



JWSAN: Japanese word similarity and association norm

Keisuke Inohara¹  · Akira Utsumi²

Accepted: 13 May 2021 / Published online: 18 June 2021
© The Author(s) 2021

Abstract We present a new Japanese dataset, Japanese Word Similarity and Association Norm (JWSAN), comprising human rating scores of similarity and association for 2145 word pairs, with a clear distinction between word similarity and word association. Computational models of human semantic memory or mental lexicon, such as distributed semantic models, must predict not only association but also similarity. People can distinguish between word similarity and association. However, although the SimLex-999 dataset is publicly available for English, there is no Japanese similarity dataset with a clear distinction between the two types of word relatedness. JWSAN is the first large Japanese dataset with similarity and association ratings, containing noun, verb, and adjective word pairs. It is also characterized by data collection from a sufficient number of age- and gender-controlled assessors, with similarity and association ratings obtained via a web-based survey conducted of 6450 native speakers of Japanese. In addition, the effects of the gender and age of the raters were also examined; these factors were only given scant consideration in the past. This dataset can act as a benchmark for improving distributed semantic models in Japanese.

Keywords Word similarity dataset · Word association dataset · Japanese word similarity · Distributed semantic model

✉ Keisuke Inohara
kei.inohara@gmail.com

¹ College of Liberal Arts and Sciences, Kitasato University, 1-15-1, Kitazato, Minami, Sagamihara, Kanagawa 252-0373, Japan

² Department of Informatics, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan

1 Introduction

A distributional semantic model is a method of representing a word as a multi-dimensional vector learned from a significant number of use cases (e.g., Mikolov et al., 2013; Pennington et al., 2014). Several distributional semantic models have been developed. These models have shown impressive performance in various fields, such as natural language processing (Mikolov et al., 2013; Turney & Pantel, 2010), cognitive science (Jones et al., 2015; Mandera et al., 2017; Utsumi, 2011), and neuropsychology (Anderson et al., 2017; Mitchell et al., 2008).

To improve distributional semantic models further, accurate evaluations of these learning models are necessary. Word similarity prediction has been frequently used to assess the performance of distributional semantic models (De Deyne et al., 2009; De Deyne, Perfors, et al., 2016; Lenci, 2018; Levy et al., 2015; Mandera et al., 2017; Rothe & Schütze, 2017). The demonstration of high performance for this task is indicated by the model learning human semantic knowledge regarding the meaning of words.

In the ensuing subsections, we first give a brief summary of the concept of similarity, focusing on the difference between similarity and association. Next, we review the similarity databases in English and Japanese. Finally, we summarize the differences between the existing datasets and the objectives of this study.

1.1 Importance of distinguishing between similarity and association

It is difficult to clearly define “similarity.” We can find similarity for almost any pair of entities if we want to (De Deyne et al., 2016). In this study, we capture similarity in contrast to “association” with reference to previous studies.

Similarity, especially semantic similarity, has traditionally been defined intuitively. That is, how close they are to a “synonym,” and humans can assess this intuitively (Miller & Charles, 1991). Indeed, using conceptual features extracted by humans (such as the concept a robin <lays eggs> and <can fly>), we have also found that semantic similarity is well captured by shared conceptual features (De Deyne et al., 2009). A more objective framework for capturing similarity has been proposed, based on information from WordNet and Roget’s Thesaurus (Jarmasz & Szpakowicz, 2003; Resnik, 1995).

The distinction between “similarity” and “association” is important. In most cases, similarity is a special case of association (Budanitsky & Hirst, 2006). To borrow an example from Hill et al. (2015), the difference between similarity and association is exemplified by the concept pairs [car, bike] and [car, petrol]. Car is said to be (semantically) similar to bike and associated with (but not similar to) petrol. Intuitively, car and bike can be understood as similar because of their common physical features (e.g., wheels), their common function (transport), or because they fall within a clearly definable category (modes of transport). In contrast, car and petrol are associated because they frequently occur together in space and language, in this case as a result of a clear functional relationship. As another example from Scheible et al. (2013), the synonyms hot and scorching and

the antonyms hot and cold are both strongly “associated,” in that they share the dimension of “temperature.” On the other hand, in the dimension “temperature,” hot and scorching are close to each other whereas hot and cold are far from each other, the former pair is strongly “similar” to each other, whereas the latter is less strongly “similar.”

It would also be useful to distinguish between the concepts that are close. The term “association” is often confused with “word-association,” which is measured by a word-association task. It can be said that there is an association between words that is often answered by the word association task, but it is not necessarily similar. Co-occurrence, which is usually calculated from the corpus, is another type of association. However, again, there is an association between frequently co-occurred words, but it is not necessarily similar.

1.2 Similarity dataset

1.2.1 English dataset

Quite a few English datasets for word similarity tasks have been released, e.g., SimLex-999 (Hill et al., 2015), WordSim353 (Finkelstein et al., 2002), MC (Miller & Charles, 1991), RG (Rubenstein & Goodenough, 1965), Stanford’s Contextual Word Similarities dataset (Huang et al., 2012), Stanford Rare Word similarity dataset (Luong et al., 2013), CARD-660 (Pilehvar et al., 2018), Verb Similarity dataset (Baker et al., 2014), and SimVerb-3500 (Gerz et al., 2016). Of these, WordSim353 (Finkelstein et al., 2002) and SimLex-999 (Hill et al., 2015) are used with a distinction between similarity and association. The SimVerb-3500 (Gerz et al., 2016) dataset is an extension of the SimLex-999 set and contains 3500 English verbs.

WordSim353 (Finkelstein et al., 2002) does not make a distinction between “similarity” and “Association” in its instruction to participants, in contrast to Hill et al. (2015), as discussed below. Therefore, although the dataset name contains the word “similar,” it can be considered as a dataset that measures “association.” Therefore, Agirre et al. (2009) isolated the pairs in WordSim353 that were classified as synonyms, antonyms, identical, or hyponym-hyperonyms as similar pairs named “WordSim353 similarity” and used others as association pairs named “WordSim353 relatedness.” Thus, the method of separating “similarity” and “association” here is post hoc, based on the linguistic relations of the word pairs.

Hill et al. (2015) pointed out that simply asking “please assign a numerical similarity score” (Finkelstein et al., 2002) causes the contamination of association to similarity because participants do not pay much attention to the detailed differences between similarity and association. Therefore, Hill et al. (2015) conducted a pure similarity rating task, with detailed instruction contrasting association and similarity, and compiled it into the SimLex-999 dataset. On the other hand, association was estimated from an existing dataset of word association tasks—specifically, University of South Florida Free Association Database (USF) (Nelson et al., 2004). For example, the SimLex-999 dataset (Hill et al., 2015) has the “new–old” pair. The “new–old” pair is associated but dissimilar because the two words are

significantly far apart on the time axis, like hot and cold in the dimension “temperature.” In fact, this pair had the 825th rank out of 999 word pairs on the similarity rank in the SimLex-999 dataset (Hill et al., 2015). In contrast, “association” refers to the degree of some type of involvement, including similarity. In the SimLex-999 dataset (Hill et al., 2015), the “new–old” word pair had the 9th highest rank out of 999 word pairs on the association index based on the USF (Nelson et al., 2004).

Thus, even in English, there are few datasets that distinguish between similarity and association. Moreover, even in the SimLex-999 dataset (Hill et al., 2015), the measurement of association is inferred from word association tasks. If a clear distinction is made between similarity and association in instruction, it would be easier to compare the two rating values. This is one of the motivations for conducting this study.

1.2.2 Japanese dataset

Similarity datasets are extremely rare in Japanese. Recently, Vulic et al. (2020) developed a unified procedure for dataset construction and applied it to 12 languages: Chinese Mandarin, Welsh, English, Estonian, Finnish, French, Hebrew, Polish, Russian, Spanish, Kiswahili, and Yue Chinese; although this work is a major recent development in the field, it does not include Japanese. To the best of our knowledge, the Japanese Word Similarity Dataset (JWSD; Sakaizawa & Komachi, 2018) is the only publicly available similarity dataset.

JWSD (Sakaizawa & Komachi, 2018) is the first similarity dataset in Japanese, and it is characterized by the inclusion of low to high frequency words and four parts of speech: Noun, Verb, Adjective, and Adverb words. The instructions are simple, showing examples and asking users to assign the degree of similarity (“We asked annotators to assign the degree of similarity for each pair using the same 10-point scale.”) like WordSim353 (Finkelstein et al., 2002). Thus, it does not make a clear distinction between similarity and association. In addition, as noted by Karpinska et al. (2018), there are several compound words comprising two or more morphemes and past-tense verbs in JWSD. This feature can cause problems when users apply distributional models to the dataset, such as the occurrence of several out of vocabulary (OOV) words.

A different, but related dataset to the similarity dataset is the dataset for the Japanese Bigger Analogy Task Set (jBATS; Karpinska et al., 2018). This dataset does not contain similarity rating data, but it specifies in detail the linguistic relationships of word pairs and contains a wealth of examples. Later, we will compare the dataset in this study with these two datasets.

1.3 Explaining similarity and association ratings via distributional models

Levy et al. (2015) distinguished and tried to explain similarity and association ratings via distributional models using WordSim similarity and WordSim relatedness (Agirre et al., 2009) and SimLex-999 (Hill et al., 2015) described above. The

distributional models that they used were positive pointwise mutual information (PPMI), PPMI with singular value decomposition (PPMI+SVD; Bullinaria & Levy, 2007), GloVe (Pennington et al., 2014), and skip-gram with negative sampling (SGNS; Mikolov et al., 2013). Using these models, it is possible to explain human rating values from the dataset. However, the correlation coefficients between these models and rating values vary. The highest correlation coefficients among these models and several parameters are approximately 0.79 for WordSim similarity, 0.69 for WordSim relatedness, and 0.43 for SimLex-999. The correlation coefficients depend on various factors of datasets. Thus, we need to construct similarity and association datasets that differ only in instruction between similarity and association tasks. In addition, this kind of study does not exist in Japanese although there is a study—specifically, Karpinska et al. (2018)—that examined JWSD (Sakaizawa & Komachi, 2018). This constitutes another of our motivations to construct a new dataset.

1.4 Purposes of this study

In this study, we developed a new Japanese word similarity and association dataset, called the Japanese Word Similarity and Association Norm (JWSAN). To the best of our knowledge, JWSAN is the first dataset¹ that contains both word similarity and association scores collected by instructions that clearly distinguish between “similarity” and “association.” JWSAN is also characterized by data collection from a sufficient number of age- and gender-controlled assessors. This study was performed by (1) choosing word pairs prior to the survey, (2) collecting similarity and association ratings using a web-based survey for a full dataset, JWSAN, (3) reporting the characteristics of JWSAN, and (4) analyzing the effects of the gender and age of the raters. Regarding (2), this study adopted a rating task not only for similarity but also for association to unify measurement methods between similarity and association. This is contrast to previous studies (Hill et al., 2015) that used the rating task for similarity and the word association task for association from USF (Nelson et al., 2004).

2 Methods

2.1 Design for selecting word pairs

We have developed a dataset containing three classes of word pairs: noun pairs (noun–noun), verb pairs (verb–verb), and adjective pairs (adjective–adjective). For the dataset to include word pairs whose degrees of similarity and association are distributed as widely as possible, we selected word pairs according to two assumptions. First, we hypothesized that two words that belong to the same

¹ Japanese datasets containing the results of either the word association task or the similarity rating task exist. However, there is no Japanese dataset containing the results of both the association (relatedness) and similarity rating tasks.

semantic category would have a higher similarity than two words that belong to a different semantic category. Thus, we created two types of word pairs using a thesaurus (Isahara et al., 2008): pairs of words that belong to the same semantic category (expected to have high similarity) and pairs of words that belong to different semantic categories (expected to have low similarity). Second, we assumed that a pair of words that frequently occurs in the same context has a high association. Thus, we created two types of word pairs using pointwise mutual information (PMI), an index that represents co-occurrence tendency: pairs of words that frequently co-occur in the same context (expected to have a high association) and pairs of words that are unlikely to co-occur in the same context (expected to have a low association). Based on these two assumptions, we created four types of word pairs: semantically similar and frequently co-occurring word pairs, semantically similar and infrequently co-occurring word pairs, semantically dissimilar and frequently co-occurring word pairs, and semantically dissimilar and infrequently co-occurring word pairs.

2.2 Selection of word pairs

Before selecting the words used for our dataset, we had to choose a proper word dictionary. This is very significant because Japanese sentences are written in a non-segmented form, and word boundaries are not explicitly marked. To process Japanese texts, which is necessary for training distributional word vectors, word segmentation must be performed using a morphological analyzer. However, different morphological analyzers use various word dictionaries; even worse, these dictionaries contain numerous inconsistencies from a linguistic perspective (Maekawa et al., 2014). Thus, we adopted UniDic (Ver. 2.1.2)² (Den et al., 2007). UniDic is a word dictionary developed to resolve the above problem and provide a proper tool for Japanese morphological analysis. The unit for identifying a word in UniDic is based on a short unit word (Maekawa et al., 2014), which does not contain compound words.

As a thesaurus for creating semantically similar–dissimilar pairs, we used Japanese WordNet (Ver. 1.1)³ (Isahara et al., 2008). WordNet is a type of thesaurus in which words are grouped into sets of semantic categories, called synset. We extracted nouns, verbs, and adjectives contained in both WordNet and UniDic.

As pre-processing, we excluded words comprising only one kanji character.⁴ Furthermore, low-frequency words were eliminated in reference to the vocabulary

² “unidic-mecab-2.1.2_src.zip” was downloaded from <https://ja.osdn.net/projects/unidic/>, and “lex.csv” was used.

³ <http://compling.hss.ntu.edu.sg/wnja/>.

⁴ Words comprising only one kanji character are highly ambiguous in that they are used not only as single words but also as affixes and abbreviations. For example, a Japanese kanji character “日” is a single word that refers to the sun; it is also used as an affix for counting days and abbreviation for Sunday. Thus, we only used words comprising two or more characters, including hiragana, katakana, and kanji characters.

list⁵ of Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014) as follows. For nouns, the bottom 75% (frequency of less than 152) were eliminated, and 4117 unique words were selected. For verbs, the bottom 50% (frequency of less than 91) were discarded, and 1589 unique words were extracted. For adjectives, because they were small in number, only those with frequencies less than four were eliminated, and 463 unique words were left.

We then constructed a set of two-word pairs using synsets of WordNet and PMI values from the word pool. First, we classified synsets into two types: one-word synsets that contain only one word in the vocabulary and multiple-word synsets that contain two or more words. Using multiple-word synsets, we created all pairs of words in the same multiple-word synset, except for those including the same kanji character, as candidates of semantically similar pairs. As candidates for semantically dissimilar pairs, we also created all pairs of words in different one-word synsets. Thus, we obtained a pool of semantically similar pairs, including 2387 noun pairs, 1577 verb pairs, and 214 adjective pairs, and a pool of semantically dissimilar pairs, including 493,053 noun word pairs, 83,132 verb pairs, and 3975 adjective pairs.

Next, we computed PMI⁶ values for all pairs of words that co-occurred in the BCCWJ corpus (Maekawa et al., 2014), because PMI cannot be computed for two words that do not co-occur. We then selected the top 400 pairs of PMI values for nouns, 100 pairs for verbs, and 50 pairs for adjectives from the pool of semantically similar pairs, as semantically similar and frequently co-occurring word pairs (550 pairs). Similarly, semantically dissimilar and frequently co-occurring word pairs (550 pairs) were determined as the top 400 pairs of PMI values for nouns, 100 pairs for verbs, and 50 pairs for adjectives from the pool of semantically dissimilar word pairs. Furthermore, we determined semantically similar and infrequently co-occurring word pairs (550 pairs) and semantically dissimilar and infrequently co-occurring word pairs (550 pairs) by randomly choosing 400 pairs for nouns, 100 pairs for verbs, and 50 pairs for adjectives that did not co-occur in the corpus from the pool of semantically similar pairs and the pool of semantically dissimilar pairs. Consequently, the four word-pair types each comprised 400 noun pairs, 100 verb pairs, and 50 adjective pairs; thus, 2200 pairs were selected.

These 2200 pairs were divided into two sets of 1100 pairs each. Two graduate students judged each set to identify pairs of words that were difficult to understand. The percentage of agreements for the two sets were 98.3% and 97.5%. From these pairs, 55 (noun pairs: 22, verb pairs: 21, adjective pairs: 12) were judged to contain a word that was difficult to understand by at least one assessor. We eliminated these 55 pairs and used the remaining 2145 pairs for the web-based survey.

⁵ We used “BCCWJ_frequencylist_suw_ver1_0.tsv” in “BCCWJ_frequencylist_suw_ver1_0.zip”, downloaded from http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html.

⁶ From the BCCWJ (Maekawa et al., 2014) library subcorpus (short unit words), all paragraphs that contained four words or greater were extracted, to which PPMI was applied. The window size was 10 words before and after the target word.

2.3 Assessment set

Consequently, 2145 pairs were divided into 21 almost equally sized sets, each of which included 102 or 103 pairs. Word class and word pair types were balanced across 21 sets.

2.4 Participants

A total of 9253 native speakers of Japanese, over a wide age range, participated in a web-based survey via an Internet research company in Japan. Table 1 shows the number of participants for 10 classes of gender (male and female) and age (20 s, 30 s, 40 s, 50 s, and 60 s).

The participants were divided into 42 roughly equal-sized groups such that all groups contained at least 10 participants for each gender–age class. Each participant group was assigned one of 21 assessment sets and one of the two tasks (i.e., similarity rating and association rating). Hence, each participant was asked to rate either the similarity or association of one assessment set.

2.5 Query design and procedure

The participants were given different task-specific instructions. For the similarity rating task, we designed the instructions in accordance with Hill et al. (2015). In the instructions, we provided a number of examples of similar word pairs, rather than attempting to define the notion of similarity because it is difficult to capture a formal characterization of similarity, and, even if it exists, various instructions are needed for different concept types. However, the instruction for the association rating task is simple, as demonstrated in previous research (e.g., Finkelstein et al., 2002). Hence, participants in the similarity rating condition were given detailed instructions with some examples to clarify the distinction between similarity and association, whereas the participants in the association rating condition were given only brief instructions.

The instructions for the similarity rating task highlighted the importance of drawing a clear distinction between word similarity and word association, and the association was explained in addition to similarity using examples. To illustrate the difference between similarity and association, we provided instructions stating that, for example, “the word *tire* is “associated” with the word *car* in the sense that a tire is a part of a car, but they do not have similar meanings (or they are not synonymous) because a tire is different from a car itself.” For the participants in the similarity rating task, a validation test was administered immediately following the instructions. The validation test comprised three multiple-choice questions to identify the most similar pair from a set of three options, all of which were associated but only one of which was clearly similar (e.g., [cutter, paper], [cutter, frying pan], [cutter, knife]). The participants were provided with feedback on whether their answer was correct or not. In the instructions for the association rating task, the notion of association was explained by the following example: the degree of association between *scholar* and *book* was “6” (considerably associated), but the

AQ1
Please rate how "similar" word-pairs is from alternatives, "Not similar at all (1)" to "Extremely similar (7)", and select numbers consistent with your choice.

		Not similar at all	Considerably not similar	Slightly not similar	Neutral	Slightly similar	Considerably similar	Extremely similar
		1	2	3	4	5	6	7
liquid	friend	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
vegetable	woman	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
solution	work	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
salty	flavor	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
cottage	log	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
small	young	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
going out	walk	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
license	authorization	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
residence	home	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
forget	bury	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Fig. 1 Rating screen for similarity rating task (originally in Japanese)

degree of association between *ground meat* and *magnet* was “2” (considerably unassociated). To ensure that participants read the instructions, they were not allowed to move to the next page until 30 s had passed for the similarity rating task and 10 s had passed for the association rating task. All instructions translated from Japanese for the similarity and association rating conditions are shown in Appendices A and B, respectively.

After going through the instructions, participants were asked to rate the degree of similarity or association of the word pairs presented on the screen by choosing the most appropriate number on a seven-point scale, ranging from not similar (associated) at all (1) to extremely similar (associated) (7). The rating screen for the similarity rating task is shown in Fig. 1.

2.6 Data screening

We eliminated, as unreliable, all the rating data of participants who met any one of the following criteria: all rated values were identical, the task completion time was much longer (i.e., it was within the longest 5% of all participants), and the task completion time was much shorter (i.e., it was within the shortest 5% of all participants). After preprocessing, the number of participants decreased to 8132.

Next, for each combination of 21 assessment sets and two rating tasks, we eliminated the data of randomly chosen participants such that 10 groups of gender (male, female) and age (20 s, 30 s, 40 s, 50 s, and 60 s) had an equal amount of rating data (i.e., participants). As a result, the data from 6450 participants remained in use for the dataset, and each combination of assessment sets and rating tasks had at least 100 participants (10 for each group) and at most 190 participants (19 for each group).

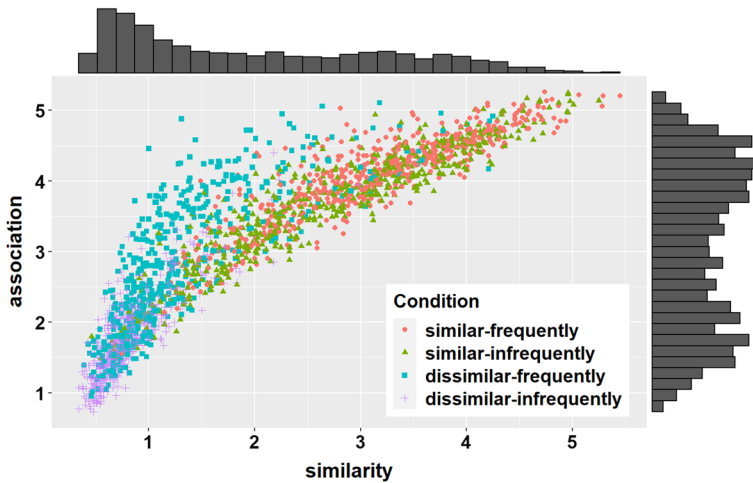


Fig. 2 Scatterplot of similarity and association ratings with histograms in JWSAN

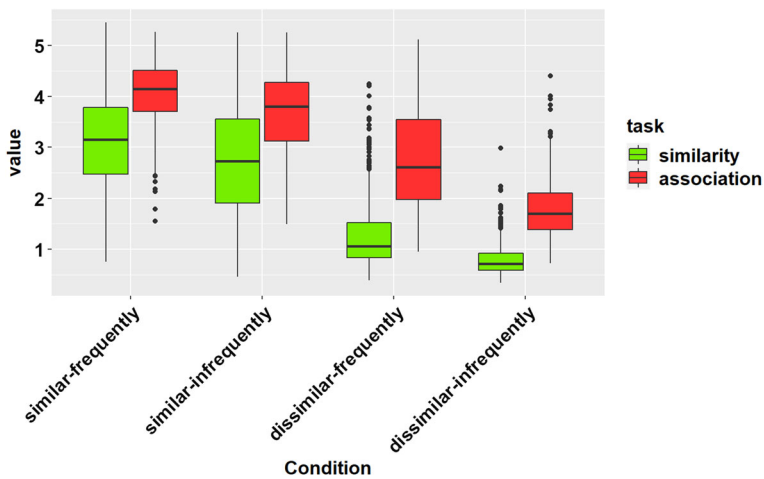


Fig. 3 Boxplots rated similarity and rated association in each condition set before survey

3 Results and discussion

We transformed the rating scale from the original 1–7 scoring to a range of 0–6 by subtracting one point from all rated scores such that the minimum value of scores was identical to other representative similarity datasets in English, such as WordSim-353 and SimLex-999. We then computed the mean similarity and association ratings for each of the 2145 word pairs, which are available in JWSAN.⁷

⁷ See the “Availability of data and material” section.

Table 1 Number of participants before and after data screening

	Age					Sum
	20–29	30–39	40–49	50–59	60–69	
Before screening ($n=9253$)						
Male	899	897	872	888	889	4445
Female	925	929	902	919	1133	4808
Sum	1824	1826	1774	1807	2022	9253
After screening ($n=6450$)						
Male	645	645	645	645	645	3225
Female	645	645	645	645	645	3225
Sum	1290	1290	1290	1290	1290	6450

Table 2 Examples of pairs with high similarity–high association, medium similarity–medium association, low similarity–high association, and low similarity–low association pairs for each POS category

Word 1	Word 2	POS	Similarity	Association	Difference
Big (でかい)	Very big (どでかい)	A	4.57	4.86	0.29
Blamed (忌ま忌ましい)	Bothersome (煩わしい)	A	2.59	3.05	0.46
Severe (きつい)	Sad (悲しい)	A	1.51	3.28	1.77
Far (遠い)	Yellow (黄色い)	A	0.57	0.93	0.36
Book seller (書店)	Book shop (本屋)	N	5.45	5.21	-0.24
One paragraph (一節)	Segment (部分)	N	3.53	3.86	0.33
Supply (供給)	Demand (需要)	N	1.31	4.88	3.57
Northwest (北西)	Conclusion (結論)	N	0.48	0.73	0.25
Plus (加える)	Add (足す)	V	4.86	4.76	-0.1
Insert (差し込む)	Interleave (挟む)	V	3.12	3.44	0.32
Borrow (借りる)	Lend (貸す)	V	1.39	4.34	2.95
Change clothes (着替える)	Fly (飛ぶ)	V	0.48	1.24	0.76

The actual Japanese words are shown in parentheses

The mean rating for all the 2145 pairs was 1.99 ($SD=1.25$) for similarity and 3.08 ($SD=1.16$) for association. The mean similarity rating was much lower than the midpoint of “3,” whereas the mean association rating was roughly equal to the midpoint. This suggests that many word pairs might be rated “dissimilar.” Table 2 shows some examples of word pairs in JWSAN and their similarity and association ratings. We selected examples of pairs with high-high, medium-medium, low–high, and low-low similarity and association values for each POS category.

3.1 Distributions of and correlation between rated similarity and association in JWSAN

Pearson's and Spearman's correlation coefficients between rated similarity and rated association were $r_s=0.91$, 0.94 ($p_s<0.01$). These results clearly show a very strong correlation between similarity and association ratings.

Figure 2 suggests one possible reason for the strong correlation. As shown in the scatterplot of Fig. 2, it is unlikely that two words are judged as semantically similar but unassociated, although the set of 2145 word pairs included infrequently co-occurring and semantically similar pairs. The mean similarity and association ratings of infrequently co-occurring and semantically similar pairs were 2.72 ($SD=1.05$) and 3.68 ($SD=0.80$). This suggests that semantically similar pairs were also judged as highly associated, which is not surprising. The similarity histogram in Fig. 2 shows a biased distribution of rated similarity in that there are many dissimilar pairs.

Next, we examined the two ratings based on the original conditions. Figure 2 shows a scatterplot by the conditions, Table 3 shows the mean and SD, and Fig. 3 shows a bar plot to clearly show the characteristics of each condition. With regard to similarity, WordNet's synsets-based procedure seems to have been successful: as can be seen in Table 3 and Figs. 2 and 3, the two semantically similar conditions are generally more similar than the two semantically dissimilar conditions. Many pairs of the semantically dissimilarity conditions are distributed near the lower limit, which may explain why the peak of the histogram of similarity in Fig. 2 is at the lower end of the similarity scale. It is clear from Table 3 and Fig. 3 that the procedure based on PPMI values for association did not have much effect on the similarity.

On the rated association, the procedures based on synsets in WordNet had an effect here as well. Overall, the two semantically similar conditions had a greater effect on rated association than the two semantically dissimilar conditions. This is to be expected, because, as discussed in the introduction, similarity is a kind of association. On the other hand, the procedure based on the value of PPMI had only

Table 3 Number of pairs and mean scores (SD) of rated similarity and rated association in each condition set before survey

	Semantically similar-frequently co-occurring	Semantically similar-infrequently co-occurring	Semantically dissimilar-frequently co-occurring	Semantically dissimilar-infrequently co-occurring
Synset	Same	Same	Different	Different
Co-occurrence	High PPMI	Never	High PPMI	Never
# Of pairs	540	540	533	532
Similarity	3.12 (0.89)	2.72 (1.05)	1.28 (0.68)	0.80 (0.32)
Association	4.05 (0.63)	3.68 (0.80)	2.78 (0.97)	1.78 (0.57)

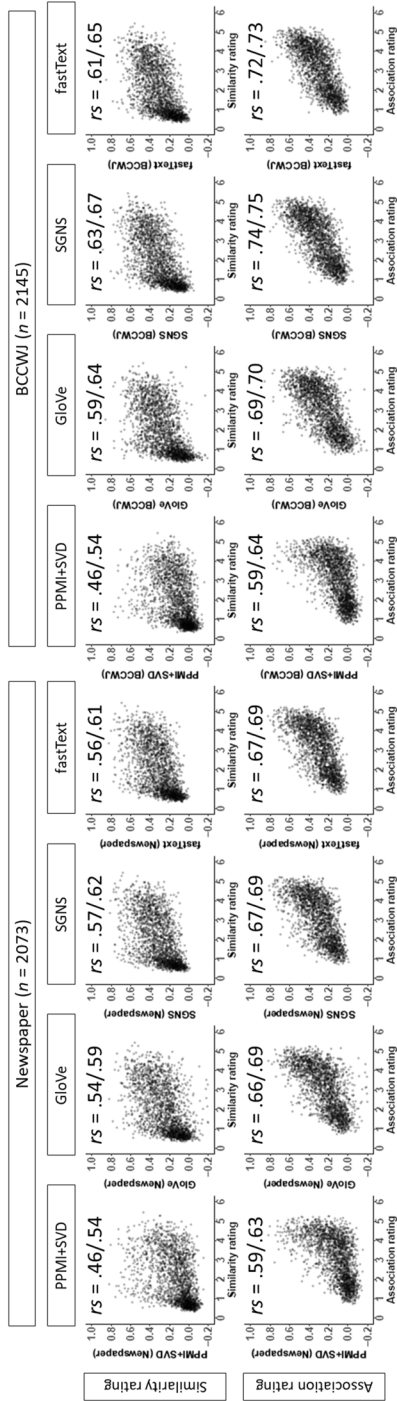


Fig. 4 Scatterplots of rated similarity and rated association in JWSAN and cosines of semantic spaces. Note: Values before and after the solidus symbol (/) are the Pearson's and Spearman's correlation coefficients, respectively. n is the number of pairs for the calculation of correlation coefficients. All correlation coefficients were statistically significant ($p_s < 0.01$)

minimal effect on the two semantically similar conditions, but had a significant effect on the two semantically dissimilar conditions, as is evident in Fig. 3.

3.2 Correlation between JWSAN and semantic spaces

In this section, we examine the applicability of JWSAN to the evaluation of distributional semantic models by analyzing the prediction performance of the distributional semantic models in estimating similarity and association ratings. Eight semantic spaces were constructed from combinations of four representative models and two training corpora. The distributional semantic models used for evaluation were positive pointwise mutual information with singular value decomposition (PPMI+SVD; Bullinaria & Levy, 2007), GloVe (Pennington et al., 2014), skip-gram with negative sampling (SGNS; Mikolov et al., 2013), and fastText (Bojanowski et al., 2017). Two corpora from which semantic spaces were trained were the Mainichi newspaper corpus (articles in 2000–2016), which includes approximately one billion word tokens, and the BCCWJ (National Institute for Japanese Language and Linguistics 2011), which includes 120 million word tokens.

The vocabulary of semantic spaces was determined as follows. For the newspaper corpus, we segmented all sentences using the morphological analyzer Mecab (ver. 0.996) with the IPAdic dictionary. As a result, the vocabulary included 750,258 unique words. For the BCCWJ corpus, we used the tags attached to the texts according to UniDic and all 175,800 unique words as vocabulary. All the parameters for training the eight semantic spaces were identical. The vector dimension was 300, and the size of the context window was ten words on either side of the target words. The size of the negative sampling of SGNS and fastText was five.

Figure 4 lists scatterplots, Pearson's and Spearman's correlation coefficients between the cosine values computed by semantic spaces and the mean similarity or association ratings for all 2145 pairs of words. The correlation coefficients between rated similarity and cosines ranged from 0.46 to 0.63 (Pearson) and from 0.54 to 0.67 (Spearman), whereas the correlation coefficients between rated association and cosines ranged from 0.59 to 0.74 (Pearson) and from 0.63 to 0.75 (Spearman). For all the semantic spaces, the correlation of association was higher than that of similarity, which is consistent with the results repeatedly observed for English word pairs (e.g., Hill et al., 2015; Levy et al., 2015).

However, in the case of JWSAN, the differences in correlation between similarity and association were not large because the correlations for similarity were relatively high. It has been generally observed in experiments using an English dataset that correlation coefficients between cosines computed from semantic spaces and similarity ratings are low. For example, Hill et al. (2015) reported that the correlation of similarity in SimLex-999 was 0.28 in the Mikolov et al. (2013) skip-gram model and 0.23 in the PMI+SVD model, whereas the correlation of association in WordSim353 was 0.44 and 0.38 for the skip-gram and PMI+SVD models, respectively. Similarly, Levy et al. (2015) also reported a maximum correlation of 0.44 between the cosine and rated similarity of SimLex-999, although they showed a maximum correlation of 0.79 between the cosine and rated

association in WordSim relatedness. All these findings suggest that our dataset JWSAN can easily predict the similarity of word pairs when compared with the English dataset SimLex-999. One possible reason for the unexpectedly high correlations between the similarity ratings of JWSAN and the model prediction is the biased distribution of similarity, as shown in Fig. 2. Many word pairs in JWSAN were judged to be dissimilar—1005 pairs (46.9%) have a similarity score of 1.5 or lower. In general, these dissimilar word pairs are likely to have low cosine values, regardless of the performance of the distributional semantic models. Such *easy* pairs would make the correlation between similarity ratings and cosines seem higher than it actually is.

3.3 Analysis of gender and age differences

We classified participants aged 20–39 ($n=2580$) as young and those aged 50–69 ($n=2580$) as old. The ratings for the participants aged 40–49 ($n=1290$) were not used when we calculated the values for each age group. The number of participants and mean (*SD*) of the rated similarity and association are listed in Table 4. Scatterplots and Pearson’s and Spearman’s correlation coefficients are shown in Fig. 5.

For rated similarity, we performed ANOVA with items as random variables ($n=2145$). The independent variables were gender (male and female; within) and age (younger and older; within). The ANOVA showed significant main effects of gender and age and interaction effect ($F_s(1, 2144)=305.99, 31.80, 147.11, p_s<0.01$). All simple main effects were significant: gender effects for young and old groups and age effects for male and female ($F_s(1, 2144)=20.90, 31.61, 10.09, 11.81, p_s<0.01$). Male participants generally rated the word pairs as more similar than did female participants; however, the effect of age differed between males and females. Old male participants rated the pairs as more similar when compared with young males,

Table 4 Numbers of participants, means, and *SDs* of rated similarity and association for each category of gender and age

			Younger (20–39)	Middle (40–49)	Older (50–69)	Total
The number of participants	Male		1290	645	1290	3225
	Female		1290	645	1290	3225
	Total		2580	1290	2580	6450
Mean (<i>SD</i>)	Similarity	Male	2.03 (1.19)	–	2.06 (1.28)	2.04 (1.22)
		Female	1.99 (1.28)	–	1.89 (1.34)	1.94 (1.29)
		Total	2.02 (1.22)	–	1.98 (1.29)	1.99 (1.25)
	Association	Male	3.03 (1.13)	–	3.02 (1.19)	3.03 (1.14)
		Female	3.12 (1.24)	–	3.14 (1.23)	3.12 (1.20)
		Total	3.08 (1.17)	–	3.08 (1.19)	3.08 (1.16)

Data from the middle-aged group were not analyzed when calculating the values of each age group. n is the number of pairs for the calculation of correlation coefficients

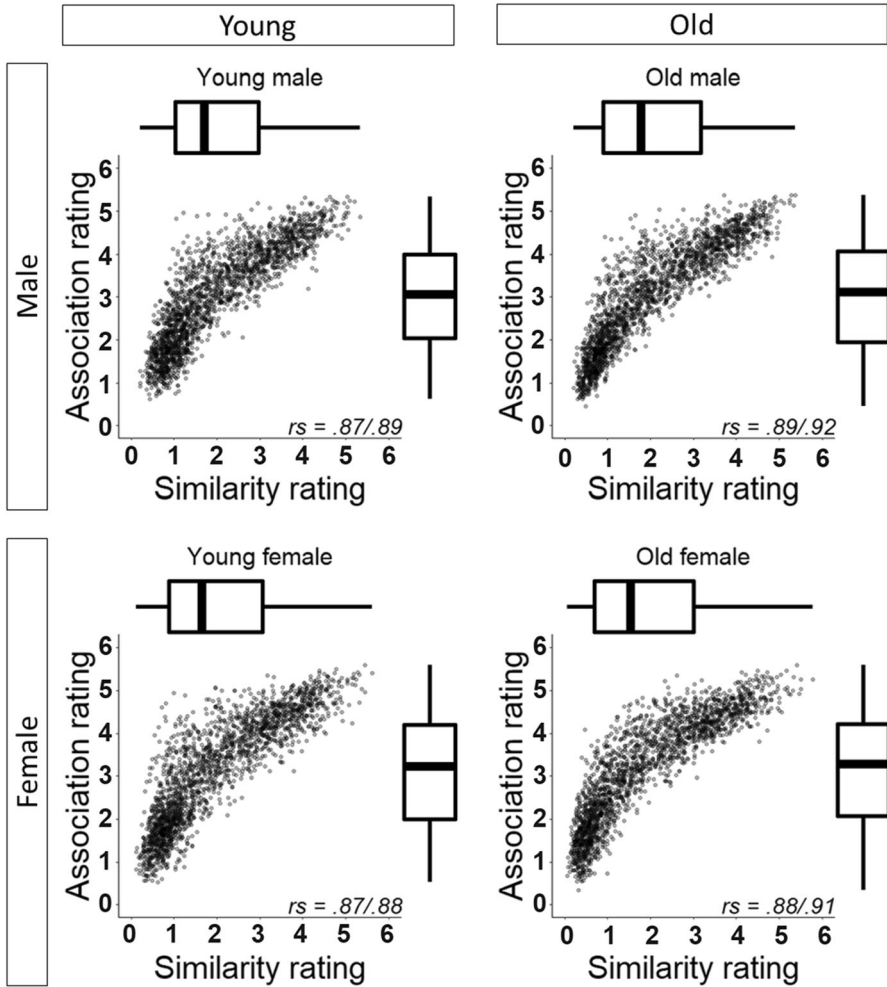


Fig. 5 Scatterplots and boxplots of rated similarity and rated association for each category of gender and age. Note: Correlation coefficients before and after slashes are Pearson's and Spearman's correlation coefficients, respectively. All correlation coefficients were statistically significant ($p < 0.01$)

whereas old females rated the pairs as less similar when compared with young females.

In the same manner, we conducted ANOVA for rated association ($n=2145$) to study the significance of gender, the effect of interaction ($F_s(1, 2144)=255.54, 7.10, p < 0.01$), and the marginal significance of the effect of age ($F(1, 2144)=3.58, p=0.059$). The simple effect of gender on young and old participants and age effect for females were significant ($F_s(1, 2144)=20.90, 31.61, 10.09, 11.81, p < 0.01$), whereas the simple effect of age for males was not significant ($F(1, 2144)=0.13, n. s.$). In contrast with the results of similarity ratings, female participants rated word

Table 5 Top 10 pairs in which similarity and association scores were rated higher by males

Rated similarity				
Word 1	Word 2	Male	Female	Difference
Intelligence (知性)	Brain (頭脳)	3.90	2.84	1.06
Expenditure (支出)	Fee (費用)	3.82	2.89	0.93
Elimination (拒絶)	Exclusion (排除)	3.34	2.43	0.91
Comeback (復帰)	Restitution (返還)	3.04	2.16	0.88
Exemplar (模範)	Sample (見本)	4.45	3.58	0.87
Sense (察知)	Feeling (感性)	2.47	1.61	0.86
Supervision (監視)	Management (観察)	3.80	2.94	0.86
Extermination (退治)	Expulsion (駆逐)	4.46	3.61	0.85
Protection (守護)	Defense (防御)	4.30	3.45	0.85
Traffic (交通)	Transportation (輸送)	3.36	2.52	0.84
Rated association				
Word 1	Word 2	Male	Female	Difference
Political (政治)	Decay (腐敗)	4.16	3.24	0.92
Abandonment (放棄)	Farming (耕作)	2.35	1.45	0.90
Bodyguard (護衛)	Aircraft carrier (空母)	4.25	3.43	0.82
Know-how (ノウハウ)	Skill (技能)	4.77	4.06	0.71
Scale (尺度)	Standard (標準)	3.96	3.31	0.65
Universe (宇宙)	Expansion (膨張)	3.68	3.05	0.63
Fall (下落)	Index (指数)	3.18	2.55	0.63
Hell (地獄)	Care (手入れ)	1.55	0.92	0.63
Walnut (胡桃)	Cheerful (陽気)	2.01	1.39	0.62
Boundary (限度)	Arrangement (配置)	2.12	1.52	0.60

Actual Japanese words are noted in parentheses

pairs as more associated when compared with males. In addition, old females rated the pairs as more associated when compared with young females.

Tables 5, 6, 7, 8 list the top 10 word pairs whose gender and age differences were the largest. In summary, under both similarity and association rating conditions, males tended to rate high economic, political, and abstract noun pairs (e.g., [expenditure (支出), fee (費用)], [political (政治), decay (腐敗)]). Females rated higher values for adjective pairs (e.g., [dazzling (眩しい), beautiful (美々しい)]) and verb pairs (e.g., [to lower (引き下げる), drop (落とす)]). Although it appears that the difference between young and old participants was not clear, there were some pairs whose absolute values of difference were greater than one point (e.g., [strange (可笑しい), curious (物珍しい)] in Table 7, [rail track (線路), railway (鉄道)] and [ladder (梯子), gradually (段々)] in Table 8).

Based on the results of the ANOVAs, we found that relatively large differences were caused by gender rather than age. Because absolute levels of rated values for

Table 6 Top 10 pairs in which similarity and association scores were rated higher by females

Rated similarity				
Word 1	Word 2	Male	Female	Difference
To lower (引き下げる)	Drop (落とす)	2.36	3.27	-0.91
Attack (攻撃)	Criticism (非難)	2.07	2.94	-0.87
To profit (儲ける)	Fight to win (勝ち取る)	1.76	2.60	-0.84
Joy (喜び)	Satisfaction (満足)	3.19	3.99	-0.80
Sorting (整理)	Adjustment (調節)	1.91	2.67	-0.76
Defense (守り)	Nursing (養護)	2.95	3.67	-0.72
Put out (出す)	Pull (引っ張る)	1.87	2.59	-0.72
Joy (喜び)	Pleasure (快楽)	3.89	4.58	-0.69
Cut (切れる)	Break (割れる)	2.25	2.92	-0.67
Selfish (さもしい)	Ugly (醜い)	2.00	2.66	-0.66
Rated association				
Word 1	Word 2	Male	Female	Difference
Dazzling (眩い)	Beautiful (美々しい)	2.66	3.60	-0.94
Defense (守り)	Nursing (養護)	3.62	4.52	-0.90
Notice (気付く)	Find out (見出だす)	3.44	4.32	-0.88
Come loose (緩む)	Fade (薄れる)	2.54	3.40	-0.86
Annoying (煩い)	Noisy (騒がしい)	3.48	4.33	-0.85
Worship (崇拜)	Love (愛好)	2.92	3.77	-0.85
Obstruction (阻止)	Disability (障害)	3.06	3.91	-0.85
Body weight (体重)	Height (身長)	3.82	4.66	-0.84
Hostility (反抗)	Objection (異議)	3.24	4.04	-0.80
Relationship (付き合い)	Come and go (行き来)	3.25	4.03	-0.78

Actual Japanese words are noted in parentheses

similarity and association would differ between males and females, we need to keep that in mind when selecting raters. However, as shown in Fig. 5, the correlation coefficients of rated values between male and female participants and between young and old participants were very high. This suggests that the relative differences in rated values among word pairs do not depend on gender and age. Furthermore, the results of the correlation coefficients of rated similarity and association listed in Fig. 6 demonstrate that both gender and age differences may not be affected by the rating method, resulting in the relative differences between rated similarity and association values.

3.4 Comparison with other Japanese datasets

It is useful to compare JWSAN with the limited Japanese datasets: JWSD (Sakaizawa & Komachi, 2018) and jBATS (Karpinska et al., 2018). Table 9

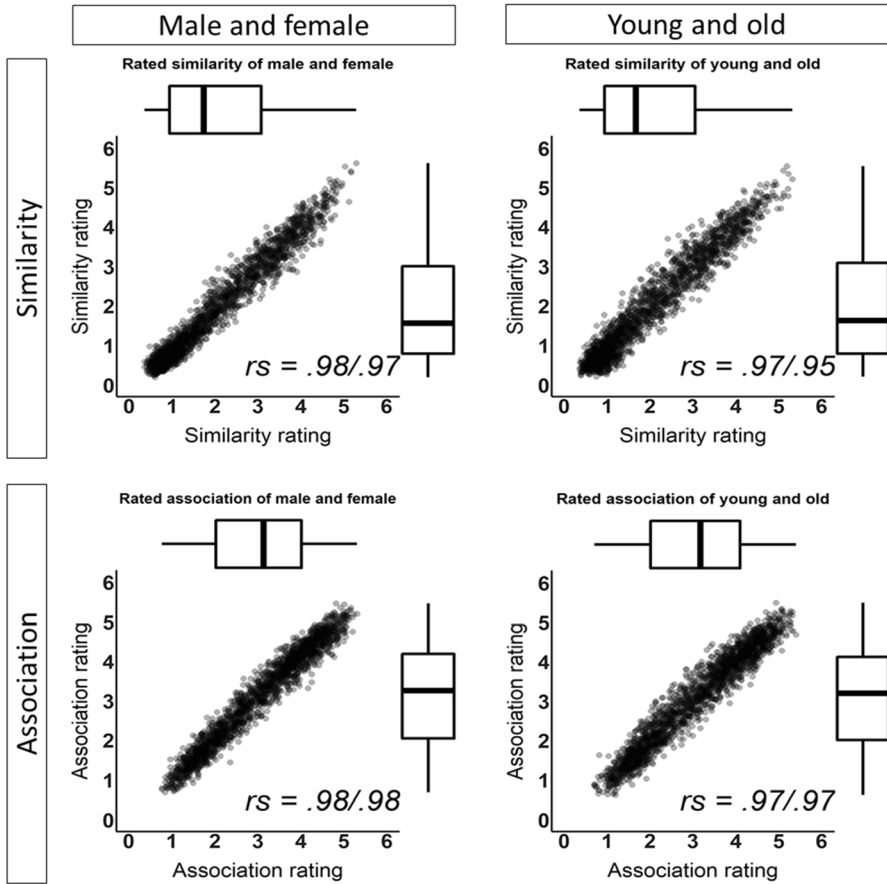


Fig. 6 Scatterplots and boxplots between rated similarities and between rated associations for male and female and for younger and older participants. Note: Values before and after the solidus symbol (/) are the Pearson's and Spearman's correlation coefficients, respectively. All correlation coefficients were statistically significant ($p < 0.01$)

compares JWSAN and JWSD in terms of the number of pairs, OOV percentages, and representative values. In terms of the number of pairs of parts of speech, JWSAN contains more nouns, while JWSD contains more verbs and adjectives. Furthermore, JWSD is unique in that it contains adverbs, which are not included in JWSAN. When we examined the common pairs of JWSAN and JWSD, we found a very small number of pairs: seven pairs of nouns and nine pairs of adjectives. These probably resulted from the fact that the procedures for creating word pairs differed between the two.

In terms of the OOV words, there are many instances in the JWSD. While we were able to calculate 2073 pairs (96.6%) for newspaper and 2145 pairs (100%) for BCCWJ in JWSAN (2145 pairs), 1928 pairs (43.5%) for newspaper and 1102 pairs (24.9%) for BCCWJ in JWSD (4429 pairs). This may be owed to the fact that the

Table 7 Top 10 pairs in which similarity and association scores were rated higher by younger participants

Rated similarity				
Word 1	Word 2	Young	Old	Difference
Strange (可笑しい)	Curious (物珍しい)	2.23	1.19	1.04
Temple (寺院)	Shrine (神殿)	3.45	2.53	0.92
Assault (加害)	Error (過失)	3.06	2.15	0.91
Test for certification (検定)	Survey (調査)	2.53	1.63	0.90
Visiting (見舞い)	Inspection (視察)	2.28	1.39	0.89
Deviation (偏差)	Weapon (武器)	1.37	0.50	0.87
Expire (切れる)	Disappear (無くなる)	3.33	2.47	0.86
Exemplar (模範)	Sample (見本)	4.42	3.56	0.86
Wording (言葉遣い)	Market (市場)	1.18	0.35	0.83
Enterprise (事業)	Duty (任務)	2.38	1.56	0.82
Rated association				
Word 1	Word 2	Young	Old	Difference
Twins (双子)	Constellation (星座)	4.33	3.35	0.98
Intoxication (酔い)	Case (事件)	2.90	2.02	0.88
Commissioned officer (将校)	Birth (生まれ)	1.91	1.04	0.87
Commit crime (犯す)	To judge (裁く)	3.88	3.05	0.83
Oppressive (重苦しい)	Refined (心憎い)	3.29	2.46	0.83
lot (一口)	Allotment (配当)	3.80	2.98	0.82
Exemplar (模範)	Sample (見本)	5.07	4.28	0.79
Visiting (見舞い)	Inspection (視察)	3.17	2.40	0.77
Manor (荘園)	Lily (百合)	2.95	2.18	0.77
Assault (加害)	Error (過失)	4.42	3.67	0.75

Actual Japanese words are noted in parentheses

JWSD contains numerous compound words consisting of two or more morphemes or past-tense verbs, as noted by Karpinska et al. (2018).

Comparing the representative values, the neutral value on the JWSAN scale is “3,” whereas it is “5” for JWSD; for the JWSAN rated similarity, most values are smaller than “3,” as can be seen from the median; for the association, the median is 3.19, which is close to “3.” On the other hand, the JWSD has a median of 6.8, which is considerably larger than “5.” These probably resulted from the differences in the word pairs used for grading and the differences in teaching.

Furthermore, we analyzed the comparisons between semantic spaces and datasets performed for JWSAN (Fig. 4) to the JWSD (Table 10). We observed that the correlation values ranged from 0.20 to 0.25 for newspapers and from 0.23 to 0.31 for BCCWJ. Since the correlation coefficients of JWSAN (similarity) shown in Fig. 4 are in the range of 0.46–0.62 for newspapers and 0.46–0.67 for BCCWJ, those

Table 8 Top 10 pairs in which similarity and association scores were rated higher by older participants

Rated similarity				
Word 1	Word 2	Young	Old	Difference
Rail track (線路)	Railway (鉄道)	2.41	3.70	-1.29
Performance (上演)	Show (興行)	2.89	3.95	-1.06
Excuse (言い訳)	Apology (謝罪)	1.28	2.32	-1.04
Graceful (奥床しい)	Modest (慎ましい)	2.97	4.01	-1.04
Entrance (入り口)	Wood door (木戸)	2.40	3.41	-1.01
Graceful (奥床しい)	Discreet (慎み深い)	3.18	4.19	-1.01
Back door (勝手)	Kitchen (台所)	2.66	3.67	-1.01
Fierce (凄まじい)	Rough (荒い)	1.97	2.96	-0.99
Traffic (交通)	Transportation (輸送)	2.41	3.38	-0.97
Traffic (交通)	Freight (運輸)	2.75	3.69	-0.94
Rated association				
Word 1	Word 2	Young	Old	Difference
Ladder (梯子)	Gradually (段々)	2.50	3.65	-1.15
Plot (趣向)	Menu (献立)	2.27	3.40	-1.13
Incident (事変)	Riot (暴動)	3.45	4.47	-1.02
Entrance (入り口)	Wood door (木戸)	3.38	4.39	-1.01
Worship (崇拜)	Praise (賛美)	3.32	4.28	-0.96
Amend (改める)	Shift (転ずる)	2.85	3.81	-0.96
Darkness (暗黒)	Constellation (星座)	2.06	3.00	-0.94
Relationship (関係)	Affair (用事)	2.05	2.98	-0.93
Alley (路地)	Tenement house (長屋)	2.83	3.73	-0.90
Tone (調子)	Echo (響き)	2.50	3.40	-0.90

Actual Japanese words are noted in parentheses

of JWSD are noted to be lower than the values for JWSAN. This might be caused by the differences in the distributions of rated similarities or parts-of-speech ratios between JWSAN and JWSD, as shown in Table 9.

In addition, it is likely that the difference in word frequency between JWSAN and JWSD might affect the results. For comparison, we show the histograms of the word frequencies (Fig. 7) by the following procedure: first, we removed the word pairs, including OOV words, from each dataset; next, we made unique word lists contained in word pairs; finally, we added the word frequencies computed for the newspaper and BCCWJ corpora. We observe from Fig. 7 that the JWSD includes more low-frequency words than JWSAN. Owing to the difficulty of explaining rare-word similarities using semantic spaces (Luong et al., 2013), this feature could be one of the reasons for the low correlation coefficients in Table 10.

jBATS is a dataset of four types of relations—inflectional_morphology, derivative_morphology, encyclopedic_semantics, and lexicographic_semantics.

Table 9 Comparison of JWSAN and JWSD

Number of pairs	JWSAN		JWSD		Common pairs
Noun	1578		1103		7
Verb	379		1464		0
Adjective	188		960		9
Adverb	0		902		0
Sum	2145		4429		16
OOV percentages					
	Newspaper	BCCWJ	Newspaper	BCCWJ	
	3.4%	0%	75.1%	56.5%	
Representative values					
	Similarity	Association	Similarity		
Possible range	0–6	0–6	0–10		
Min	0.34	0.73	0		
Median	1.66	3.19	6.8		
Mean	1.99	3.08	6.46		
Max	5.45	5.26	10		

We calculated the out of vocabulary (OOV) percentages to compute the ratio of the number of word pairs whose words both appeared at least once in the corpus to the total number of pairs in each dataset. We used the mean (remove_extreme_annotator) column in JWSD as the rated similarity values

Table 10 Correlation coefficients of rated similarities in JWSD and cosines of semantic spaces

	PPMI+SVD	GloVe	SGNS	fastText
Newspaper (n=1928)	.24/.24	.20/.21	.25/.23	.24/.22
BCCWJ (n=1102)	.30/.31	.23/.24	.30/.30	.28/.28

Values before and after the solidus symbol (/) are the Pearson's and Spearman's correlation coefficients, respectively. We used the mean (remove_extreme_annotator) column in JWSD as the rated similarity values for the mean (n is the number of pairs for the calculation of correlation coefficients). All the correlation coefficients were statistically significant ($ps < 0.01$)

Further, each type has 10 categories, and each category has 50 examples. For example, in the category L07 [synonyms—intensity] of lexicographic_semantics, the word young (若い) has paired candidate words: immature (幼い), small (小さい), youth (年少), childish (幼稚), naïve (稚い), child (子供), baby (赤ちゃん). For all examples of all types and categories, when these were paired (e.g., young–immature, young–small, ...), the result was 13,897 pairs. Regardless of the order of the pairs, the common pairs with JWSAN were only 13 pairs (noun: 6, verb: 1, adjectives: 6), all of which belonged to the lexicographic_semantics type. Specifically, we have the following: seven pairs in L07 [synonyms—intensity], three pairs in L08 [synonyms—exact], one pair in L02 [hypernyms—misc], one pair

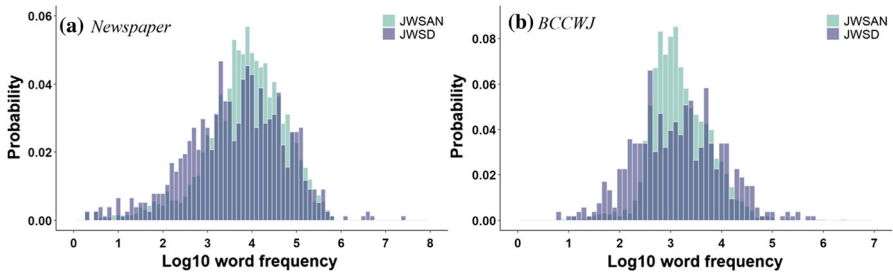


Fig. 7 Histograms of the Log₁₀ word frequencies for JWSAN and JWSD computed from the Newspaper and BCCWJ corpora

in L03 [hyponyms—misc], and one pair in L09 [antonyms—gradable]. To give specific pairs, JWSAN has a small–young pair with rated similarity and rated association of 1.92 and 3.02, respectively. There is also an immature–young pair, with rated similarity and rated association of 3.75 and 4.10, respectively.

As with JWSD, there were also very few common pairs with jBATS. In future study, in terms of increasing comparability, it would be worthwhile to assign rated similarity and rated association to JWSD and jBATS in the same manner as in the present study.

4 Conclusion

In the present study, JWSAN, a dataset of similarity and association ratings for Japanese word pairs, was created. JWSAN is the first Japanese dataset that includes both similarity and association ratings for Japanese word pairs. An additional characteristic feature of JWSAN is that it has a sufficient number of age- and gender-controlled participant-rated word pairs with instructions that clearly distinguish similarity from association. We hope that the proposed dataset will be widely used to assess the performance of Japanese distributional semantic models in the future.

Funding This research was supported by JSPS KAKENHI Grant Numbers JP15H02713 and 20H04488.

Data availability The rated similarity and association scores are freely available in JWSAN. The data can be accessed at <http://www.utm.inf.uec.ac.jp/JWSAN/en/>, downloadable as csv files. Sheets are sorted by the number of word pairs (pairID). They include the word pair (word1 and word2), part of speech (POS), mean similarity rating (similarity), mean association rating (association), and number of participants who rated similarity and association of a word pair (n_sim and n_asso).

Declarations Conflict of interest All authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as

you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix A: Similarity rating condition (translated from Japanese)

Please read the explanation below and assign a *similarity* rating for each word pair strictly as per the instruction. There are approximately 100 pairs.

Explanation

In this study, it is critical to make a clear distinction between “similarity” and “association.”

First, we illustrate “similarity.” We regard two synonyms as an example of a similar word pair. The below given word pairs are synonyms.

cup mug.

hammer kanaduti (*hammer made from metal in Japanese).

envy jealousy.

There are significantly similar word pairs that are not synonymous. The following word pairs are significantly similar word pairs. We call these almost synonymous pairs.

shiba-dog akita-dog (*Both are names of dog breeds in Japanese).

love affection.

frog toad.

Next, we illustrate “association.” The following word pairs are “associated” but not “similar.” That is, the two words in the following word pairs mean completely different things.

car tire.

car highway.

car accident.

Car and *tire* are associated because tires are integral parts of cars. However, they are not similar because tires are not cars themselves. *Car* and *highway* are associated because cars run on highways. However, a car as a vehicle is not similar to a highway as a road. *Car* and *accident* are associated because cars sometimes cause accidents. However, a car is an object and is not similar to the accident as a phenomenon.

Note that words that are “associated” with each other are not necessarily “similar.”

If you are not confident in this, go back to examples of synonyms (hammer kanaduti) and.

consider how two words are similar (or different) in meaning.

Example

The following examples further illustrate the difference between similarity and association.

The three pairs below are all associated. However, there is only one pair in which the two words are similar. Please select a similar word pair.

bread butter.

bread toast (correct).

bread mold.

Please re-read the EXPLANATION page if you do not understand why “bread toast” is the correct answer.

Have you read the entire explanation above?

Yes, I have. (*Participants must have checked a checkbox inserted next to this sentence.)

Validation test

The following section illustrates the tests for validation. The three pairs given below are all associated. However, there is only one pair in which the two words are similar.

Please select a similar word pair.

1 cutter paper.

2 cutter frying pan.

3 cutter knife.

Your answer is “(*Chosen number).”

Explanation for the correct answer

Option 3 is the correct answer.

Option 1: “cutter” and “paper” are associated in terms of the frequency because we cut paper using a cutter. However, “cutter” and “paper” are not similar. The two are different in terms of their function and form.

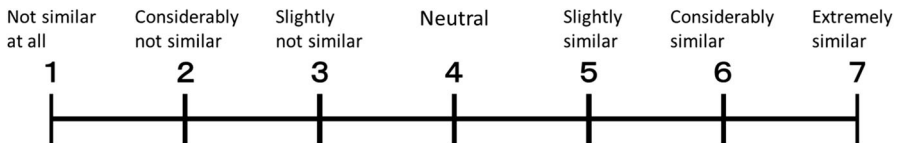
Option 2: “cutter” and “frying pan” are associated because both are tools that humans use. However, both are not similar in terms of their form differences.

Option 3: “cutter” and “knife” are both made of solid material, have forms that are long and thin, and have sharp parts that are able to cut. Words “cutter” and “knife” are more similar than different.

Please read the EXPLANATION section again and make sure that you understand the differences between “similar” and “association” if you selected incorrect answer. Next, rate how “similar” are the word pairs on the next page.

Instructions

Please rate how similar are the given word pairs. Alternatives range from “Not similar at all (1)” to “Extremely similar (7).” You are required to select the numbers that best matches your answer.



For example, “toad” is considerably similar to “frog” because toad is a type of frog. Therefore, you should select 6 (Considerably similar) for the frog–toad pair. In contrast, “scholar” is not similar at all to “study” because scholar is a human and study is not a human, although both are strongly associated. Therefore, you should select 1 (Not similar at all) for the “scholar–study” pair.

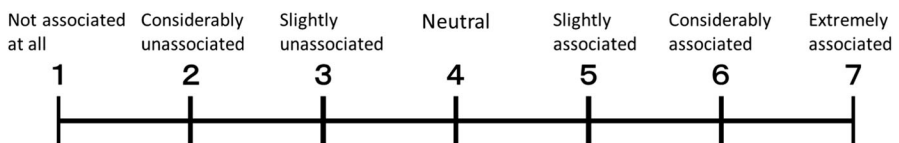
There is no correct answer for these ratings. It is best to go with your hunch. Alternatively, please go with your feeling as a native Japanese speaker. Please remember these things, particularly when you rate dissimilar word pairs.

Appendix B: Association rating condition (translated from Japanese)

Please read the explanation below and assign an *association* rating for each word pair strictly as per the instruction. There are approximately 100 pairs.

Explanation

Please rate how associated are the given word pairs. Alternatives range from “Not associated at all (1)” to “Extremely associated (7).” You are required to select the numbers that best matches your answer.



For example, because a “Scholar” is likely to buy a “Book” frequently, the “Scholars–Book” pair should have a relevance of 6 (Considerably associated).

Although it is not impossible for “Minced meat” and “magnet” to find match points because they are used by humans and are not very expensive, the degree of their association is considered to be low. Thus, the degree of association of “Minced meat–Magnet” should be 2 (Considerably unassociated).

There is no correct answer for these ratings. It is best to go with your hunch. Alternatively, please go with your feeling as a native Japanese speaker. Please remember these things, especially when you rate word pairs that are not associated at all.

Have you read the entire explanation above?

Yes, I have. (*Participants must check a checkbox next to this sentence.)

References

- Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009* (pp. 19–27): Association for Computational Linguistics
- Anderson, A. J., Kiela, D., Clark, S., & Poesio, M. (2017). Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5, 17–30. https://doi.org/10.1162/tacl_a_00043
- Baker, S., Reichart, R., & Korhonen, A. (2014). An unsupervised model for instance level subcategorization acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014* (pp. 278–289): Association for Computational Linguistics. doi:<https://doi.org/10.3115/v1/D14-1034>.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.
- Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1), 13–47.
- Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3), 510–526.
- De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2016a). Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General*, 145(9), 1228.
- De Deyne, S., Peirsman, Y., & Storms, G. (2009). Sources of semantic similarity. In *Proceedings of the Annual Meeting of the Cognitive Science Society, 2009* (Vol. 31, No. 31).
- De Deyne, S., Perfors, A., & Navarro, D. J. (2016). Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016b* (pp. 1861–1870)
- Den, Y., Ogiso, T., Ogura, H., Yamada, A., Minematsu, N., Uchimoto, S., et al. (2007). Ko-pasu nihongogaku no tameno gengoshigen: Keitaisokaiseikiyoudennshikazisho no kaihatu to sono ouyou. *Nihongo Kagaku*, 22, 101–123.
- Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., et al. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1), 116–131. <https://doi.org/10.1145/503104.503110>
- Gerz, D., Vulic, I., Hill, F., Reichart, R., & Korhonen, A. (2016). SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of EMNLP*, (pp. 2173–2182). doi:<https://doi.org/10.18653/v1/D16-1235>.
- Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*. https://doi.org/10.1162/COLI_a_00237

- Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. Paper presented at the *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, Jeju Island, Korea.
- Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., & Kanzaki, K. (2008). Development of the Japanese WordNet. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, (pp. 2420–2423).
- Jarmasz, M., & Szpakowicz, S. S. (2003). Roget's thesaurus and semantic similarity. In *Proceedings of the RANLP-2003, 2003*: Citeseer
- Jones, M. N., Willits, J., & Dennis, S. (2015). Models of semantic memory. *The Oxford handbook of computational and mathematical psychology* (pp. 232–254). Oxford University Press.
- Karpinska, M., Li, B., Rogers, A., & Drozdz, A. (2018). Subcharacter information in Japanese embeddings: When is it worth it? In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP, 2018* (pp. 28–37).
- Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, 4(1), 151–171. <https://doi.org/10.1146/annurev-linguistics-030514-125254>
- Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3, 211–225. https://doi.org/10.1162/tacl_a_00134
- Luong, T., Socher, R., & Manning, C. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning* (pp. 104–113): Association for Computational Linguistics.
- Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., et al. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48, 345–371.
- Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92, 57–78. <https://doi.org/10.1016/j.jml.2016.04.001>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference of Learning Representations*.
- Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1–28. <https://doi.org/10.1080/01690969108406936>
- Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., et al. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880), 1191–1195. <https://doi.org/10.3758/BF03195588>
- National Institute for Japanese Language and Linguistics. (2011). *Modern Japanese balanced written language corpus manual (ver.1.1)*. Tachikawa: National Institute for Japanese Language and Linguistics.
- Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3), 402–407.
- Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014* (pp. 1532–1543): Association for Computational Linguistics
- Pilehvar, M. T., Kartsaklis, D., Prokhorov, V., & Collier, N. (2018). Card-660: Cambridge rare word dataset - a reliable benchmark for infrequent word representation models. In *Proceedings of EMNLP*, (pp. 1391–1401). doi:<https://doi.org/10.18653/v1/D18-1169>.
- Resnik, P. (1995). Using information content to evaluate semantic similarity. In *Proc. 14th International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada*, (pp. 448–453).
- Rothe, S., & Schütze, H. (2017). Autoextend: Combining word embeddings with semantic resources. *Computational Linguistics*, 43(3), 593–617. https://doi.org/10.1162/COLI_a_00294
- Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10), 627–633. <https://doi.org/10.1145/365628.365657>
- Sakazawa, Y., & Komachi, M. (2018). Construction of a Japanese word similarity dataset. *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC 2018)*.
- Scheible, S., Im Walde, S. S., & Springorum, S. (2013). Uncovering distributional differences between synonyms and antonyms in a word space model. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing, 2013* (pp. 489–497)

- Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 141–188.
- Utsumi, A. (2011). Computational exploration of metaphor comprehension processes using a semantic space model. *Cognitive Science*, 35(2), 251–296. <https://doi.org/10.1111/j.1551-6709.2010.01144.x>
- Vulic, I., Baker, S., Ponti, E. M., Petti, U., Leviant, I., Wing, K., & Majewska, O. (2020). Multi-SimLex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity. *Computational Linguistics*, 46(4), 1–51.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.