Semantic vector evaluation and human performance on a new vocabulary MCQ test

Joseph P. Levy (j.levy@roehampton.ac.uk)
Department of Psychology, University of Roehampton
London, UK

John A. Bullinaria (j.a.bullinaria@cs.bham.ac.uk)
School of Computer Science, University of Birmingham,
Birmingham, UK

Samantha McCormick (Samantha.McCormick@roehampton.ac.uk)
Department of Psychology, University of Roehampton
London, UK

Abstract

Vectors derived from patterns of co-occurrence of words in large bodies of text have often been used as representations of some aspects of the meanings of different words. Generally, the distance between such vectors is used as a measure of the semantic similarity between the word meanings they represent. One important way of evaluating the performance of these vectors has been to use them to answer vocabulary multiple choice questions (MCQs) where the participant is asked to judge which of several choice words is closest in meaning to a stem word. The existing vocabulary MCQ tests used in this way have been very useful but there are some practical problems in their use as general evaluation measures. Here, we discuss why such tests remain useful evaluation measures, introduce a new vocabulary test, evaluate several current sets of semantic vectors using the new test and compare their performance to human data.

Keywords: Distributional semantics; vocabulary MCQ.

Introduction

There are many potential applications for a method that can reliably form the basis for measuring the semantic distance between words or concepts. Many methods achieve this by placing each word/concept in a multidimensional space where the dimensions are defined by the way in which words co-occur in corpora of real language use (Schütze, 1993; Bullinaria & Levy, 2007; Turney & Pantel, 2010). The simplest such methods place a target word in a space defined by the count of how many times this word co-occurs with other words in the corpus with each of these context words defining a dimension. The resulting semantic vectors may also be smoothed by some kind of dimensionality reduction technique. Most current techniques retain only a small proportion of the number of initial dimensions (often 300) and refer to these dense vector sets as "word embeddings" (e.g., Mikolov et al., 2013). Two words are then judged to be semantically similar if the vector distance between them is small.

Various validation approaches have been used, but a particularly convenient way of evaluating such techniques is to measure how well they perform in a vocabulary multiple choice question (MCQ) test where a participant is asked to choose which of a number of *choice* words is closest in meaning to a *stem* word (not to be confused with a morphological stem). Often this amounts to choosing a ("key") *synonym,* or the word that is closest to being a synonym, from the other choice words that act as *distractors*. For human participants, these tests are used to measure vocabulary knowledge. Such tests are ideal methods to use to evaluate co-occurrence techniques which construct semantic vectors for each word such that their distances are related to how substitutable the words are for each other. A linguistic intuition would be that if two words are substitutable for each other in everyday language then they are synonyms or at least very closely semantically related.

Landauer and Dumais (1997) used the performance of their LSA (Latent Semantic Analysis) method on the retired items of a once commercially available test of English vocabulary the TOEFL (Teaching of English as a Foreign Language). They reported a performance of 64.4%, well above chance and equivalent to acceptable performance for entry to a US University. This MCQ test has been the most widely used vocabulary test to date for evaluating distributional semantic vectors[1]. Turney (2001) also describes a commercially available test, the ESL, and this has been used to evaluate such methods. Another candidate MCQ test would be the one used by Adelman et al. (2014) from Shipley (1940). This is a 40 item vocabulary MCQ using some now somewhat archaic US English usages. It has the advantage of being freely available in the appendix of a historic journal article. Such tests are valuable evaluation measures for semantic representations in that they are independently designed to measure the performance of human participants. However, they have a number of disadvantages if they are used as the only evaluation:

[1]The ACL Wiki lists the performance of some key approaches: https://www.aclweb.org/aclwiki/index.php?title=TOEFL_Synonym_Questions_(State_of_the_art)

- They are at least relatively commercially sensitive as making a real-word test freely available would render it useless as a fair measure of human vocabulary knowledge. We have always found that researchers who report the use of such tests have been helpful and generous in making the test items available to other researchers but, inevitably, the commercial/practical sensitivity of the items is an obstacle for widespread open publication of them. This can prevent the clear reporting of slight changes made to the items due to low frequency word forms (or USA/UK spelling variants) not appearing, or not appearing frequently enough, in the corpora used, which can lead to papers reporting results of slightly different tests.
- They are relatively small as they are designed to be completed in a reasonably short period of time by the human testees.
- The individual questions may not be uniform in terms of difficulty or kind of semantic relationship being tested. A question may test knowledge of near-synonymy or one of a degree of some other kind of semantic similarity such as category membership.

In this paper, we describe a new 200 item MCQ vocabulary which we will make freely available. The test has been constructed using the lexicographical judgements implicit in the construction of the noun entries in WordNet (Fellbaum, 1998). Half of the stem words in the test are relatively high frequency (in the psycholinguistically relevant sense) and half are low frequency. Word frequency is a dominant lexical variable for human language processing and especially so in instruments, such as this one, that are designed to measure vocabulary knowledge. The 200 MCQ set is large enough to be performed comfortably by human participants and to be split into subsets for training and testing when using machine learning techniques.

As noted above, vocabulary MCQ tests have frequently been used as evaluation measures for distributional semantic vectors. However, some of the most recent methods for generating such semantic vectors (e.g., Mikolov et al., 2013; Pennington et al., 2014) have emphasised evaluation using sets of analogy problems. We would argue that both vocabulary and analogy tests are important in evaluating the semantic *competence* of distributional semantic vectors, as well as being useful in models of human semantic *performance*. Here, we therefore evaluate three promising recent semantic vector techniques using our new vocabulary test.

In constructing our new test, we use WordNet (Fellbaum, 1998), a freely available lexicographic database organised around lists of synonyms (synsets) for the different senses of each word in the database. This allows us to use the independent linguistic judgements of the WordNet team as a standard for competence in tests of synonymy judgements on the vocabulary MCQ items that we construct.

We consider that MCQ vocabulary tests are interesting psychological tasks in their own right. It is likely that word frequency measures will dominate any quantitative model of relative question difficulty and that word familiarity (and proxies for this such as level of education or experience with English in the case of the data described here) is likely to dominate models of individual differences in performance on these tasks. If a participant has never or very rarely come across a stem or synonym then, apart from the possibility of sensitivity to form-meaning symbolism (e.g., Levy & Thompson, 2008; Monaghan et al., 2014), they are unlikely to perform well on test items containing these words. However, there remains the strong possibility that semantic distances between stem and synonym, stem and distractors and synonym and distractors will affect question difficulty and the choice of distractor when an MCQ item is answered incorrectly. Semantic vector techniques are a good resource for calculating these distances. Thus, vocabulary MCQs are useful measures for evaluating the competence of semantic vector techniques, and semantic vectors are likely to be components of any complete model of human vocabulary MCQ performance.

In the rest of this paper, we outline how we constructed a new 200 item vocabulary MCQ test, show how well three recent methods for generating distributional semantics vectors perform on the test, and compare the performance of the semantic vectors with human performance on the same test items. We intend to make the MCQ items and human data freely available as a research resource.

Construction of New Vocabulary Test

We constructed a set of 200 vocabulary MCQ items. This is larger than most of the evaluation benchmarks that have been suggested (allowing the set to be potentially split into independent training and testing components for reliable model selection purposes) but still a manageable number of items for individual human participants.

The words in the MCQs were chosen by using the entries for nouns in WordNet. All words considered for selection appeared in both the SUBTLEX-UK (Van Heuven et al., 2014) and WordNet databases, and were dominantly tagged in their noun form in both databases. SUBTLEX-UK is a database constructed from BBC TV subtitling records and so this ensured that the words chosen were in common usage in the UK.

We chose to generate stem-synonym pairs by using the synsets in WordNet because this gives us an independent benchmark for lexical semantic relations. By dividing the MCQ items into two subsets where one has relatively high stem frequencies and the other has relatively low ones, the vocabulary test controls one of the most important influences on human linguistic performance.

The potential candidate list of stem words was divided into lower-frequency (LF) and higher-frequency (HF) subsets using the "Zipf" scale (van Heuven et al., 2014), which is based on the $\log_{10}$ transform of word frequency. Those authors argue that this scale is a better way to control frequency in a psycholinguistically relevant way than frequency per million word (fpmw) counts. For example, using these counts to select stimuli results in an

underrepresentation of relatively low frequency words that are familiar to human participants.

These candidate lists were randomly sorted. Stems and synonyms (taken from the synsets associated with each stem noun) were selected from this list such that the final HF and LF subsets consisted of pairs that were matched for stem word length, synonym frequency and synonym length. Hyphenated stem words or synonyms were excluded from selection. Three distractor words were selected at random from the remaining nouns in WordNet and SUBTLEX-UK with a Zipf frequency greater than two. The mean distractor length and frequency (over the three distractors) was pairwise matched to the synonym. Mean stem, synonym and distractor characteristics are shown in Table 1.

Table 1: Mean Zipf Frequency (F) and word length (L) for Stem, Synonym (Syn) and Distractor (Distr) words

MCQs	Stem F	Stem L	Syn F	Syn L	Distr F	Distr L
LF	3.0	6.3	3.6	6.4	3.6	6.3
HF	4.8	6.4	3.7	6.5	3.7	6.6

Human Performance on the Vocabulary Test

The vocabulary MCQ test was given to 85 monolingual English speaking undergraduate student participants and 77 non-monolingual students. Their performance is summarised in Table 2.

Table 2: Performance (% correct) of monolingual and non-monolingual participants

MCQs	mean	SD	range
Monolingual			
All 200	79%	10%	56% - 97%
100 LF	72%	13%	46% - 96%
100 HF	86%	8%	63% - 98%
Non-monolingual			
All 200	71%	10%	49% - 92%
100 LF	61%	12%	32% - 88%
100 HF	81%	10%	52% - 96%

Mean performance is high but does not appear to be close to ceiling. The very best performance is close to maximal demonstrating that it is possible for humans to achieve a close to perfect score.

The mean scores for monolingual participants are higher than that for non-monolingual participants for the complete MCQ set and the two subsets.

As would be expected for human performance, performance for the high frequency stems exceeds that for the low frequency ones with the non-monolingual participants demonstrating a larger deficit for low frequency stems. Clearly, human performance has been affected by our manipulation of stem frequency whilst matching the synonym and distractor frequencies between the low- and high-frequency subsets. The correlation between MCQ question item difficulty (as measured by overall percentage correct) and stem SUBTLEX-UK frequency is 0.41, suggesting that factors other than word frequency may be affecting human vocabulary test performance.

Three participants achieved overall scores of 97% - six errors from the 200 MCQ items. These few errors were sometimes made on the same question by more than one participant and were also sometimes also made by the semantic vector methods.

Semantic Vectors for Evaluation

In addition to testing our new MCQ vocabulary test on human participants, we also used it to evaluate three available sets of semantic vectors, all derived from large text corpora but using contrasting methods to capture the patterns of word usage regularities. Our aim here is to illustrate how well a few recent methods that have been most successful on other semantic tasks are able to perform on this task, and not to make any claims about optimal methods or parameters.

We compared the levels of success on the new vocabulary test using three different kinds of semantic vectors that span the range of approaches available: vectors derived from the methods described by Bullinaria & Levy (2012), the publicly available semantic vectors that were generated using one of the word2vec (Mikolov et al., 2013) methods (available at: https://code.google.com/archive/p/word2vec/) and the GloVe (Pennington, Socher & Manning, 2014) vectors derived from the co-occurrence matrix from 6, 42 and 840 billion word corpora available at: http://nlp.stanford.edu/projects/glove/.

The Bullinaria & Levy (B&L) vectors are derived from the ukWaC (Baroni et al., 2009) (2 billion words) corpus by counting word co-occurrences in a context window of size one and using those counts to generate a starting matrix of positive pointwise mutual information (PPMI) values for about 50,000 target words and 50,000 context words. This choice of window size and co-occurrence statistic was previously shown to be optimal for a range of semantic tasks (Bullinaria & Levy, 2007) and is now widely used. Singular Value Decomposition (SVD) dimensionality reduction is then used to generate orthogonal matrices U and V and diagonal singular value matrix S such that $M = USV^T$, and dimensionality reduction is performed by taking the principal components of $US^P = MVS^{P-1}$ to produce semantic vectors with a dimensionality of 5,000 with the components weighted using a Caron (2001) P value of 0.25. These parameter values were optimised so as to perform well on four different semantic evaluation measures including a version of the Landauer & Dumais (1997) TOEFL MCQ vocabulary test, and achieved the current state-of-the art performance on the TOEFL MCQ[1] test.

The word2vec (W2V) vectors were generated using a 100 billion word sample of the Google News dataset. Word2vec uses supervised learning algorithms to train a simple but

very large neural network model to predict either which words will appear in a window around the current word (the context given the current word) or which word will appear given the current context words. There are a number of different methods and parameters that can be varied in what amounts to a family of techniques. We made use of the publicly available vectors which have 300 dimensions.

The GloVe (G6, G42, G840) vectors were extracted from the files linked to on the GloVe website. The G6 vectors were generated from a 6-billion-word corpus derived from Wikipedia. The 42B and 840B vectors were generated from 42 billion and 840 billion word corpora derived from Common Crawl archives (obtained by an automated process of systematically browsing the web). All the GloVe vectors used here have 300 dimensions. The GloVe method uses regression modelling to learn semantic vectors from the non-zero entries of a word co-occurrence matrix such that the dot product between the vectors for a pair of words equals the logarithm of their probability of co-occurrence. Pennington et al. (2013) show that their vectors perform well on the analogy problem set that was also used to evaluate the word2vec methods.

Levy, Goldberg and Dagan (2015) have argued that the three broad semantic vector techniques used here have similar levels of overall performance when appropriately tuned.

17 of the 1,000 words within the 200 MCQs did not appear in the word2vec vector sets due to differences in UK and USA English. For these words, we used the vectors derived from the USA spelling variants. 7 words did not appear in the GloVe vectors derived from their 6 billion word corpus. For these words, we substituted related words that did appear in the vector set. The other semantic vector sets contained all the 1,000 words used in this MCQ vocabulary set.

Clearly, there are a number of differences in the corpora and parameters used for the three methods and so this exercise cannot reliably compare the success of the different methods in general, but serves as a comparison of a number of different off-the-shelf semantic vector sets.

Semantic Vector Performance on the Vocabulary Test

We compared the performance of the five different vector sets on all 200 items and on the LF and HF subsets. Mean performance is summarised in Table 3.

Table 3: Performance (% correct) of the five different vector sets.

MCQs	B&L	W2V	G6	G42	G840
All 200	91.0	87.0	86.5	89.5	92.0
100 LF	93.0	87.0	86.0	92.0	95.0
100 HF	89.0	87.0	87.0	87.0	89.0

All three types of semantic vector perform well but not perfectly. None of them match the competence standard of the WordNet-based MCQ test. The GloVe vectors trained on an 840 billion word corpus comes closest to matching the very best performance of human participants. However, all the vector sets exceed the mean performance of the human participants by large margins.

The LF and HF subsets are distinguished by the SUBTLEX-UK frequencies of their stem words. The frequencies (and word lengths) of synonyms and distractors were matched between subsets. Unsurprisingly, the human participants performed better on the HF subset than on the LF subset, presumably reflecting the association between word frequency and the familiarity that participants had with the stem words. However, in three of the five sets, the semantic vectors performed better on the LF subset than the HF one. Since even the smallest corpus used for generating the semantic vectors was 2 billion words in size, it is likely that all the words used in the vocabulary MCQ test were sampled a great many times and that this overcame any difference in the reliability of the semantic representations due to frequency differences. For the 2 billion word ukWaC corpus that we used (by far the smallest of the corpora used to train the methods compared here), Table 4 gives the frequency statistics for the vocabulary MCQ test synonyms. There is a very large variance in frequency values within each subset. The mean frequency for the LF subset is 3993 with the lowest stem frequency being sampled 98 times in the corpus. The correlation between the $\log_{10}$ ukWaC corpus synonym frequencies and their SUBTLEX-UK Zipf frequencies is 0.93.

Table 4: ukWaC frequencies for the stem words in the vocabulary MCQ test

MCQs	mean	SD	range
All 200	119,809	198,796	98 – 1,057,891
100 LF	3993	5099	98 – 36,571
100 HF	235,625	228,725	12,291 – 1,057,891

We suppose that any differences in performance for the semantic vectors on the LF and HF subsets is due to an inadvertent bias in the distribution of semantic distances between the MCQ words that is revealed when the statistical reliability related to word frequency differences is made irrelevant due to very high degrees of corpus sampling. It is likely that there is a higher degree of semantic ambiguity for the high frequency stems and this may have affected the MCQ results. We will explore these issues in further detail in modelling work in a future paper.

The corpora used to train the vector generation methods ranged from 2 to 840 billion words. Although performance of the different methods differed by only a few percent, it is notable that the B&L vectors achieved one of the higher scores using a corpus of 2 billion words and that the GloVe vectors achieved higher scores as the corpus size used increased from 6 to 42 and then 840 billion words. The B&L method was tuned for previous work on a different

vocabulary MCQ test whilst W2V and GloVe have been reported as having notable success on the rather different domain of analogy problems. It is likely that further parameter tuning would increase the scores of all three methods on this specific task if not in general for other tasks.

Landauer & Dumais (1997) reported that their LSA semantic vectors performed at a level of 64.4% on a TOEFL vocabulary MCQ test. This matched the performance of a large sample of applicants to colleges in the USA from non-English speaking countries who averaged a score of 64.5%. This is close to the performance of our non-monolingual group on the 100 low frequency MCQs (61.4%). These LSA vectors were trained on a much smaller corpus than the other semantic vectors describe here (6.4 million words) and so, arguably, are a better psychological model of attaining a degree of semantic competence from a realistic scale of linguistic input. However, they did not approach the high scores required to claim to be a model of idealised semantic competence.

As noted above, there were several MCQs where the same errors were made by some of the highest performing human participants and some of the vector methods. In some of these cases, it appears that the questions were made more difficult than expected by the random choice of distractor items leading to one of the distractors potentially being more closely semantically related to the stem than the intended synonym. An example is the intended stem, synonym, distractor1, distractor2, distractor3 MCQ: *benefit*, *welfare*, *flask*, *advantage*, *lipstick*. Here, two of the three highest performing participants and four of the five semantic vector methods made an error. Mean human accuracy was at below chance level. Because the vector methods have captured synonymy well, they show the potential for automatically measuring vocabulary MCQ difficulty in terms of semantic similarity over and above the effect of word frequency.

Discussion

In cognitive science, we are often interested in building idealised or technologically useful models of intelligent behaviour as well as psychologically valid ones. Ideally, these are complementary aims. The development of methods to generate distributional semantic vectors over the past 20 years is an interesting example of the possible tensions between these two types of model. Landauer & Dumais (1997) proposed LSA as a model of human semantic performance. LSA was partly validated by its success in matching human non-native performance in the TOEFL MCQ test. However, LSA was not capable of approaching perfect performance on the task. Current techniques have achieved very high levels of success on that task and similar ones such as the test proposed here. However, the amount of data used to train these models is very far in excess of the amount of text that a human could read in a lifetime. The use of distributional semantic vectors in the modelling of human performance (e.g., Pereira et al., 2015; Mandera et al., 2017) and human brain activity (e.g., Mitchell et al., 2008; Bullinaria & Levy, 2013) is becoming more widespread due to better and more easily available semantic vectors. However, it remains unclear how the balance between idealised competence and realistic human performance can be modelled by such techniques and which corpora and parameter settings should be used. Some of these issues can be addressed in the straightforward arena of vocabulary MCQ tests.

In this paper, we introduce a vocabulary test, based on WordNet synsets, that is both challenging enough for human participants not to be performed at ceiling and large and uniform enough to be useful as an evaluation measure for corpus-derived semantic vectors.

We argue that we are within reach of developing distributional semantic vectors that can demonstrate competence in the important, if narrow, domain of synonymy judgement. However, there is much to be done in using such representations as components of models that can successfully account for actual human performance on these same tasks.

Although the semantic vectors we tested were close to an idealised level of competence, they do not reflect the clear effect of synonym frequency in the human data. However, the ability of the semantic vectors to provide reliable measures of semantic similarity does show promise for modelling aspects of vocabulary MCQ question difficulty that are left after the influence of word frequency is accounted for.

A single set of semantic vectors cannot both account for idealised semantic competence as defined by a resource such as WordNet and provide a model of average imperfect human performance. For tasks such as vocabulary MCQ tests, it may be necessary to use semantic vectors as models of competence and account for varying performance using simple psychologically valid variables such as word frequency or familiarity. Certainly, current methods for the generation of semantic vectors only obtain very high performance scores after training on enormous corpora, orders of magnitude larger than any human would experience. This may make them poor or partial models of human semantic *learning* but useful technological tools and cognitive modelling components. It remains to be seen whether semantic vectors with somewhat lesser levels of competence, perhaps trained on much smaller corpora, are better tools for modelling ordinary levels of human performance.

Vocabulary MCQ tests have been useful measures of human word knowledge. In the past they have proved their worth as evaluation measures for semantic vector generation. They are psychological tasks in themselves and we have suggested here that semantic vector methods may allow us to model aspects of question difficulty that are related to relative semantic distances and this may also prove useful for the design of such instruments.

Vocabulary MCQ tests are an important component in the evaluation of representations of lexical semantics. We have

argued that it is important that such representations can account for idealised performance and so reach perfect performance in these tests. Current techniques have not yet reached this level of competence. It would also be highly desirable if these techniques contributed to our ability to model the imperfect performance of human participants on semantic tasks. We argue that vocabulary MCQ tests serve as useful psychological tasks to model. By making our new test freely available along with human data, we hope to stimulate further research.

Acknowledgments

We acknowledge the funding given by our Universities and thank the student participants in our vocabulary MCQ experiment and our colleagues who facilitated this data collection. We also thank the authors of the ukWaC corpus, word2vec methodology and GloVe methodology for making their data and code available to us.

Availability of the MCQ word list

We have made the new vocabulary MCQ test words, and various earlier test sets, available to be downloaded from: http://www.cs.bham.ac.uk/~jxb/corpus.html

References

Adelman, J. S., Johnson, R. L., McCormick, S. F., McKague, M., Kinoshita, S., Bowers, J. S., et al. (2014). A behavioral database for masked form priming. *Behavior Research Methods*, *46*(4), 1052-1067.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The waCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226. http://doi.org/10.1007/s10579-009-9081-4.

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.

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods*, *44*(3), 890-907.

Bullinaria, J. A., & Levy, J. P. (2013). Limiting factors for mapping corpus-based semantic representations to brain activity. *PLoS ONE*, *8*(3), e57191. http://doi.org/10.1371/journal.pone.0057191

Caron, J. (2001). Experiments with LSA scoring: Optimal rank and basis. In: M. W. Berry (Ed.), *Computational Information Retrieval*, 157-169. Philadelphia, PA: SIAM.

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

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

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.

Levy, J. P., & Thompson, N. (2008). Using distributional methods to explore the systematicity between form and meaning in British Sign Language. *From Associations to Rules - Connectionist Models of Behavior and Cognition - Proceedings of the Tenth Neural Computation and Psychology Workshop*, 100–111.

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. http://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. *arXiv preprint arXiv:1301.3781*.

Mitchell, T. M., Shinkareva, S. V, Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191–5. http://doi.org/10.1126/science.1152876

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B*. 369: 20130299. http://dx.doi.org/10.1098/rstb.2013.0299

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2015). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*, 175-190. http://dx.doi.org/10.1080/02643294.2016.1176907

Shipley,W. C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. The *Journal of Psychology, 9*, 371–377.

Schütze, H. (1993). Word space. In: S. J. Hanson, J. D. Cowan & C. L. Giles (Eds.) *Advances in Neural Information Processing Systems 5,* 895-902. San Mateo, CA: Morgan Kauffmann.

Turney, P.D. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, Freiburg, Germany, pp. 491-502.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188. http://doi.org/10.1613/jair.2934

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176-1190.