Proceedings of the 15th Workshop on Building and Using Comparable Corpora (BUCC 2022) @LREC2022 , pages 1–7
Marseille, June 25, 2022

Evaluating Monolingual and Crosslingual Embeddings on Datasets of Word Association Norms

Trina Kwong[1], Emmanuele Chersoni[2], Rong Xiang[2]
King George V School[1], The Hong Kong Polytechnic University[2]
King George V School, Ho Man Tin, Kowloon, Hong Kong [1]
The Hong Kong Polytechnic University, Yuk Choi Road 11, Hung Hom, Kowloon, Hong Kong [2]
{trinakwong,emmanuelechersoni,xiangrong0302}@gmail.com

Abstract

In *free word association* tasks, human subjects are presented with a stimulus word and are then asked to name the first word (the response word) that comes up to their mind. Those associations, presumably learned on the basis of conceptual contiguity or similarity, have attracted for a long time the attention of researchers in linguistics and cognitive psychology, since they are considered as clues about the internal organization of the lexical knowledge in the semantic memory.
Word associations data have also been used to assess the performance of Vector Space Models for English, but evaluations for other languages have been relatively rare so far. In this paper, we introduce word associations datasets for Italian, Spanish and Mandarin Chinese by extracting data from the Small World of Words project, and we propose two different tasks inspired by the previous literature. We tested both monolingual and crosslingual word embeddings on the new datasets, showing that they perform similarly in the evaluation tasks.

Keywords: Word Associations, Distributional Semantic Models, Crosslingual Embeddings

1. Introduction

With the expression "semantic memory", linguists and psychologists tend to refer to the people's memory for conceptual and linguistic meanings, and the way in which this knowledge is encoded and organized has always been a common point of interest. A commonly used metaphor is that of a network, where nodes represent words and the lines linking them are the connections between those words (Fitzpatrick, 2012). When it comes to the specific problem of the organization of word meanings, the procedure known as *word association norms* is probably the most typical mean of investigation: a stimulus word is presented to a human participant, who is simply required to produce the first word coming to mind (McRae et al., 2012). Most authors agree that word associations are learned by contiguity (Church and Hanks, 1990; Wettler et al., 2005; Rapp, 2014), and that they play a fundamental role in language learning (McRae et al., 2012). Some of the modern theories of linguistic and conceptual processing even assume that they capture most of the semantic representations in the language system (Barsalou et al., 2008; De Deyne and Storms, 2008).

One of the strongest paradigm in computational semantics research, on the other hand, has been focusing on the representation of words as distributional vectors, and on the assessment of their semantic similarity on the basis of the similarity of the linguistic patterns of co-occurrence, extracted from large scale textual corpora (Turney and Pantel, 2010; Lenci, 2018). Given the success of *Vector Space Models* (henceforth VSMs) such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), researchers in cognitive science successfully tested them on a variety of psycholinguistic tasks, including the prediction of word associates (Mandera et al., 2017; Nematzadeh et al., 2017), the modeling of human-elicited cloze completion of sentences (Hofmann et al., 2017) and of association ratings (Hofmann et al., 2018). Interestingly, VSMs that are trained directly on word associations have been shown to outperform those trained on textual corpora in predicting human similarity and relatedness judgements, suggesting that such associations are providing a more accurate reflection of the structure of the mental lexicon (De Deyne et al., 2016). Although new benchmarks for modeling word associations with VSMs have recently been introduced (Evert and Lapesa, 2021), however, this kind of evaluation task has almost always been done in English, also because of the lack of similar word association datasets for other languages.

In this paper, we describe the creation of three comparable word association datasets for Italian, Spanish and Mandarin Chinese, which were manually compiled by extracting the data from the interface of the website of the Small World of Words project (De Deyne et al., 2019), and we propose a first evaluation with word embedding models. [1] In addition to monolingual word embeddings for each language, we also used *crosslingual embeddings* (Ruder et al., 2019) that represent the lexicon of two or more languages in the same semantic space. Our results show some differences between languages, but in general the crosslingual embeddings perform comparably to monolingual ones. [2]

[1] https://smallworldofwords.org/en/project/explore

[2] The datasets described in this work will be available upon publication. For more information, contact emmanuelechersoni@gmail.com.

2. Related Work

2.1. Free Association Data for VSMs Evaluation

Using free association data, two types of evaluation tasks can be designed: in the *forward association* task, a model is given a stimulus word and it has to produce the first associate (*lucky* $\rightarrow$? *fox* $\rightarrow$?), while in the *backward* or *reverse* association task the model is presented with one or more response words, and it has to identify the originary stimulus (for example, it would have to guess that *cloud, pizza, drug, kingdom* and *chewy* are responses to the stimulus *mushroom*). The evaluation in the first type of task is typically challenging, since there is a high amount of variation in word production (Rapp, 2008; Rapp, 2014) and the model would have to pick the right answer out of thousands of possible alternatives (Evert and Lapesa, 2021). Some tasks based on the forward associations of the Edinburgh Associative Thesaurus (Kiss et al., 1973) were introduced first in the ESSLLI 2008 Workshop on Lexical Semantics (Baroni et al., 2008). Among the others, the authors proposed a multiple choice discrimination task: given tuples composed by a cue word, a first associate, a hapax associate (e.g. a response produced only once for a given stimulus) and a random associate, a VSM had to assign a higher similarity score to the cue-first pair. They also introduced the more challenging open-vocabulary access task where, for each cue word, a VSM had to retrieve the first associate from an open set of possible response words.

More recently, a much larger free associations dataset for word embeddings evaluation in English has been created by Evert and Lapesa (2021), with more than 12000 association tuples extracted from the Edinburgh Associative Thesaurus and from the Southern Florida Association Norms (Nelson et al., 2004). Evert and Lapesa proposed a multiple choice task and an open vocabulary access task, similarly to Baroni et al. (2008), and reported that in the former the best performance is obtained by first-order models, based on collocations, while VSMs do better in the latter one. The two types of models seemed to have complementary strengths since their combination further improved the global accuracy scores.

As for *reverse associations*, a task example was instead proposed by Rapp (2014): given a list of response words, a system has to predict the stimulus word leading to their production. Reverse association can be also seen as related to lexical access issues, such as the so-called *tip-of-the-tongue* problem, when a person cannot recall a particular word but can still think to its features and associates (Zock and Bilac, 2004; Zock et al., 2010; Zock and Schwab, 2011). Moreover, as suggested by Zock (2002), an automatic tool that is able to efficiently retrieve a target word from its associates could be potentially very useful for navigating lexical resources. Following Rapp's proposal, the CogALex Shared Task 2014 (Rapp and Zock, 2014) introduced an evaluation dataset of responses and stimuli, also based on the Edinburgh Associative Thesaurus. The best results were reported by Ghosh et al. (2014), who used vector similarities from a Word2Vec vector space (Mikolov et al., 2013) to generate cue candidates and then ranked them on the basis of Pointwise Mutual Information scores (Church and Hanks, 1990).

2.2. Predicting Norms with Word Embeddings

A common criticism of VSMs is that, as a semantic representation, they are not grounded in perception and word meanings are only defined in relation to each other (Glenberg and Robertson, 2000; Fagarasan et al., 2015). Several works, for this reason, proposed to map word embedding features onto interpretable norms of different types via regression or neural network methods, e.g. conceptual (Fagarasan et al., 2015; Li and Summers-Stay, 2019), modality exclusivity (Chersoni et al., 2020) or neurocognitive norms (Utsumi, 2018; Utsumi, 2020; Chersoni et al., 2021).

Chersoni et al. (2020) recently reported that norms for a new language can be decently predicted with a machine learning classifier trained on English norms and *crosslingual word embeddings*, a kind of VSM that represents the lexicon of two or more languages in the same semantic space. This kind of crosslingual prediction could be an interesting application for psycholinguistic research relying on norms, as norms are generally available only for a few languages other than English and their collection is typically time-consuming. Being able to automatically predict norms for under-resourced languages via crosslingual transfer, on the other hand, would certainly represent a big advantage.

In our work, we decided to test also crosslingual word embeddings in word association tasks, to assess to what extent word association knowledge can be modeled with multilingual semantic spaces. According to some previous studies (Brainerd et al., 2008; McRae et al., 2012), word associations are to be understood in terms of semantic relations, and those relations could be at least partially shared across languages. However, it should also be considered that responses to a cue might depend on language-specific patterns, and such cases are expected to be more challenging for models aligning multiple languages in the same semantic space. [3]

3. Experimental Settings

3.1. Dataset Creation

For the Italian, Spanish and Mandarin Chinese datasets, we manually collected word associations data by querying the `https://smallworldofwords.org/en/project/explore`, as it contains data for many different languages and words that are filtered by a minimum frequency threshold. Each dataset includes 300 stimuli words. At the beginning, we tried

[3] In this work, the expressions *Vector Space Models* (VSMs) and *word embeddings* are used interchangeably.

to select the 300 words of the original ESSLLI 2008 dataset (Baroni et al., 2008) and to translate them in the other languages; however, we found out that the coverage was low, i.e. different stimuli have been used for different languages. Therefore, we just selected the stimuli for each language by using the random selection function of the project interface. It should also be noticed that the Small World of Words is an ongoing project, and new data gets continuously added: the datasets described here refer to the status of the collection as of December 2021.

For each stimulus, we generated a tuple of words <FIRST HIGHER RANDOM>, where:

- FIRST is the first associate word, the one that was produced more frequently as a response to a stimulus word;
- HIGHER is a higher-rank associate word, i.e. a word that is not the first but the *n*-th in a rank based on the decreasing number of subjects that produced it as a response. This word will still be related to the stimulus, but is likely to reflect a weaker association strength. For all datasets, we always sampled HIGHER words with the minimum frequency that was available on the Small World of Words website for the given stimulus, i.e. 2 for most words, meaning that all those words have been produced by at least 2 subjects;
- RANDOM was a word that was randomly picked out of the pool of the first associates of the other stimuli in the same language. The sampling was carried out by using the Python RANDOM package, and the same word could have been sampled multiple times.

Examples of the generated tuples for each language can be seen in Table 1.

The words in the Small World of Words interface are not lemmatized, and therefore the frequencies are split over morphologically-related forms. For our datasets, we considered the unlemmatized forms, that is, the frequencies were kept separate for different morphological forms of the same word.

Finally, for each tuple we added an association score between the stimulus and the FIRST associate, to be used for extra analysis. Following Baroni et al. (2008), this score was computed by taking the number of the responses for the FIRST associate of a given stimulus and dividing it by the total number of responses for that stimulus. For example, if a FIRST associate has been produced 5 times out of 10 responses, the association score for the tuple will be 0.5. This score could be eventually used to design other evaluation tasks, for example by assessing the correlation between the association and similarity scores produced by a word embedding model.

A noticeable difference between our datasets and the previous ones is represented by the HIGHER associates. In the works by Baroni et al. (2008) and Evert and Lapesa (2021), the tuples contained HAPAX associates, i.e. words that were produced only once as a response to the stimulus. However, the Small World of Words website does not include such responses, as the minimum frequency is 2. Evert and Lapesa (2021) used the HAPAX associates as distractors, in order to make the task more challenging for VSMs. Since our HIGHER associates have been produced more than one subjects, we expected them to have a higher association strength with the original stimulus, and thus, to be more difficult to discriminate from the FIRST associates.

Lang	Stimulus	First	Higher	Random
ITA	linea (line)	retta (straight)	lunga (long)	prete (priest)
SPA	bueno (good)	malo (bad)	dulce (sweet)	verde (green)
ZH	活 (live)	死 (die)	人生 (life)	人才 (talent, talented person)

Table 1: Examples of the tuples for each language

Model	Corpus	Type
FastText Wiki	Wikipedia	Monolingual
FastText WikiAlign	Wikipedia	Crosslingual (2 languages)
Numberbatch	ConceptNet, Word2Vec, Glove OpenSubtitles 2016	Multilingual (78 languages)

Table 2: Summary of word embedding types.

3.2. VSMs

For each language, we used three 300-dimensional off-the-shelf word embedding models, which are summarized in Table 2. One of them is a monolingual model, i.e. the publicly available FastText vectors (Bojanowski et al., 2017; Grave et al., 2018) trained with a Skip Gram model on Wikipedia (*FastText Wiki*). [4], [5]

Together with the monolingual *FastText Wiki* vectors, we also tested the crosslingual vectors of *FastText WikiAlign* (Joulin et al., 2018). In the *FastText WikiAlign* models [6], the embeddings of English a source language have been aligned to the embeddings of a target language, using a mapping function that minimizes the distances between words that are reciprocal translations, and maximizes the margin between correct translations and other candidate words.

Finally, we also experimented with the multilingual *Numberbatch* embeddings (Speer and Lowry-Duda, 2017), which are obtained by retrofitting different types of word embeddings with a subgraph of ConceptNet (Speer et al., 2017). We used the more recent release

[4] https://fasttext.cc/docs/en/pretrained-vectors.html

[5] All the hyperparameters are the default ones of the Word2Vec package, see Mikolov et al. (2013) for details.

[6] https://fasttext.cc/docs/en/aligned-vectors.html

of Numberbatch, where the sources of the retrofitted embeddings are Word2Vec, GloVe and the OpenSubtitles2016 corpus (Lison and Tiedemann, 2016).

3.3. Tasks and Metrics

For the evaluation tasks, we follow the design of the two tasks proposed by the previous literature (Baroni et al., 2008; Evert and Lapesa, 2021). In the **multiple choice task**, given a stimulus and a tuple <FIRST HIGHER RANDOM>, the word embedding model should be able to determine which one of the words in the tuple is the FIRST associate. For each embedding space, we simply compute the cosine similarity of the stimulus vector and the vectors of the three words in the tuple, and assign a hit whenever the similarity score with the FIRST word is the highest. Performance is assessed using the standard *Accuracy* metric.

In the **open-vocabulary access task**, for each stimulus in the dataset, a word embedding model has to retrieve the right FIRST associate out of a list of candidates including all the other FIRST associates in the dataset (e.g. for each language, there will be around 300 candidates). For each stimulus, we measure the cosine similarity with all the other FIRST associates in the dataset and we compile a ranking based on decreasing similarity values. We then assess the performance with the following metrics:

- *Top-N Accuracy*: we assign a hit whenever the right FIRST associate for a stimulus is in the top-N of the rank. We reported Accuracy values for $N = 1, 5, 10$;
- *Mean Rank*: we compute the average rank of the right FIRST associate for each stimulus (see Equation 1). For $rank_i$, we use directly the index of instance i if the right FIRST associate is in the top 10 of the rank, and 10 otherwise.

$$MeanRank = \frac{1}{n} * \sum_{i=1}^{n} rank_i \quad (1)$$

Notice that for the latter metric, the lower the score the better, as we want a model to push the right FIRST associates at rank 1 (or as close as possible).

4. Results and Analysis

Table 3 reports the scores for the Multiple Choice Task in the three target languages. The models have full or almost full coverage for the Spanish and Italian dataset, while the Wikipedia-based models for Chinese have several missing words.

For the two European languages, it can be noticed that Wikipedia-based models are the better performing ones, with the monolingual and the crosslingual model achieving similar accuracy scores. As for the Chinese dataset, the situation is reversed: Numberbatch is the model achieving the highest Accuracy scores, and it also shows a better coverage of the dataset vocabulary.

The scores might not seem particularly high, especially in comparison with previous evaluation of this task on English data (Evert and Lapesa, 2021), but besides the limit of our evaluation (e.g. we are also testing VSMs, but no first-order models based on collocations), it should also be considered that our HIGHER distractors are likely to be much more difficult to disentangle from FIRST associates than the HAPAX words of the previous datasets. The reason is that the HAPAX words were associates being produced only by one subject in response to a stimulus, while our HIGHER associates have been produced by two or more subjects, and thus they are likely to reflect less sporadic associations in the mental lexicon. A partial proof of this can be seen in Table 4, 5 and 6, which report, for each dataset, the number of highest cosine scores per condition. At a glance, it is clear that for all models the stimulus-FIRST pair has the highest number of highest cosine scores, but HIGHER words are efficient distractors, leading to a consistent number of errors.

To assess how good the models are at discriminating between the three conditions, we also ran a Kruskal-Wallis test by means of the the R statistical software.The scores for all models show significant differences by condition ($p < 0.001$). We then ran Wilcoxon tests with Bonferroni correction for the pairwise comparisons, and we found strongly significant differences ($p < 0.001$) for almost all of them, with just a small exception, i.e. a weaker effect ($p < 0.05$) for FastText-Wiki for Chinese.

The results of the Open-Vocabulary Access Task for each language can be seen in Tables 7, 8 and 9. They follow similar patterns: for Italian and Spanish, FastText-Wiki and FastText-WikiAlign are the best models in terms of Top1-Accuracy and MeanRank, with the Numberbatch model clearly lagging behind. It should also be mentioned that the Numberbatch model has generally low scores for Top-1 Accuracy, but on the other hand is quite consistent across languages in retrieving the FIRST candidate in the first 5-10 rank positions, and thus its scores for MeanRank, Top-5 and Top-10 Accuracy are closer to the other models.

As for Chinese, Numberbatch is the best model for all metrics, except for the Top-1 Accuracy, where it is topped by FastText WikiAlign. Interestingly, both crosslingual models achieve higher scores on the Chinese data.

A general observation can be made: the crosslingual embeddings are always competitive with the monolingual ones, or even slightly better. In Task 1, FastText-WikiAlign even achieves the top score for Spanish, and in Task 2 the crosslingual models outperform the monolingual models for all metrics on Italian and Spanish. Chinese was expected to be more difficult, as it is a more typologically distant language from English than Spanish and Italian are. However, the Numberbatch embeddings still do better than the monolingual model in all metrics.

Task 1	Italian		Spanish		Chinese	
	Accuracy	Missing Words/ Vocab Size	Accuracy	Missing Words/ Vocab Size	Accuracy	Missing Words/ Vocab Size
FastText-Wiki	**0.723**	0/718	0.657	1/789	0.590	65/809
FastText-WikiAlign	0.717	0/718	**0.673**	1/789	0.630	65/809
Numberbatch	0.623	0/718	0.647	2/789	**0.670**	23/809

Table 3: Results for the Multiple Choice Task in terms of Accuracy for all languages (the top scores are **in bold**). Missing words and vocabulary size are also reported.

	FIRST	HIGHER	RANDOM
FastText-Wiki	216	82	2
FastText-WikiAlign	214	83	3
Numberbatch	187	88	25

Table 4: Number of highest similarity scores with the stimulus in Task 1 for Italian.

	FIRST	HIGHER	RANDOM
FastText-Wiki	193	94	13
FastText-WikiAlign	198	88	14
Numberbatch	194	88	18

Table 5: Number of highest similarity scores with the stimulus in Task 1 for Spanish.

	FIRST	HIGHER	RANDOM
FastText-Wiki	177	97	26
FastText-WikiAlign	189	90	21
Numberbatch	201	95	4

Table 6: Number of highest similarity scores with the stimulus in Task 1 for Chinese.

Task2-Italian	Acc1	Acc5	Acc10	Mean Rank
FastText-Wiki	0.257	**0.533**	**0.630**	5.577
FastText-WikiAlign	**0.263**	**0.533**	**0.630**	**5.523**
Numberbatch	0.120	0.423	0.520	6.580

Table 7: Results for the open-vocabulary task for Italian. Top-N Accuracy for $N = 1, 5, 10$ and Mean Rank are reported (for the latter metric, the lower the better).

Task2-Spanish	Acc1	Acc5	Acc10	Mean Rank
FastText-Wiki	0.268	0.482	0.572	5.823
FastText-WikiAlign	**0.281**	**0.528**	**0.609**	**5.569**
Numberbatch	0.144	0.475	0.548	6.204

Table 8: Results for the open-vocabulary task for Spanish. Top-N Accuracy for $N = 1, 5, 10$ and Mean Rank are reported (for the latter metric, the lower the better).

Also because of the small size of the datasets, the performance differences between models are not significantly different. However, we still think our results can be taken as preliminary evidence that the alignment of embeddings in multilingual spaces does not detract too much from their ability of modeling word associations data in the target language.

Task2-Chinese	Acc1	Acc5	Acc10	Mean Rank
FastText-Wiki	0.170	0.253	0.317	7.757
FastText-WikiAlign	**0.203**	0.347	0.413	7.050
Numberbatch	0.183	**0.537**	**0.617**	**5.657**

Table 9: Results for the open-vocabulary task for Mandarin Chinese. Top-N Accuracy for $N = 1, 5, 10$ and Mean Rank are reported (for the latter metric, the lower the better).

5. Conclusions

In this paper, we have presented an evaluation of Vector Space Models on word associations tasks for languages other than English, after generating three new datasets for Italian, Spanish and Mandarin Chinese from the association data of the Small World of Words project.

Inspired by the previous literature, we tested Vector Space Models on a multiple choice task and on the more challenging open-vocabulary access task. We have included both monolingual and crosslingual embeddings in the evaluation, and we observed that they perform comparably, and in many settings the crosslingual model even do slightly better than their monolingual competitors. We plan to release the three datasets upon publication, in order to encourage further research on the topic.

Our finding might have interesting future applications, such as the automatic prediction of norms for other languages, using multilingual embedding spaces and/or supervised training based on data from high-resource languages (Chersoni et al., 2020). Another necessary step will be to increase the size of the data collections and to include more new languages in the word association evaluation.

Acknowledgements

The current project was carried out during Trina Kwong's internship at the Hong Kong Polytechnic University, under the program "Linguistic Training and Internship for Gifted Students" (PJTD).

The authors would like to thank Simon De Deyne for the availability to share the data extracted from the Small World of Words project.

6. Bibliographical References

Baroni, M., Evert, S., and Lenci, A. (2008). Lexical Semantics: Bridging the Gap between Semantic Theory and Computational Simulation. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*.

Barsalou, L. W., Santos, A., Simmons, W. K., and Wilson, C. D. (2008). Language and Simulation in Conceptual Processing. *Symbols, Embodiment, and Meaning*, pages 245–283.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brainerd, C. J., Yang, Y., Reyna, V. F., Howe, M. L., and Mills, B. A. (2008). Semantic Processing in "Associative" False Memory. *Psychonomic Bulletin & Review*, 15(6):1035–1053.

Chersoni, E., Xiang, R., Lu, Q., and Huang, C.-R. (2020). Automatic Learning of Modality Exclusivity Norms with Crosslingual Word Embeddings. In *Proceedings of *SEM*.

Chersoni, E., Santus, E., Huang, C.-R., and Lenci, A. (2021). Decoding Word Embeddings with Brain-Based Semantic Features. *Computational Linguistics*, 47(3):663–698.

Church, K. W. and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

De Deyne, S. and Storms, G. (2008). Word Associations: Network and Semantic Properties. *Behavior Research Methods*, 40(1):213–231.

De Deyne, S., Perfors, A., and Navarro, D. J. (2016). Predicting Human Similarity Judgments with Distributional Models: The Value of Word Associations. In *Proceedings of COLING*.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., and Storms, G. (2019). The "Small World of Words" English Word Association Norms for over 12,000 Cue Words. *Behavior Research Methods*, 51(3):987–1006.

Evert, S. and Lapesa, G. (2021). FAST: A Carefully Sampled and Cognitively Motivated Dataset for Distributional Semantic Evaluation. In *Proceedings of CONLL*.

Fagarasan, L., Vecchi, E. M., and Clark, S. (2015). From Distributional Semantics to Feature Norms: Grounding Semantic Models in Human Perceptual Data. In *Proceedings of IWCS*.

Fitzpatrick, T. (2012). Word Associations. *The Encyclopedia of Applied Linguistics*.

Ghosh, U., Jain, S., and Paul, S. (2014). A Two-stage Approach for Computing Associative Responses to a Set of Stimulus Words. In *Proceedings of the COLING Workshop on Cognitive Aspects of the Lexicon (CogALex)*.

Glenberg, A. M. and Robertson, D. A. (2000). Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning. *Journal of Memory and Language*, 43(3):379–401.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of LREC*.

Hofmann, M. J., Biemann, C., and Remus, S. (2017). Benchmarking N-Grams, Topic Models and Recurrent Neural Networks by Cloze Completions, EEGs and Eye Movements. In *Cognitive Approach to Natural Language Processing*, pages 197–215. Elsevier.

Hofmann, M. J., Biemann, C., Westbury, C., Murusidze, M., Conrad, M., and Jacobs, A. M. (2018). Simple Co-Occurrence Statistics Reproducibly Predict Association Ratings. *Cognitive Science*, 42(7):2287–2312.

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., and Grave, E. (2018). Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. In *Proceedings of EMNLP*.

Kiss, G., Armstrong, C., Milroy, R., and Piper, J. R. I. (1973). An Associative Thesaurus of English and Its Computer Analysis. pages 153–165. Edinburgh University Press.

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4:151–171.

Li, D. and Summers-Stay, D. (2019). Mapping Distributional Semantics to Property Norms with Deep Neural Networks. *Big Data and Cognitive Computing*, 3(2):30.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of LREC*.

Mandera, P., Keuleers, E., and 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.

McRae, K., Khalkhali, S., and Hare, M. (2012). Semantic and Associative Relations in Adolescents and Young Adults: Examining a Tenuous Dichotomy. *Psychology Publications*, 115.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Nelson, D. L., McEvoy, C. L., and 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.

Nematzadeh, A., Meylan, S. C., and Griffiths, T. L. (2017). Evaluating Vector-Space Models of Word Representation, or, the Unreasonable Effectiveness of Counting Words Near Other Words. In *Proceedings of CogSci*.

Pennington, J., Socher, R., and Manning, C. (2014).

Glove: Global Vectors for Word Representation. In *Proceedings of EMNLP*.

Rapp, R. and Zock, M. (2014). The CogALex-IV Shared Task on the Lexical Access Problem. In *Proceedings of the COLING Workshop on Cognitive Aspects of the Lexicon*.

Rapp, R. (2008). The Computation of Associative Responses to Multiword Stimuli. In *Proceedings of the COLING Workshop on Cognitive Aspects of the Lexicon*.

Rapp, R. (2014). Corpus-Based Computation of Reverse Associations. In *Proceedings of LREC*.

Ruder, S., Vulić, I., and Søgaard, A. (2019). A Survey of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, 65:569–631.

Speer, R. and Lowry-Duda, J. (2017). ConceptNet at SemEval-2017 Task 2: Extending Word Embeddings with Multilingual Relational Knowledge. In *Proceedings of SemEval*.

Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of AAAI*.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Utsumi, A. (2018). A Neurobiologically Motivated Analysis of Distributional Semantic Models. In *Proceedings of CogSci*.

Utsumi, A. (2020). Exploring What Is Encoded in Distributional Word Vectors: A Neurobiologically Motivated Analysis. *Cognitive Science*, 44(6):e12844.

Wettler, M., Rapp, R., and Sedlmeier, P. (2005). Free Word Associations Correspond to Contiguities Between Words in Texts. *Journal of Quantitative Linguistics*, 12(2-3):111–122.

Zock, M. and Bilac, S. (2004). Word Lookup on the Basis of Associations: From an Idea to a Roadmap. In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*.

Zock, M. and Schwab, D. (2011). Storage Does Not Guarantee Access. The Problem of Organizing and Accessing Words in a Speaker's Lexicon. *Journal of Cognitive Science*, 12(3):233–259.

Zock, M., Ferret, O., and Schwab, D. (2010). Deliberate Word Access: An Intuition, a Roadmap and Some Preliminary Empirical Results. *International Journal of Speech Technology*, 13(4):201–218.

Zock, M. (2002). Sorry, What Was Your Name Again, or How to Overcome the Tip-of-the-Tongue Problem with the Help of a Computer? In *Proceedings of the Workshop on Building and Using Semantic Networks*.