturing important aspects of words through their associate sets and patterns of connectivity. Figure 3 illustrates the basic concept of the lexical association network maps with an example for the English word 'planet' based Nelson and McEvoy (2005). At the heart of the lexical association network map is the set of core associates (responses provided by two or more respondents) that the target word elicits, together with the strengths of those associations indicated by response frequency. In the cases of the two Japanese examples shown in Figures 1 and 2, the sets would consist of 5 and 11 associates respectively. The lexical association network maps also feature backward associations—the associates that elicit the target word as a word association response—in terms of both number and associative strength. In Figure 3, these are indicated by the dark gray arrows. Beyond the forward and backward associations between the target word and its set of associates, which are clearly independent, association density—the levels and strengths of associations between all members of an associate set—can also be represented in the maps (indicated by the light gray arrows in Figure 3).

The lexical association network maps can complement other approaches to capturing aspects of lexical knowledge. As Steyvers and Tenenbaum (2005) speculate, the similarities that they identified between the semantic network based on word associations (Nelson, McEvoy, & Schreiber, 1998) and the one based on WordNet (Fellbaum, 1998) probably reflect pervasive and deep features of semantic knowledge. The possibility of tapping into such features through free word association responses suggests that they could be a realistic alternative to the expertise of the lexicographer required to define synonym sets for WordNet. The natural emergence of structure in the network maps from free associations also avoids possible concerns for associative concept dictionaries and ontologies where the hierarchical structure is theory-driven or limited to specific pre-defined relationships. The finding reported by Steyvers, Shiffrin, and Nelson (2005), noted earlier, that measures based on a word association space were better predictors of three kinds of episodic memory tasks than the LSA-based measures also suggest that word association normative data can be more sensitive to certain aspects of lexical knowledge than corpora-based collocation data.

3.2. Cognitive Science

3.2.1. Memory Research

As Nelson and McEvoy (2005) remark, their database of word association norms has made it possible for them to investigate the effect of a word's existing associative structure on recall and recognition memory for recent experiences, by allowing them to systematically manipulate the association structures of words in terms of set size, resonance (or backward association) and connectivity to examine how these features influence memory performance. Word association normative data is also playing an important role in false memory research. For example, employing the so-called Deese-Roediger-McDermott (DRM) paradigm, Roediger and McDermott (1995) provide evidence of false recognition. They found that after being presented with 15 words (e.g., *bed*, *rest*, *awake*, etc.) that are strongly associated to a word that was not presented (e.g., *sleep*), participants would often recall, recognize, and even claim to remember seeing the non-presented word even though they had not. The Japanese word association database is a valuable resource that will make it possible to conduct similar memory research using Japanese language stimuli.

3.2.2 Visual Word Recognition and Connectionist Modeling

With compounding being so productive in Japanese, the language is particularly suitable for investigating the extent of morphological involvement in the organization of the mental lexicon. Joyce (2002, 2004) has specifically addressed this issue by examining the lexical representation and retrieval of two-kanji compound words in a series of constituent-morpheme priming experiments. Comparing the facilitation in the lexical decision task from component kanji across different types of compound words, such as modifier + modified, verb + complement, complement + verb combinations, he has found that generally both constituents facilitate responses and, in most cases, at similar levels, indicating that morphology is important in the organization of the mental lexicon.

Moreover, in experiments that manipulated the positional frequency of the verbal constituents in verb + complement and complement + verb compound words, reversed patterns of priming were observed in high positional frequency conditions, indicating that verbal information may also be important within the mental lexicon. However, while this research suggests a central role for morphological information, possible confounding factors, such as association effects, need to be investigated. The Japanese word association database will also make it possible to run experiments that manipulate word associations in further investigating the role of morphological information within the mental lexicon.

Based on constituent-morpheme priming experiments, Joyce (2002, 2004) has proposed adapting for the Japanese mental lexicon a version of the multilevel interactive-activation framework. A special feature of the model is the incorporation of lemma-unit representations that mediate the links between orthographic, phonological and semantic information (and, potentially, syntactic representations as well), as an appealing way of handling the complex nature of the Japanese writing system. Proposed primarily as a model of visual word recognition, to date relatively little attention has been given to the rather primitive semantic representations within the Japanese lemma-unit model, but an extremely interesting approach to developing the semantic aspects of the model will be to incorporate lexical association network maps within the model.[5]

3.3. Japanese Lexicography and Language Instruction

3.3.1 Japanese Lexicography

There are also direct and interesting Japanese lexicographical and language instruction applications of the word association database and the lexical association network maps (Joyce, 2005). For instance, the inclusion of word association norms within the lexical entry would greatly enrich the variety of lexical information presented to the dictionary user. So, in addition to listing information relating to the entry word's pronunciation, its definitions, its inflectional and derivational forms, and idiomatic expressions, the dictionary could also provide word association information, which together with response frequency data, would assist in identifying for a given entry word the relative importance of different kinds of associative relations, such as synonyms and antonyms, as well as common modifiers and complements and various other relations.

Supplementing electronic dictionaries with the word association normative data and the lexical association network maps could also support user-friendly search functions (Zock & Bilac, 2004). Current electronic dictionaries are of little help to the user experiencing the common 'tip of the tongue' phenomenon (Brown & McNeill, 1966). While the individual in this situation is unable to retrieve the desired word from their mental lexicon, typically they are aware of various kinds of semantic information related to the target word, such as the associative relations the word has with other words. If the word association normative data and the lexical association network maps were incorporated into electronic dictionaries, then inputting an associated word (either single or multiple word entry) could provide access to the relevant lexical association network map, from which the user could follow appropriate association links until the target word is identified.

[5] The basic suggestion here is quite similar in nature to the attempt by Ijuin, Fushimi, and Tatsumi (2003) to merge a hierarchically-organized thesaurus (Ikehara, et al, 1999) with a Japanese version of the triangle model (Fushimi, Ijuin, Patterson, & Tatsumi, 1999). However, as already noted, the fact that the Japanese thesaurus classifies nouns differently to other word classes suggests that it will be difficult to handle the entire lexicon consistently with that approach.

3.3.2. Japanese Language Instruction

Memory researchers have long demonstrated that the categorization and semantic organization of stimulus materials have dramatic effects on retrieval performance (e.g., Bower, Clark, Winzenz, & Lesgold, 1969). Such findings suggest that the lexical association network maps for basic Japanese vocabulary can open up effective strategies for Japanese language instruction. For example, in the context of second language teaching, Morin and Goebel (2001) have reported effects of using semantic mapping techniques with beginner level college students of Spanish. After engaging in semantic mapping activities—essentially semantic clustering based themes and associations, students had greater familiarity for the vocabulary and were better at classifying words under appropriate headings than a control group. In a recent review of vocabulary instruction, Blachowicz and Fisher (2000) stress the importance of (1) students being active in developing an understanding of words and how to learn them, (2) personalizing their word learning, (3) being immersed in the vocabulary, and (4) having access to multiple sources of information to enhance learning through repeated exposures. The lexical association network maps, depicting sets of associatively-related words based on free word association responses from native Japanese speakers, represent a kind of authentic study material and a valuable reference source for the second language learner in judging the naturalness of lexical combinations.

In pursuing these Japanese lexicographical and language learning applications the research project is also working to create a comprehensive kanji database and integrated kanji instruction system. The key concept behind the system is an electronic study notebook that would build into a personalized dictionary by drawing on the reference database through various learning assignments, ranging from basic *look-up tasks* to more advanced *expansion tasks* that would help the learner develop a deeper understanding of kanji structure, of the morphological structures of compound words, and of the rich networks of associations between words revealed through the large-scale database of Japanese word associations.

In summary, this paper has reported on a research project to construct a large-scale database of word association norms for basic Japanese vocabulary. The results from the first stage of data-collection, through two major questionnaire surveys in which up to 50 word associations were obtained for a sample of 2,100 basic Japanese kanji and words, are being made available as version 1.0 of the Japanese word association database. After noting plans for the continuing development of the database, this paper has also discussed some interesting applications of the word association nominative data, from the development of lexical association network maps, cognitive science experimentation, to Japanese lexicography and language instruction.

References

Amano, S., & Kondō, T. (1999). *Nihongo no goitokusei* [Lexical properties of Japanese]. (Vols. 1-6, NTT database series). Tokyo: Sanseidō.

Bower, G. H., Clark, M. C., Winzenz, D., & Lesgold, A. (1969). Hierarchical retrieval schemes in recall of categorized word lists. *Journal of Verbal Learning and Verbal Behavior*, *8*, 323-343.

Brown, R., & McNeill, D. (1966). The “tip of the tongue” phenomenon. *Journal of Verbal Learning and Verbal Behavior*, *5*, 325-337.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*, 22-29.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*, 407-428.

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

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

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

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

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.

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

Ijuin, M., Fushimi, T., & Tatsumi, I. (2003). Hyōki no chigai ga imi no keisan ni ataeru eikyō ni tsuite: Konekushonisuto moderu ni yoru yomi no shimyurēshon [On the effects of orthographic variations on the calculation of meaning: A connectionist model simulation of reading]. *Proceedings of the 67th Japanese Psychological Association's Annual Meeting*, 737.

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

Ishizaki, S. (2004). *Rensō gainen jisho Verison 1.0* [Associative concept dictionary, Verison 1.0]. CD.

Joyce, T. (2002). Constituent-morpheme priming: Implications from the morphology of two-kanji compound words. *Japanese Psychological Research*, *44*, 79-90.

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. (2005). Lexical association network maps for basic Japanese vocabulary. In V. B. Y. Ooi, A. Pakir, I. Talib, L. Tan, P. K. W. Tan, & Y. Y. Tan, (Eds.), *Words in Asia cultural contexts*. (Proceedings of the 4th Asialex conference, 1-3 June 2005). (pp. 114-120). Singapore: Department of English Language and Literature, Faculty of Arts and Social Sciences, & Asia Research Institute, National University of Singapore.

Morin, R., & Goebel, J., Jr. (2001). Basic vocabulary instruction: Teaching strategies or teaching words? *Foreign Language Annals*, *34*, 8-17.

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

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York, NY: Newbury House.

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. (1984). *Nihongo kyōiku no tame no kihon goi chōsa* [A survey of fundamental vocabulary for Japanese language teaching]. Tokyo: Shuei Shuppan.

Nelson, D. L., & McEvoy, C. L. (2005). Implicitly activated memories: The missing links of remembering. In C. Izawa, & N. Ohta, (Eds.), *Human learning and memory: Advances in theory and application*. (The 4th Tsukuba International Conference on Memory). (pp. 177-198). Mahwah, NJ & London: 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 August 31, 2005, from http://www.usf.edu/FreeAssociation.

Okamoto, J., & Ishizaki, S. (2001). Gainenkan kyōri no teishikika to kizon denshi jisho to no hikaku [Construction of associative concept dictionary with distance information, and comparison with electronic concept dictionary], *Shizen Gengo Shori*, *8*, 37-54.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 803-814.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group, (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1. Foundations*. Cambridge, MA: MIT Press.

Sanseidō Henshūjo. (1991). *Atarashii kokugo hyōki handobukku* [A new handbook of Japanese orthography]. (Fourth edition). Tokyo: Sanseidō.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy, (Ed.), *Experimental cognitive psychology and its applications*. (Decade of behavior). (pp. 237-249). Washington, D.C.: 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.

Tamamura, F. (2003). Chūkyūyō goi: Kihon 4000 go [Intermediate vocabulary: Basic four thousand words], *Nihongo Kyōiku*, *116*, 5-28.

Umemoto, T. (1969). *Rensō kijunhyō: Daigakusei 1000 nin no jiyū rensō ni yoru* [Table of association norms: Based on the free associations of 1,000 university students], Tokyo: Tokyo Daigaku Shuppankai.

Yokoyama, S., Sasahara, H., Nozaki, H., & Long. E. (1998). *Shinbun denshi media no kanji: Asahi Shinbun CD-ROM ni yoru kanji hindohyō* [Electronic newspaper media kanji: Kanji frequency lists based on Asahi Newspaper CD-ROM]. Tokyo: Sanseidō.

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