CrossMark

Association and response accuracy in the wild

Sudeep Bhatia[1] · Lukasz Walasek[2]

Abstract

We studied contestant accuracy and error in a popular television quiz show, "Jeopardy!" Using vector-based knowledge representations obtained from distributional models of semantic memory, we computed the strength of association between clues and responses in over 5,000 televised games. Such representations have been shown to play a key role in memory and judgment, and consistent with this work, we find that contestants are more likely to provide correct responses when these responses are strongly associated with their clues, and more likely to provide incorrect responses when correct responses are weakly or negatively associated with their clues. This effect is stronger for easier questions with low monetary values, and for questions in which contestants compete to respond quickly. Our results show how distributional models of semantic memory can be used to predict human behavior in naturalistic high-level judgment tasks with skilled participants and significant monetary and social incentives.

Keywords Distributional semantics · Associative judgment · Response accuracy · Big data · Field data

Introduction

Distributional models of semantic memory provide a powerful computational approach to understanding how people represent knowledge about real-world objects, individuals, and events. These models describe knowledge representations using high dimensional vectors, trained on natural language word-co-occurrence data, and subsequently specify the association between any two words using the distance between their corresponding vectors (Dhillon, Foster, & Ungar, 2011; Griffiths, Steyvers, & Tenenbaum, 2007; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mikolov et al., 2013; Pennington, Socher, & Manning, 2014).

The idea that knowledge representations are derived from the distribution of words in natural language has a long history in psychology, linguistics, and other areas of cognitive science (Firth, 1957; Harris, 1954). However, with advances in computer technology, as well as the availability of large online datasets of natural language corpora, this insight has been translated into the development of tools and techniques for uncovering the actual knowledge representations possessed by individuals. Such representations have been shown to successfully predict behavior in a wide range of cognitive tasks, including similarity judgment, categorization, cued recall, and free association (for a review see Bullinaria & Levy, 2007 or Jones, Willits, & Dennis, 2015). These representations are also highly successful at modelling language use in humans, and for this reason, are also commonly applied to problems involving the automated understanding of language in computational linguistics (Turney & Pantel, 2010).

Although most of the above work is focused on relatively low-level cognition, recently Bhatia (2017) has shown how this approach can be extended to model high-level judgment. Many such judgments are associative (Kahneman, 2003; Sloman, 1996), and distributional models can provide a quantitative measure of the strength of association between questions and feasible responses. For example, Bhatia (2017) finds that measures of association derived from distributional semantic models accurately predict participant responses to probability judgment and factual judgment questions, with participants being most likely to select responses that are highly associated with the content of the question. This relationship holds both when the associative response is correct and when it is incorrect, showing that distributional semantic models accurately describe both adaptive and fallacious judgment.

✉ Sudeep Bhatia
bhatiasu@sas.upenn.edu

[1] Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA

[2] Department of Psychology, University of Warwick, Coventry, UK

All of the results discussed above have been documented in a controlled lab setting. However, the recent computational and societal developments that have made large natural language datasets available for model training have also made similarly large datasets of human behavior available for model testing (for a discussion of such datasets and the need to use these datasets in cognitive research, see Griffiths, 2015 or Jones, 2017). Thus, it is now possible to apply distributional models of semantic memory to predict high-level cognitive phenomena observed in a variety of real-world circumstances.

In this paper, we attempt such a test, using a dataset of questions from the "Jeopardy!" game show. We apply existing distributional models to obtain vector-based knowledge representations for each of the words in the questions in our dataset. Subsequently, we are able to compute a measure of the associative strength between the clue in each question and the correct response to the question. We use this measure to predict whether contestants are able to successfully provide the correct response. If associations are at play in high-level judgment, and if distributional models accurately quantify these associations, we should expect higher contestant accuracy in questions where correct responses are strongly associated with their clues. This would be the case despite the fact that the "Jeopardy!" game show involves highly skilled contestants in real-world environments with complex stimuli and large monetary and social incentives. Thus, our goal is not to build a question-answering system capable of providing correct responses (e.g., Ferrucci, 2012), but rather to study contestant accuracy and error in the wild, using a theoretically grounded model of knowledge and association.

Methods

Overview of data

The "Jeopardy!" game show presents contestants with clue-based questions. Contestants must respond to these questions with the correct response (typically a single word or phrase) to the clue. The questions have varying monetary values, and contestants earn money or lose money based on the accuracy of their responses. There are three contestants in each game show, and these contestants typically compete to respond to the clue as quickly as possible after it has been read. Thus, responses are made under considerable time pressure.

The clues, as well as the correct responses to the clues, have been compiled by "Jeopardy" fans on www.j-archive.com. This website contains transcripts for the game shows from 1984 to the present time, and we scraped this website to obtain 298,820 questions, across 5,082 different games. For each of these questions we had both the clue text as well as the correct response text. We also had various question and game-level data, including the monetary value of the question,

whether the question was in the first round (Jeopardy!), second round (Double Jeopardy!), or the third round (Final Jeopardy!), whether the question was a Daily Double question, and when the game was played. Importantly, we also obtained data on whether contestants were able to respond to the question correctly. Note that there are some differences between the structure of regular questions in the first two rounds, Daily Double questions in the first two rounds, and Final Jeopardy! questions, and so we analyzed each of these three sets separately. The Online Supplemental Materials describe the "Jeopardy!" game structure and our dataset in detail.

Overview of analysis

We used a prominent prebuilt set of vector representations in order to examine the relationship between contestant accuracy and the association between the words in the questions' clues and the words in the corresponding responses. The representations we used were generated by the Global Vectors for Word Representation (GloVe) model (Pennington et al., 2014), which performs a dimensionality reduction on word co-occurrence matrices, emphasizing the use of the ratios of word-word co-occurrence probabilities. We obtained publicly available GloVe vectors from Pennington et al.'s online repository (http://nlp.stanford.edu/projects/glove/). These vectors were trained on a six billion-word corpus combining English language Wikipedia with the English Gigaword corpus, and have a vocabulary of 400,000 words. Bhatia (2017) found that these vectors described participant responses in high-level judgment tasks with considerable accuracy, and so we restrict the main text of this paper to only the analysis of the GloVe vectors. In the Online Supplemental Materials we replicate the results of our analysis using the Word2Vec and Eigenwords vector representations (Dhillon et al., 2011; Mikolov et al., 2013), also considered in Bhatia (2017).

We computed the association between each clue and response in our dataset, as assessed by the vector representations. These representations specify a word i as a 300-dimensional vector w_i. For a given question, we first generated an aggregate representation of the question clue by taking the average of its words' vectors, weighted by the frequency of the words in the clue (excluding highly common "stop words" and words that were not present in GloVe's vocabulary). The vector for a clue, c, can be written as $c = \frac{\sum_i n_i \cdot w_i}{\sum_i n_i}$, where n_i is the number of times word i occurs in the clue. We can use the same method to build a vector representation of the correct response r, and in turn specify the association between the clue and the response based on the distance between c and r. As in prior work, we used cosine similarity to specify distance, so that the association between c and r is $A(c,r) = c \cdot r/(\|c\| \cdot \|r\|)$. $A(c,r)$ ranges between -1 and +1, with higher values corresponding to clues and responses that are more closely

associated. The Online Supplemental Materials in Bhatia (2017) provide additional details about the computational techniques used in our analysis.

Results

Summary statistics

Table 1 presents the summary statistics for the first round (Jeopardy!) and second round (Double Jeopardy!) questions in our dataset, separated by the question value. For each type of question, it presents the total number of such questions in the dataset, and the mean and standard deviation of contestant accuracy on these questions. This table also presents the total number of such questions in the dataset for which we were able to compute the association between the question clue and the correct response with the GloVe representations, as well as the mean and standard deviation of these association scores. Here contestant accuracy is a binary variable which, for each question, calculates whether or not at least one of the contestants managed to provide the correct response. Association, in contrast, is a continuous variable ranging from -1 to +1, and is calculated by measuring the cosine similarity between the question clue and its corresponding correct response. The reason why we are unable to calculate associations for some questions is because either their clues or their

responses are composed entirely of words absent from the GloVe vocabulary.

Table 1 illustrates a number of regularities in our data. Firstly, contestant accuracy is fairly high, averaging between 66% and 97% based on the type of question in consideration. Likewise, the association measure is also relatively high. Unsurprisingly the correct response for a question is associated with the content of the question clue. More importantly, however, we see both contestant accuracy and association vary systematically with the value of the question in consideration. Question value depends on question difficulty, and we find that contestants tend to answer low-valued easy questions more accurately than high-valued difficult questions. The low-valued questions are also the ones for which the association of the clue and correct response is particularly high. This suggests that there may be a systematic relationship between association and the ability of contestants to give correct responses.

Contestant accuracy in regular questions

The goal of this section is to rigorously test this relationship. More specifically, we examine the correlation between the association of a question clue and its correct response, and contestant accuracy for the question (whether or not one of the contestants managed to provide the correct response). Overall, we find a very strong positive relationship between

Table 1 Summary statistics for different types of Jeopardy! questions. Here "Total # Quest.", "Con. Acc. Mean", and "Conc. Acc. Std." describe the total number of each type of question, as well as the mean and standard deviation of contestant accuracy on the questions. "Assoc. #

Quest.", "Assoc. Mean" and "Assoc. Std." describe the total number of each type of question for which we were able to compute associations, as well as the mean and standard deviation for the associations for the questions

Question value	Total # Quest.	Con. Acc. Mean	Con. Acc. Std.	Assoc. # Quest.	Assoc. Mean	Assoc. Std.
First round (Jeopardy!)						
$100	11,172	0.97	0.17	10,730	0.34	0.16
$200	29,953	0.95	0.21	28,892	0.33	0.16
$300	10,647	0.90	0.30	10,255	0.32	0.17
$400	28,936	0.90	0.30	27,946	0.32	0.17
$500	10,097	0.78	0.41	9,748	0.30	0.17
$600	18,066	0.89	0.31	17,461	0.31	0.17
$800	17,601	0.85	0.35	16,999	0.30	0.17
$1,000	17,578	0.76	0.43	17,006	0.29	0.17
Second round (Double Jeopardy!)						
$200	10,987	0.96	0.20	10,704	0.35	0.16
$400	29,433	0.93	0.25	28,605	0.33	0.16
$600	9,871	0.86	0.35	9,638	0.32	0.17
$800	27,415	0.86	0.35	26,704	0.31	0.16
$1,000	9,614	0.69	0.46	9,398	0.30	0.17
$1,200	16,930	0.84	0.37	16,424	0.30	0.16
$1,600	16,080	0.77	0.42	15,626	0.29	0.17
$2,000	16,803	0.66	0.47	16,275	0.28	0.17

association and contestant accuracy. This is illustrated in Fig. 1, which plots the average contestant accuracy as a function of association, as assessed by the GloVe vectors. Here we have divided all our questions into ten equal portions based on the strength of the association measure for the questions, and pooled contestant accuracy for each of these portions. For the reasons discussed above we exclude Daily Double and Final Jeopardy! questions, as well as questions for which we were unable to compute association (those whose component words are not in the GloVe vocabulary). This leaves us with $N = 272{,}412$ regular questions for the analysis in this section. The histogram nested within Fig. 1 shows the distribution of associations for all questions. As can be easily seen, these association scores are distributed normally.

Figure 1 shows that contestant accuracy gets, on average, progressively higher as the association of the question clue and the correct response increases. Overall, contestant accuracy is at its lowest (around 82%) for the questions whose correct responses are unassociated with the question clues (the first decile), and at its highest (around 87%) for the questions whose correct responses are highly associated with the clues (the ninth decile). It seems that contestant accuracy does drop for the last decile of questions. This could be due to a ceiling on the effect of associative strength on accuracy (as further increases to association after reaching a certain level no longer facilitate increased recall, and there is eventually a regression to the mean). Alternatively, this may capture the effect of
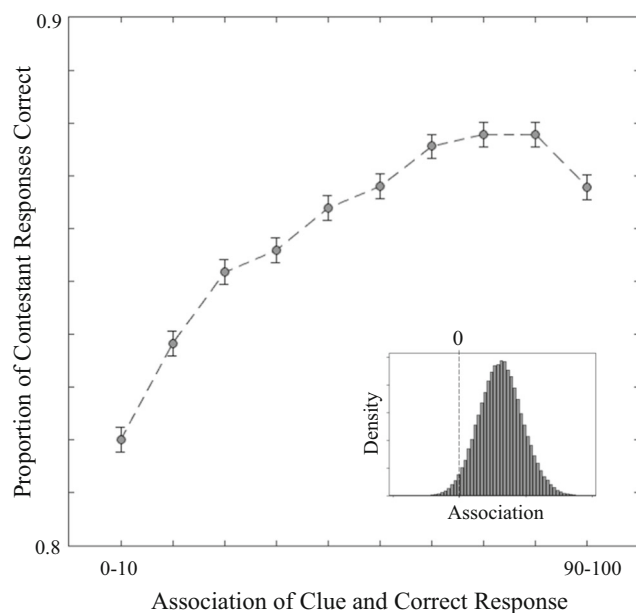
Fig. 1 Average contestant accuracy for questions with different strengths of association between clues and correct responses. The x-axis indicates the association decile (ranging from weakest association to strongest association) for each group of questions, whereas the y-axis indicates the proportion of the questions that are answered correctly by some contestant. The nested histogram shows the distribution of associative strength across all our questions. Error bars indicate standard error

questions with multiple highly compelling intuitive answers (out of which only one is correct). In the Online Supplemental Materials we provide exploratory analysis suggesting that the latter explanation may be correct.

We first examined this relationship statistically using a simple logistic regression. In this regression, our dependent variable was the contestant accuracy for a given question (1 if it was answered correctly by at least one of the three contestants; 0 otherwise), and our primary independent variable was the association between the question clue and correct response, as measured by cosine similarity on our GloVe vectors. This regression revealed a strong positive effect of association on contestant accuracy ($\beta = 0.78$, $z = 23.46$, $p < 0.001$, 95% CI = [0.72–0.85], $OR = 2.18$). We also ran a more rigorous variant of this analysis. This second regression had controls for the monetary value of the question (a dollar amount ranging from $100 to $2,000), in order to ensure the relationship observed in the regression and in Fig. 1 is not confounded by question difficulty. This second regression also included controls for whether the question was part of the first round or second round (1 if in Double Jeopardy!; 0 otherwise), and the year in which the game was played (between 1984 and 2016). This regression also permitted random intercepts for the game in consideration, in order to accommodate game-level effects on contestant accuracy. Finally, as we suspected that the effect of association on contestant accuracy varies across easy and difficult problems, we also included an interaction effect term between association and question value.

Our second regression again found a strong positive relationship between association and contestant accuracy ($\beta = 0.70$, $z = 10.66$, $p < 0.001$, 95% CI = [0.57–0.82], $OR = 2.01$). In addition to this, we also found a strong negative effect of question value, showing that contestant accuracy drops for harder questions ($\beta = -0.14 \times 10^{-2}$, $z = -61.47$, $p < 0.001$, 95% CI = [$-0.15 \times 10^{-2} - -0.14 \times 10^{-2}$], $OR = 0.9986$). Our analysis also revealed positive effects for both Double Jeopardy! ($\beta = 0.38$, $z = 26.20$, $p < 0.001$, 95% CI = [0.36–0.42], $OR = 1.46$) and for year ($\beta = 0.31 \times 10^{-1}$, $z = 26.69$, $p < 0.001$, 95% CI = [$0.29 \times 10^{-1} - 0.34 \times 10^{-1}$], $OR = 1.03$), indicating that contestants are more accurate in the second round of the game show (once question value has been controlled for) and for more recent game shows. Finally, we noted a negative interaction effect between question value and association ($\beta = -0.26 \times 10^{-3}$, $z = -4.29$, $p < 0.001$, 95% CI = [$-0.37 \times 10^{-3} - -0.14 \times 10^{-3}$], $OR = 1.0003$), indicating that the positive effect of association on accuracy drops as the questions get harder.

The effect of association on contestant accuracy for different types of questions is shown in Fig. 2. As in Fig. 1, questions are pooled based on association (this time using quartiles rather than deciles), and the average contestant accuracy for each set of questions is calculated and plotted separately based on the monetary value of the question and whether or not the question was in the first or second round of the game show.
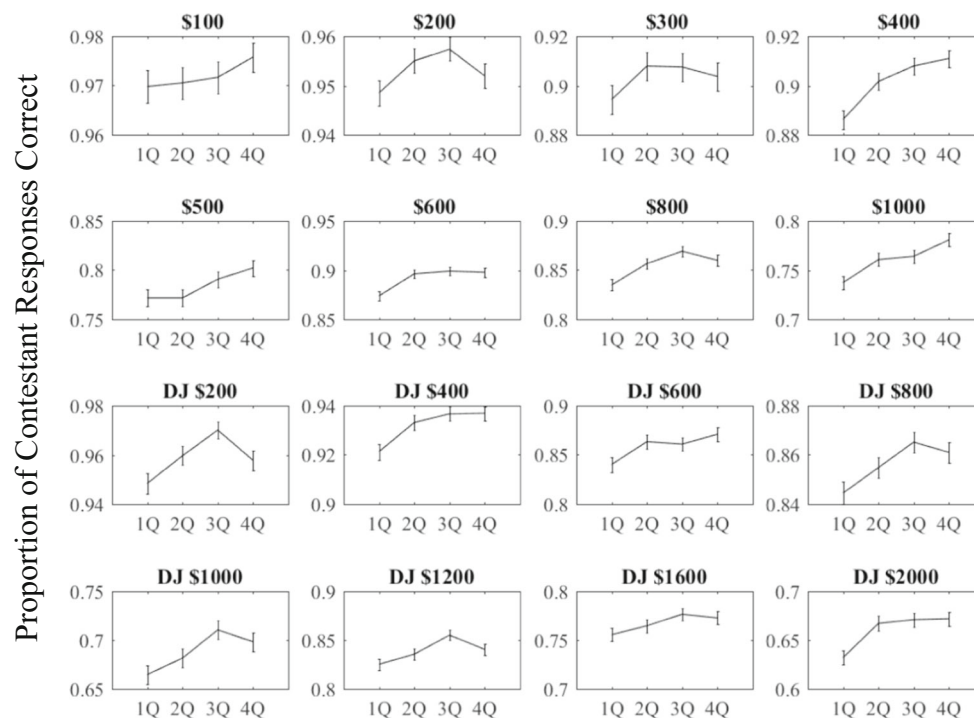
Fig. 2 Average contestant accuracy for questions with different strength of association between clues and correct responses, for different question types (here "DJ" corresponds to the Double Jeopardy! round). The x-axis indicates the association quartile (ranging from weakest association to strongest association) for each group of questions, whereas the y-axis indicates proportion of the questions that are answered correctly by some contestant. Error bars indicate standard error

We repeat the analysis in this section with Word2Vec and Eigenwords representations in our Online Supplemental Materials. In the Supplementary Materials, we also repeat our analysis after excluding questions in which the correct answer is actually present in the clue text (to ensure that such questions are not driving our results).

Contestant accuracy in daily double questions

We also tested the above effects for the Daily Double questions. Note again that these questions have a different format to the regular questions, in that contestants do not have to compete to provide the response first, and are additionally able to specify the amount of money they wish to wager on the question. For the Daily Double questions ($N = 14,584$) we again ran a logistic regression with contestant accuracy as the main dependent variable and the association between the clue and the correct response as the main independent variable. We found a significant positive relationship between these two variables, both with a simple logistic regression ($\beta = 0.30$, $z = 2.88$, $p < 0.01$, 95% CI = [0.10–0.51], $OR = 1.35$) and with a more extensive regression with the multiple controls and random intercepts used in the prior section ($\beta = 0.35$, $z = 2.03$, $p < 0.05$, 95% CI = [0.01–0.70], $OR = 1.42$). Unlike in our previous analysis, however, question value had a

positive relationship with contestant accuracy ($\beta = 0.12\times10^{-3}$, $z = 4.58$, $p < 0.001$, 95% CI = [0.07×10^{-3}–0.17×10^{-3}], $OR = 1.0001$). This likely reflects the contestants' confidence, which correlates positively with both wagered amounts and accuracy for Daily Double questions. For this reason, we also fail to find an interaction effect between question value and association ($p > 0.10$).

It is useful to note that the magnitude of the effect of associative strength on contestant accuracy is much smaller for the Daily Double questions compared to the regular questions in the prior section. This may reflect the fact that contestants do not have to compete to provide responses, and thus need not rely as strongly on associative cues (which are likely to be disproportionately used when under time pressure). We tested this formally by combining our Daily Double questions with the regular questions from the previous section, and performing a logistic regression to predict contestant accuracy. This regression included main effects for association and Daily Double, as well as an interaction between these two variables. Like our previous regressions it also had controls for the year and the value of the question, and random effects for the game. As expected, this regression showed a positive effect of association on accuracy ($\beta = 0.61$, $z = 17.77$, $p < 0.001$, 95% CI = [0.54–0.67], $OR = 1.84$). More interestingly, however, we obtained a negative interaction effect between

association and Daily Double, indicating that contestants are less likely to use association for such questions ($\beta = -0.41$, $z = -3.39$, $p < 0.001$, 95% CI = [-0.65– -0.17], $OR = 0.66$).

The Online Supplemental Materials report a similar analysis for Final Jeopardy! questions. This round is different from the others in that all three contestants must provide an answer to the question. Here we found no significant correlation between cue and response association and contestant accuracy. This could reflect the fact that Final Jeopardy! questions are some of the hardest questions in the game show and associations are less useful for these types of questions (as evidenced by the negative interaction effect between question value and association, shown previously). It could also be due to the fact that contestants do not have to compete to provide responses, and thus need not rely as strongly on associative cues. Indeed, this is the case for the Daily Double questions analyzed above. Both these issues are likely compounded by the relatively small sample sizes in our dataset for Final Jeopardy! questions (there is only one such question per game).

Discussion

We used distributional models of semantic memory to specify the strength of association between clues in the Jeopardy! game show and their corresponding correct responses. We found that contestants are more likely to provide the correct response if this response is strongly associated with the clue. This relationship weakens when questions increase in their difficulty (as with high monetary value Jeopardy! questions) and when contestants are not under time pressure to respond first (as with Daily Double questions).

Our results provide strong support for the predictive power of distributional models of semantic memory (Dhillon et al. 2011; Griffiths et al., 2007; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mikolov et al., 2013; Pennington et al., 2014), showing that such models can be successful even in the context of high-level associative judgment (Kahneman, 2003; Sloman, 1996; also see Bhatia, 2017). In addition, they showcase a novel method for analyzing high-level cognition in the real-world. Such analyses ensure the robustness and generalizability of existing theories in settings with much more data, complexity, and realism than those achievable in the laboratory. They are also valuable for understanding the ways in which cognitive mechanisms (such as those involving associative judgment) manifest in everyday life, thereby facilitating the development of richer theories of human cognition and behavior (Griffiths, 2015; Jones, 2017).

Some readers may note a similarity between the dataset used in this paper and that used to train IBM Watson's groundbreaking Jeopardy-playing computer (see Ferrucci, 2012). Note, however, that, unlike IBM, our goals are not to answer Jeopardy questions accurately, but rather to study the psychological determinants of human Jeopardy responses (responses that are both correct and those that are incorrect). Of course, future work could adopt some of the computational advancements developed for question-answering systems such as Watson.

Such work could also attempt to integrate the proposed approach with more sophisticated psychological theories of question-answering (e.g., Anderson et al., 2004; Reder, 1987), which are able to process complex relations between the clues and the responses, while also specifying metacognitive processes for controlling memory search and response generation. We look forward to research that exploits these new and exciting data sources and techniques, to further integrate the analysis of large-scale human data into the study of cognition and behavior.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4), 1036.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, 124(1), 1-20.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39, 510–526.

Dhillon, P., Foster, D. P., & Ungar, L. H. (2011). Multi-view learning of word embeddings via CCA. In *Advances in neural information processing systems* (pp. 199-207).

Ferrucci, D. A. (2012). Introduction to "this is watson". *IBM Journal of Research and Development,* 56(3.4), 1. Chicago

Firth, J. R. (1957). Papers in linguistics. London, England: Oxford University Press.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, 135, 21-23.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114(2), 211.

Harris, Z. S. (1954). Distributional structure. *Word*, 2, 146–62.

Jones, M. N. (2017). *Big data in cognitive science*. Psychology Press: Hove.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114(1), 1.

Jones, M. N., Willits, J. A., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer & J. T. Townsend (Eds.) *Oxford handbook of mathematical and computational psychology* (pp. 232–254).

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58,* 697–720.

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).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical methods on natural language processing* (pp. 1532-1543).

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology, 19*(1), 90-138.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1).

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1), 141-188.