Vector Space Semantic Models Predict Subjective Probability Judgments for Real-World Events

Sudeep Bhatia (bhatiasu@sas.upenn.edu)
Department of Psychology, University of Pennsylvania
Philadelphia, PA.

Abstract

We examine how people judge the probabilities of real-world events, such as natural disasters in different countries. We find that the associations between the words and phrases that constitute these events, as assessed by vector space semantic models, strongly correlate with the probabilities assigned to these events by participants. Thus, for example, the semantic proximity of "earthquake" and "Japan" accurately predicts judgments regarding the probability of an earthquake in Japan. Our results suggest that the mechanisms and representations at play in language are also active in high-level domains, such as judgment and decision making, and that existing insights regarding these representations can be used to make precise, quantitative, a priori predictions regarding the probability estimates of individuals.

Keywords: Judgement and decision making; Subjective probability; Semantic representation; Semantic space models

Introduction

Subjective probability judgment plays an important role in everyday cognition and behavior. Our interactions with the world around us are guided by the probability estimates we place on its largely uncertain events. These estimates, in turn, stem from our knowledge of the world, and the cognitive mechanisms that we possess for learning, representing, and applying this knowledge.

The study of subjective probability judgment has yielded a number of valuable insights regarding how individuals assign probabilities to uncertain events (Kahneman & Tversky, 1973; Lichtenstein, Fischhoff & Phillips, 1977; Tversky & Koehler, 1994; Wallsten & Budescu, 1983). One of the most important of these insights involves the use of simple heuristics, such as those relying on associations between the various objects or concepts involved in the judgment task (Kahneman & Frederick, 2002; Sloman, 1996). The use of association-based heuristics is relatively effortless, though it can lead to biases in specific settings. This is one reason why, for example, individuals commit the conjunction fallacy in the Linda problem (Tversky & Kahneman, 1983), which asks them to judge whether Linda, a female activist concerned with issues of social justice, is more likely to be a bank teller or a feminist bank teller. Here the description of Linda is strongly associated with feminism, making participants believe that the probability of Linda being a feminist bank teller is higher than her being a bank teller, despite the fact that all feminist bank tellers are in fact also bank tellers.

The events considered in most research on associative judgment are abstracted or artificial. These types of tasks are valuable, as they allow for rigorous tests of scientific hypotheses. However, in order for association-based heuristics to be considered good models of subjective probability judgment, they should be able to predict the specific probabilities individuals assign to the occurrence of real-world events, that is, the types of events that individuals encounter and evaluate on a day-to-day basis. Thus we should not only be able to state that individuals place a higher probability on Linda being a feminist bank teller compared to a bank teller, but also predict the explicit probabilities individuals attach to, for example, various outcomes in current affairs or popular culture. These types of events are often of the form "X happens to Y" (e.g. an earthquake occurs in Japan), and associative heuristics predict that the association between X and Y (e.g. "earthquake" and "Japan") is used by individuals to judge the probabilities of these types of events.

Predicting real-world judgments is not trivial: Although associative heuristics are easy to apply, the information that these heuristics utilize is fairly complex. Thus, even though decision makers may use the strength of association between "earthquake" and "Japan" to predict the probability of there being an earthquake in Japan, it is not immediately clear what determines these associations, and in turn what the decision maker's actual probability judgment about an earthquake in Japan will be.

Associative processing is also of interest in the study of language, and there have been many recent advances in understanding the determinants of association, or more generally, semantic relatedness, as it applies to people's comprehension and use of words (Bullinaria & Levy, 2007; Griffiths, Steyvers, & Tenenbaum, 2007; Jones & Mewhort, 2007; Landaur & Dumais, 1997; Lund & Burgess, 1996). The key insight underlying these advances is that the representation of words depends on the statistical structure of the environment in which these words occur (see also Firth, 1957 and Harris, 1954). Studying the distribution of words in the types of settings people encounter on a day-to-day basis can uncover the representations that people have of everyday words, and in turn the semantic relationships and associations between these words, and the objects and concepts they represent.

Models that build semantic representations using the distribution of words often characterize each word in their vocabulary as a vector in a highly multidimensional space. The proximity between the vectors of two words corresponds to the relatedness or association of the words, so that synonyms and other closely related words are frequently located near each other. Vector space semantic

models are typically trained on very large natural language text corpora, and subsequently have large vocabularies, which can be used to make predictions regarding judgments of word similarity, the strength of word priming, and related psycholinguistic phenomena, for nearly all the words commonly used in a given language. The predictions of these models have been shown to be highly accurate, suggesting that the representations recovered by these models provide a good characterization of the representations underlying semantic processing in language use. For this reason, these models are also popular in machine learning and artificial intelligence, particularly in applications related to computer processing of natural language (see Turney & Pantel, 2010).

Hare, Jones, Thomson, Kelly, and McRae (2009) have shown that the word associations captured by vector space approaches are able to account for priming effects regarding event representation. Relatedly, Paperno, Marelli, Tentori, and Baroni (2014) have found that word association correlates very strongly with explicit probability judgments of word co-occurrence. These results suggest that the representations and associations that guide word use and comprehension may also be the ones involved in making probability judgments for real-world events, and that vector space semantic models could in turn be used to predict these probability judgments. Thus, for example, we could obtain an estimate of the actual probability individuals assign to there being an earthquake in Japan by examining the (linguistic) association between "earthquake" and "Japan" generated by vector space semantic models.

In this paper we test this idea by studying subjective probability judgments about different countries and different famous people. Our tests utilize vector representations released by Google Research, which are based on the recurrent neural network methods proposed by Mikolov and coauthors (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Across eight studies we ask participants to assign probabilities to various natural events happening in these countries (e.g. earthquake in Japan) and to these people (e.g. Jon Stewart becoming president), and we predict judged probabilities using the distance between the vectors for the various words and phrases that make up the events.

Methods

Participants

Our tests involve eight distinct studies with 200 participants each in Studies 1-4 and 100 participants each in Studies 5-8, leading to a total of 1,200 participants (overall mean age = 34.81, 51% male). These participants were recruited from Amazon Mechanical Turk, and performed the studies online, for which they were each compensated $0.50. Our studies included an attention check question and we excluded the 31 participants who failed this attention check across the studies. The number of participants in the above studies was determined prior to running the studies, and the specific numbers were chosen as they were round numbers that allowed us to obtain a sufficient number of estimates for each event.

Materials and Procedure

The first four of our studies asked participants to judge the probability that various man-made and natural disasters would happen in the different countries of the world. For each of the countries offered to the participants, they were asked to assess the probability that the country would experience a terrorist attack in the next week (Study 1), be in a state of war at the start of 2016 (Study 2), experience an earthquake over the next year (Study 3), or experience an epidemic over the next year (Study 4). Each participant in each study was given a list of 30 countries chosen at random from the 193 countries that were members of the United Nations when the studies were run.

The remaining four studies asked participants to make judgments about various famous people in the United States. For each person offered to the participant, he or she was asked to assess the probability that the person would be the U.S. president in 2020 (Study 5), win a Nobel Prize in 2020 (Study 6), win a Grammy Award in 2020 (Study 7), or win an Academy Award in 2020 (Study 8). The list of famous people used in this study was obtained from a separate pool of MTurk participants (mean age = 36.81, 56% male) who were each asked to write the names of ten highly recognizable people in the United States that were still alive. The 50 most frequent names generated by these participants were used in Studies 5-8. Again, each participant in Studies 5-8 was asked to make judgments about 30 people chosen randomly from our list of 50 people.

Probability judgments for each of the countries in Studies 1-4 and each of the people in Studies 5-8 were made on a slider scale between 0-100%. The 30 events offered to each participant were presented one after the other, on separate screens, in a random order. With the 30 judgments for each of the 200 participants, we obtained approximately 30 probability estimates for each of the events considered in Studies 1-4. Likewise, with the 30 judgments for each of the 100 participants, we obtained approximately 60 probability estimates for each of the events considered in Studies 5-8.

Overview of Analysis

In this paper we utilize a set of pretrained vector representations recently released by Google Research (code.google.com/p/word2vec). These vectors have been trained on a 100 billion word subset of the Google News corpus, by continuous bag-of-words (CBOW) and skip-gram techniques of Mikolov et al. (2013a, 2013b). This approach relies on a recurrent neural network that, for the CBOW technique, attempts to predict words using other words in their immediate context, and for the skip-gram technique, attempts to the do the inverse of this. These representations have a vocabulary of 3 million words and phrases, including countries and names with two or more component words, such as "United States" and "Jon Stewart". The recent successes of Mikolov et al.'s methods,

the large amount of training data and resulting vocabulary used in the word representations, and the fact that these representations have been obtained from a news corpus, make them particularly valuable for the tests we are conducting.

Each of the 3 million vectors we use is described on 300 dimensions, and the linguistic association between any two words or phrases can be computed using the distance between their corresponding vectors in this 300 dimensional space. In this paper we use the distance between "terrorism", "war", "earthquake", and "epidemic" and the words corresponding to the 193 countries to predict the probabilities that participants assign to the disasters happening in the countries in Studies 1-4. Likewise we use the distance between "president", "Nobel Prize", "Grammy Award", and "Academy Award" and the names of the 50 famous people to predict the probabilities that participants assign to the people winning these awards in Studies 5-8. Thus for example, we can calculate the association of "earthquake" and "Japan" or of "President" and "Jon Stewart" using the distance between each of these pairs of vectors, and in turn use this distance to predict the probability people assign to there being an earthquake in Japan, or to Jon Stewart becoming president. The metric of distance we consider is cosine similarity, so that the distance between any two vectors a and b is given by $dist(a,b) = a \cdot b/(||a|| \cdot ||b||)$. This metric varies between -1 and +1 (with -1 capturing orthogonal vectors and +1 capturing vectors with identical directions). Additional details about Mikolov et al.'s Word2Vec training techniques can be found in Mikolov et al (2013a, 2013b). Note that in analyzing Studies 5 and 6, we remove famous people who have previously won Nobel Prizes or have previously served two terms as the president of the United States (these awards or positions cannot be won again in the future). We also exclude participant judgments regarding St. Vincent and Grenadines, as this country is not represented in the set of word vectors released by Google Research.

Results

Overview of Data

Recall again that there are about 30 participant judgments of event probability for each of the 193 events in Studies 1-4, and about 60 participant judgment of event probability for each of the 50 events in Studies 5-8. In this paper our main dependent variable will be the probability estimate for an event obtained by averaging all the probability estimates made by participants for that event. We find that these average probability estimates vary substantially with the event that participants are required to judge, with, for example, the average probability assigned to there being an earthquake in Japan over the next year being 55.03% (N = 33, SD = 25.24) and the average probability assigned to there being an Earthquake in Norway being only 11.88% (N = 18, SD = 16.78).

Average event probabilities are highly dispersed for judgments regarding Grammy and Academy awards in Studies 7 and 8. In these studies, the average probabilities assigned to different people winning these awards appear to be roughly uniformly distributed between 0% and 70%. In contrast we observe the lowest dispersion in average estimates of epidemic and earthquake probabilities in Studies 3 and 4, in which average event probabilities are clustered between 20% and 40%.

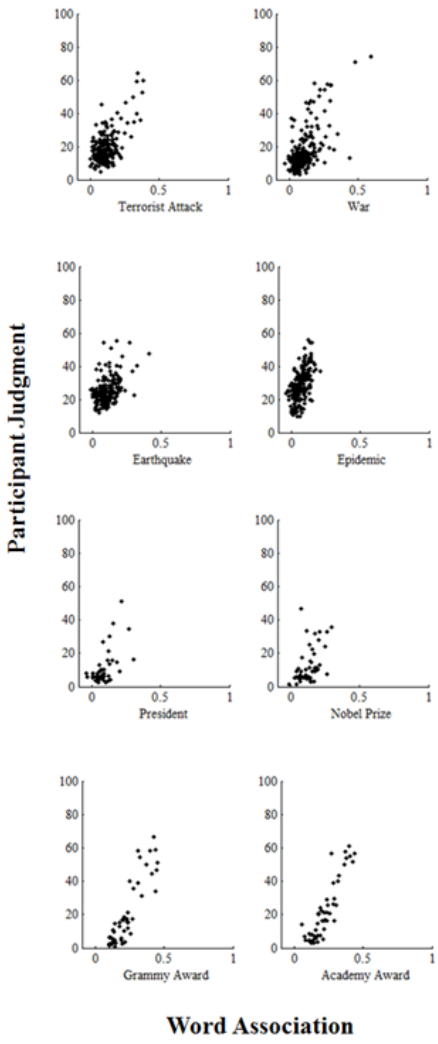
The main independent variable in this paper will be the linguistic association between the word and phrases in an event. Again, this measure is the cosine similarity between the disasters and countries (in Studies 1-4) or the awards and people (in Studies 5-8). Although cosine similarity can vary between -1 and +1, associations between the words in 941 out of the 968 events in our studies were positive. The distributions of these associations were, as with probability estimates, most dispersed for events involving Grammy and Academy awards in Studies 7 and 8, and least dispersed for events involving epidemics and earthquakes in Studies 3 and 4. The distribution of average event probabilities and word associations for the different events in our studies can be observed in Figures 1A-1H.

The Predictive Power of Word Associations

Can associations predict event probabilities? Recall that for each of the 968 events across our studies we have both the average probability assigned to the event by our participants, and the association (or, more formally, cosine similarity) between the words in the event, specified by our set of vector representations. A first step in our analysis is examining the correlation between word associations and the average estimates of participants. A standard test using Pearson's correlation reveals positive, significant correlations between these two variables in each of our studies ($p < 0.001$). Overall our approach does best in Studies 7 and 8, which involve judgment of popular culture, particularly the probability of various people winning Grammy Awards and Academy Awards. Here word associations and average participant probability estimates have correlations of 0.89 and 0.90 respectively. Our approach is also highly successful at predicting judgments of terrorism and war in different countries, in Studies 1 and 2, with correlations of 0.62 and 0.64 respectively. Predictions regarding judgments of presidential victories and Nobel Prizes for different people in Studies 5 and 6 achieve correlations of 0.59 and 0.48 respectively. The model does worst in predicting judgments of natural disasters in different countries, with correlations for earthquakes in Study 3 and epidemics in Study 4, being 0.44 and 0.57. Scatter plots displaying the relationship between word associations and participant judgments can be seen in Figures 1A-1H , and the correlations outlined here are summarized in Table 1.

It is interesting to note the above analysis does not involve any model fitting, and the word associations we use are already built in to the vector representations released by Google Research. Ultimately, the results discussed here (such as correlations of 0.89 and 0.90 between word

associations and participant probability judgments for popular culture events) emerge from what is essentially a zero parameter model.



Figures 1A-H: Scatter plot of the word associations (in terms of cosine similarity) generated by the model and average participant probability estimates (in terms of percentage) for the events in Studies 1-8 respectively.

Some model fitting could, however, help us better understand the properties of the approach we are proposing. For this purpose we first consider a simple linear model, which transforms the cosine similarity measure of word association, which ranges from -1 to 1, into a probability judgment scale, which ranges from 0 to 100. Note that such a linear fit would not change correlations or their significance levels, but would allow for a better understanding of the degree of variance in the data explained by our approach. After fitting linear models for the eight studies, using a basic linear regression, we unsurprisingly find highly significant relationships between word associations and participant judgments ($p < 0.001$ for each of the studies). Overall the R^2 statistics for these fits vary between 0.19 (for earthquake judgments in Study 3) and 0.80 (for Academy Award judgments in Study 8). More details about these fits are provided in Table 1.

The differences in the correlations and fits across the eight studies could be attributed to varying uses of association-based heuristics in different domains. Perhaps participants are just more likely to apply associative processing when making judgments of pop culture, as in Studies 7 and 8, compared to judgments regarding natural disasters in different foreign countries or winners of Nobel Prizes, as in Studies 3, 4, and 6. Alternately, it is possible that we have the highest correlations in Studies 7 and 8 because our participants have more knowledge about popular culture than they do about natural disasters or Nobel Prizes. Due to their increased knowledge they are more likely to make fine grained probability assessments in Studies 7 and 8, allowing for a cleaner dataset on which to predict probability judgments. Indeed, as discussed above, average probability estimates for the events in Studies 7 and 8 have a much higher spread than average estimates in Studies 3 and 4.

The above analysis has only attempted to predict the average probability estimate placed on the events by our participants. This type of aggregation is desirable for many reasons (see e.g. Wallsten, Budescu, Erev & Diederich, 1997 for a review). However it ignores the variance across participants in their judgments of the probabilities. Do word associations provide a good account of probability judgment when we allow for participant-level heterogeneity? We can test this on participant-level data using a linear regression model with participant-level random intercepts. As many individual participant estimates, unlike aggregate estimates, lie at the boundaries of the probability scale (i.e. at 0% and 100%) the regression we use permits a censored dependent variable using the Tobit method (Tobin, 1958). With these controls we find that cosine similarity has a strong positive significant relationship with probability estimates in each study ($p < 0.001$ for each study), showing that associations can predict not just aggregate probability judgments but also probability judgments on an individual level.

Study	Event	Target	ρ_{linear}	ρ_{log}	$\rho_{logistic}$	R^2_{linear}	R^2_{log}	$R^2_{logistic}$
1	Terrorist attack next week	Country	0.62	0.40	0.65	0.38	0.16	0.42
2	Being at war at start of 2016	Country	0.63	0.47	0.63	0.40	0.22	0.40
3	Earthquake next year	Country	0.44	0.37	0.44	0.19	0.14	0.20
4	Epidemic next year	Country	0.57	0.48	0.57	0.32	0.24	0.32
5	US president in 2020	Person	0.59	0.48	0.59	0.35	0.23	0.34
6	Nobel Prize in 2020	Person	0.47	0.41	0.49	0.23	0.17	0.24
7	Grammy Award in 2020	Person	0.89	0.85	0.88	0.79	0.72	0.78
8	Academy Award in 2020	Person	0.90	0.81	0.91	0.81	0.65	0.84

Table 1: Summary of correlations (ρ) and R^2 values from linear, logarithmic, and logistic model fits for the events in Studies 1-8. Note that all model fits involve two free parameters, and that the correlations correspond to those displayed in Figures 1A-H. All of these correlations are highly significant ($p < 0.001$).

Testing Non-linear Relationships

Psychophysical judgments are often characterized by non-linear relationships, as with the Weber-Fechner law, and it is possible that the associations decision makers perceive between the words in the event at hand are transformed non-linearly before being mapped onto probability judgments. We can test this by comparing the fits of the above linear model with a group of similarly parameterized non-linear models. The first model we consider takes a natural-log transformation of the associations for each event. These transformed values are then fit by minimizing mean-squared error. As with the untransformed linear regression, this is a two parameter model. Thus if we write the average probability estimates as y and the word associations for an event as x, our logarithmic model would attempt to find parameters β_0 and β_1 to fit $y = \beta_0 + \beta_1 \ln(x)$. Note that the log transform cannot be used on negative numbers. A very small minority the cosine similarity values for the events are in fact negative. These have been ignored in our analysis (none of the results change if we use more complex log-based functions for transforming these negative numbers).

The second model we consider is a logistic curve. Such sigmoidal (s-shaped) curves are frequently used to obtain choice probability estimates in discrete choice experiments, as their outputs are bounded by 0 and 1, and the logistic curve is perhaps the most commonly used of all of these (E.g. with Luce's choice rule). We fit the two-parameter logistic curve by minimizing mean-squared error, with the cosine similarity values as our independent variable and the average participant probability estimates as our dependent variable. Here if we write the average probability estimates as y and the word associations for an event as x, our logistic model would attempt to find parameters β_0 and β_1 to fit $y = 1/(1+\exp\{-\beta_0 - \beta_1 x\})$.

As both our logarithmic and our logistic models involve two parameters, their predictive accuracy can be directly compared with those of the linear model described above. Ultimately we find that the logistic and the linear models perform about equivalently, providing nearly identical correlations and R^2 values. The logarithmic model, in contrast provides a much more inferior fit than the linear and logistic models, with correlations and R^2 values as low as 0.41 and 0.17 for the Nobel Prize judgments in Study 6. Despite this fact, the predictions of all models have statistically significant relationships with the judgments of participants ($p < 0.001$ for all models in all studies).

A brief examination of the parameter values generated by our logistic fits explains why they perform as well as the linear models. These parameters are typically such that the inputs to the logistic function fall within its middle, linear range, implying that the logistic function behaves roughly like a linear model. Ultimately, it seems that our measure of word association maps linearly onto the probability judgments of our participants. A summary of the fits for the logarithmic and logistic models is provided in Table 1.

Discussion and Conclusion

In this paper we find that vector semantic space approaches can predict the probability judgments that people make about various events in the world. More specifically, the associations between words and phrases, as assessed by a set of vector representations released by Google Research and trained using the recurrent neural network methods proposed by Mikolov et al. (2013a, 2013b), correlates very heavily with the probabilities that people assign to natural events described using those words and phrases. Furthermore, model fitting indicates that this relationship is linear, rather than logarithmic or sigmoidal.

There are some limitations to the approach we have proposed. For example, vector space semantic models cannot by themselves modify their output to control for the length of time the events are supposed to occur. Thus, the approach described above would give the same probability estimate for an earthquake in Japan in the next one year, as it would for a similar earthquake in the next five years. There is some evidence that human probability judgment doesn't sufficiently account for magnitudes, such as durations of the events (see Fredrickson & Kahneman, 1993), but there is no doubt that our predictions could be improved by allowing for a secondary system that adjusts the estimates obtained through semantic relatedness based on event duration, as well as other non-sematic features of the event at hand.

Performance could also be improved by fitting the vector space models to the data. Recall that the above analysis only performs a transform of cosine similarity to predict probability judgment. It alters neither the number of dimensions used in the model, nor the size of the context window to train the models, nor the weights placed on these dimensions to judge semantic distance (both of which are specified a priori). A more sophisticated approach that trains the vector space models on the probability estimates of participants, would no doubt provide better predictions regarding their subjective probability estimates.

There are also boundary conditions. For example, it is unlikely that the approach outlined in this paper would be able to successfully describe probability judgments involving symbolic reasoning or more complex delilberative processing. However, despite these limitations, our results have some important implications. Firstly they provide new techniques for predicting real-world judgments. These predictions are quantitative, in that they attempt to capture the exact numerical probability assigned to an event. These predictions are also domain-general, in the sense that they can be applied to a number of different types of real-world events. Ultimately, the vector space models that we use have very large vocabularies, and are able to provide a precise measure of association between any two words or phrases in their vocabularies, and subsequently precise probability judgments for simple events composed of these words and phrases. Existing judgment models do not have this important property.

In addition to numerous practical applications to areas such as risk perception and communication (Slovic, 2000), these quantitative predictions can be used to more rigorously study the processes already known to characterize probability judgment (Kahneman & Tversky, 1974; Lichtenstein et al. 1977; Tversky & Koehler, 1994; Wallsten & Budescu, 1983), and also potentially uncover novel effects and regularities. They can also be used to predict everyday decisions involving these events, such as, for example, the purchasing of insurance. These types of real-world decisions are of considerable scholarly interest.

Our results also highlight the power of association-based heuristics in making probability judgment (Kahneman & Frederick, 2002; Sloman, 1996). Though the descriptive power of these heuristics is accepted, this paper is the first to show that a semantic instantiation of these heuristics can be used to predict the actual probabilities that decision makers assign to natural everyday events. Likewise, these results illustrate the power of vector semantic space models. These approaches not only predict judgments of word meaning and use in linguistic domains, but also judgments involving complex events in the real-world. It is important to note that many of the results could not be obtained by simpler approaches that use, for example, only the co-occurrence of words to judge associations. "Jay Z" and "Joe Biden" may never directly co-occur with "Academy Award", but participants nonetheless ascribe a higher probability to Jay Z, a musician, winning an Academy Award, relative to the Joe Biden, a politician.

Finally, the link between the representations and associations used to assess word meaning and the representations and associations used to make probability judgments, observed in this paper, suggests the existence of a single system of learning, storing, and retrieving knowledge for both language use and for high-level judgment. Subjective probability judgment does not rely on knowledge that is fundamentally different from the type of knowledge used to understand and generate language. This in turn implies a close connection between two important and influential areas in psychology. Future work should attempt to build more general models of semantic cognition; models which are not only able to explain how people acquire the knowledge of word meanings, but also how people use these word meanings to form their beliefs about the objects and events they encounter in the world.

References

Bullinaria, John A., and Joseph P. Levy. "Extracting semantic representations from word co-occurrence statistics: A computational study." Behavior Research Methods 39.3 (2007): 510-526.

Firth, J. R. (1957). A synopsis of linguistic theory 1930 – 1955. In Studies in linguistic analysis (pp. 1–32). Oxford, England: Blackwell Publishers.

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. Journal of Personality and Social Psychology,65(1), 45-68.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological Review, 114(2), 211-235.

Hare, M., Jones, M., Thomson, C., Kelly, S., & McRae, K. (2009). Activating event knowledge. Cognition, 111(2), 151-167.

Harris, Z. (1954). Distributional structure. Word, 10, 146 – 162.

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

Kahneman, D., & Frederick, S. (2002). Representativeness revisited. In Heuristics and Biases: The Psychology of Intuitive Judgment.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80(4), 237.

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).

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). Calibration of probabilities: The state of the art (pp. 275-324). Springer Netherlands.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, 28(2), 203-208.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

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.

Paperno, D., Marelli, M., Tentori, K., & Baroni, M. (2014). Corpus-based estimates of word association predict biases in judgment of word co-occurrence likelihood. Cognitive Psychology, 74, 66-83.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. Psychological Bulletin, 119(1), 3.

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

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. Psychological Review, 90(4), 293.

Tversky, A., & Koehler, D. J. (1994). Support theory: a nonextensional representation of subjective probability. Psychological Review, 101(4), 547.

Wallsten, T. S., & Budescu, D. V. (1983). State of the art—Encoding subjective probabilities: A psychological and psychometric review. Management Science, 151-173.

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. Journal of Behavioral Decision Making, 10(3), 243-268.