

# Inferring a Probabilistic Model of Semantic Memory from Word Association Norms

Mark Andrews (m.andrews@ucl.ac.uk)  
David Vinson (d.vinson@ucl.ac.uk)  
Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)

Cognition, Perceptual and Brain Sciences  
University College London,  
26 Bedford Way  
London, WC1H 0AP  
United Kingdom

## Abstract

In this paper, we introduce a method of data-analysis for word association norms. The defining characteristic of this method is that is based upon the inference of a probabilistic generative model of word-associations. The inferred model can in principle provide a clear and intuitive representation of the semantic knowledge inherent in word association data, facilitate an understanding of the process by which word associations are generated, extrapolate beyond the observed data to make reasonable inferences about new word associations, and facilitate an understanding of the process underlying false recall in memory experiments. Finally, the nature and form of the probabilistic model inferred using this method is directly comparable to the so-called Topics-model of Griffiths, Steyvers, and Tenenbaum (2007). As such, a potential future application of this work is the analysis and validation of the semantic knowledge inferred from the distributional statistics of text by direct comparison with the semantic knowledge inherent in association norms.

## Introduction

In a typical word association task, participants are provided with a set of approximately 100 *stimulus* words and for each of these, they produce a single *response* word that they judge to be meaningfully related or otherwise strongly associated. Aggregating these responses across a large number of stimuli and participants, word association norms can be collected. One such set is provided by Nelson, McEvoy, and Schreiber (1999). This set provides counts of the responses to a set of 5018 stimulus words, aggregated over more than 6000 participants.

Word association norms comprise a rich set of statistical data from which details of the structure and organization of semantic memory can be inferred. The use of word associations in this respect can be traced back to the pioneering psychometric work of Sir Francis Galton. Galton (1879) introduced word association as a basic instrument of psychometric research, and used it to classify mental concepts according to, for example, their developmental origin and sensory-motor modality. The more contemporary usage of association norms as a measure of the structure of semantic organization is largely attributed to the work of Deese, e.g. Deese (1962, 1965). Deese used factor analysis to discover the principal dimensions along which semantic memory is organized. He demonstrated that from different sets of norms, it is apparent that semantic memory is organized

in terms of coherent semantic classes or clusters, such as animacy, inanimacy, color, sound, etc., with each individual word being differentially weighted along these various dimensions. Following in this tradition, more recently, Steyvers, Shiffrin, and Nelson (2004) derived what they termed a *word association space* model of semantic memory by applying singular value decomposition (SVD) to the sparse word association matrix of the Nelson norms. As with factor analysis, SVD will lead to a minimal set of linear dimensions along which semantic memory can be characterized. From this, as well as from a metric multidimensional scaling (MDS) analysis of a dissimilarity matrix derived from the word associations, Steyvers et al. (2004) were able to successfully predict inter-word semantic similarity ratings and cued recall memory rates.

In this paper, we take an explicitly probabilistic approach to the problem of inferring the structure of semantic memory from association norms. In particular, we assume that word associations are the observed data from a probabilistic generative model whose parameters are unknown, and are thus to be inferred from the data. The system so inferred specifies a model of semantic memory that can reveal major dimensions of semantic structure as evidenced by association norms, and can allow us to simulate characteristics of a semantic memory system such as the generation of new word associations, and the occurrence of false-recall in memory experiments. In addition, the nature and structure of the probabilistic model makes it directly comparable to currently existing probabilistic models of semantic representation, such as that of Griffiths et al. (2007).

## A Generative Model for Word Associations

Association norms can be described as a sparse  $n \times m$  matrix  $A$ , where  $n$  is the total number of stimulus word types,  $m$  is the total number of response word types and  $A_{ij}$  denotes the number of times that response word  $w_j$  is produced given stimulus word  $w_i$ . It is natural to view this data as arising from a probabilistic generative process whereby responses to stimulus  $w_i$  are seen as draws from a conditional probability distribution over the  $m$  discrete word types. The probability that response  $w_j$  will be produced given stimulus  $w_i$  is given by  $P(w_j|w_i)$ , and hence  $A_{ij}$  denotes the number of times, out of a total of  $\sum_{j=1}^m A_{ij}$ , that  $w_j$  is drawn from this distribution.

Interdependences between the conditional distributions can be captured by assuming the mixture model

$$P(w_j|w_i) = \sum_{k=1}^K P(w_j|x=k)P(x=k|w_i), \quad (1)$$

whereby each stimulus  $w_i$  specifies a discrete distribution over  $K$  latent variables. Each latent variable itself corresponds to a probability distribution over the  $m$  possible response words and as such, these distributions will come to represent elementary patterns that characterize the inter-correlations in the association norms.

Although it is possible for each stimulus  $w_i$  to correspond to a distribution over the latent variables that is independent of the other stimulus words (e.g. each  $w_i$  could correspond to a  $K$  dimensional discrete distribution  $\pi_i$ ), it is also possible to represent  $P(x=k|w_i)$  inversely through Bayesian inference, i.e.

$$P(x=k|w_i) = \frac{P(w_i|x=k)P(x=k)}{\sum_{k'=1}^K P(w_i|x=k')P(x=k')}. \quad (2)$$

Doing so has at least three desirable consequences:

1. Every stimulus word is potentially a response word, and vice versa. There is no dichotomy between a stimulus word vocabulary and the response word vocabulary. All words are elements of a single vocabulary.
2. Any observation that, for example,  $w_j$  is a response to stimulus  $w_i$ , also potentially provides information about how  $w_j$  will act as a stimulus word, (i.e. what words are likely as responses given  $w_j$  as a stimulus word) or about how  $w_i$  will act as a response word (i.e. what stimulus words are likely to elicit  $w_i$  as a response)<sup>1</sup>.
3. This model is directly comparable to the probabilistic model described by Griffiths et al. (2007) that learns semantic representations from text-based distributional statistics. In that model, semantic knowledge is characterized by latent probability distributions that correspond to coherent semantic topics. Any given word's semantic representation is characterized as a distribution over these latent topics, and the inter-word relationship between word  $w_i$  and word  $w_j$  is characterized by the conditional probability of  $w_j$  given the semantic representation of  $w_i$ . As such, the latent distributions are identical in form to the  $K$  component distributions of the mixture model we employ here. The semantic representation of a given word is calculated in a manner (almost) identical to that given by Equation (2), and consequently the inter-word relationships are calculated in a manner (almost) identical to that in Equation (1).

The parallel between these models can arguably facilitate a thorough comparison of the semantic representations and knowledge acquired from text-based

<sup>1</sup>It should be emphasized, however, that this characteristic does *not* enforce any symmetry to the associations between words, which is something that is rarely observed in association norms.

distributional statistics with those as evidenced from word-association norms. As word association norms are routinely used to externally validate distributional models of semantic memory this is an important comparison to pursue. Given the space limitations of the current paper, we do not deal directly with this issue here, but instead reserve it as the primary area of application for future work.

## Fitting the Model to Association Norms Data

As is apparent from the description, the mixture model defined in Equations (1) and (2) are parameterized by two set of parameters:  $\phi = \phi_1, \phi_2 \dots \phi_k \dots \phi_K$  and  $\pi$ . Each  $\phi_k$  is a discrete probability distribution over the  $V$  words in the vocabulary<sup>2</sup> such that  $\phi_{kj} \doteq P(w_j|x=k)$ , and  $\pi$  is a discrete probability over  $K$  elements such that  $\pi_k \doteq P(x=k)$ .

As already mentioned, association norms such as the Nelson data-set take the form of stimulus-response word pairs, i.e. each participant is given a subset of stimulus words and for each of which they produce a single response word. Aggregating across subjects, we can represent the entire data-set  $\mathcal{D}$  as a set of a total of  $T$  stimulus-response word-pairs, i.e.  $\mathcal{D} = \{s_t, r_t\}_{t=1}^T$ . This manner of representations is essentially a convenience, and leads to no loss of data, i.e. the number of times response word  $w_j$  is elicited from stimulus  $w_i$  is  $A_{ij} = \sum_{t=1}^T \mathbb{I}(s_t = w_i, r_t = w_j)$ , where  $\mathbb{I}$  is an indicator function taking the value 1 if its argument is true, and 0 otherwise.

The objective of model fitting is to infer the posterior distribution over the parameters  $\phi$  and  $\pi$  given the association norms data-set  $\mathcal{D}$ , i.e.

$$P(\phi, \pi | \mathcal{D}) \propto P(\mathcal{D} | \phi, \pi) P(\phi, \pi), \quad (3)$$

where the likelihood function is

$$P(\mathcal{D} | \phi, \pi) = \prod_{t=1}^T P(s_t, r_t | \phi, \pi), \quad (4)$$

$$\begin{aligned} &= \prod_{t=1}^T \sum_{k=1}^K P(s_t, r_t | x_t = k, \phi) P(x_t = k | \pi), \\ &= \prod_{t=1}^T \sum_{k=1}^K P(s_t, | x_t = k, \phi) P(r_t | x_t = k, \phi) P(x_t = k | \pi). \end{aligned} \quad (5)$$

$$(6)$$

As is the case with any mixture model, the presence of the unobserved latent variables, i.e.  $x_1, x_2 \dots x_t \dots x_T$ , entails that the posterior in Equation (3) is analytically intractable and must be approximated by Monte Carlo sampling. An efficient method in this respect is the fol-

<sup>2</sup>Recall that we no longer make a distinction between a stimulus word vocabulary and response word vocabulary, and hence  $V$  is the total number of words that were either a stimulus or response word.

fire	two	idea	think	smoke	eggs	dog	break	chocolate	friend
candle	one	words	mind	cigarette	egg	cat	glass	candy	enemy
burn	twice	opinion	understand	tobacco	chicken	kitten	fragile	icecream	husband
flame	once	dictionary	know	ashtray	duck	animal	can	vanilla	wife
blaze	single	word	brain	cigar	goose	puppy	jar	flavor	boyfriend
wick	double	sentence	thought	ash	hen	hound	breakable	taste	girlfriend
inferno	three	thought	comprehend	smoking	quack	panther	snap	sweet	pal
fireplace	couple	phrase	aware	pipe	yolk	meow	shatter	fudge	buddy
torch	pair	theory	ponder	nicotine	ducks	pets	opener	cone	foe
hot	four	thesaurus	knowledge	lighter	geese	pet	fracture	yogurt	spouse
wood	twin	vocabulary	alert	ashes	bacon	leopard	plates	cookie	companion
log	individual	notion	concentrate	cancer	omelet	bark	dishes	lollipop	guest
match	unique	hypothesis	awareness	puff	rooster	cougar	delicate	treat	visitor
chimney	number	saying	wonder	dust	scrambled	paw	crack	mint	partner
wax	alone	statement	attention	chimney	gander	stray	window	mousse	lover
hell	solo	concept	learn	butt	pond	wolf	crackle	pudding	mate
smoke	unit	paragraph	conscious	lung	scramble	animals	crunch	bar	adversary
aim	chapter	definition	realize	camel	chick	breed	cookie	dessert	opponent
ashes	duplicate	theme	intuition	habit	boiled	tiger	bottle	milk	neighbor
ash	zero	fact	awake	soot	fried	flea	lid	good	trust

Table 1: A random sample of 10 latent component distributions from a model with a total of  $K = 300$  components. Note that each component distribution is a probability distribution over the entire vocabulary, but each will place the majority of its probability mass on a small subset of words. Shown in this table are (in descending order) the 20 words in each component with the highest probability. It is evident that each component is a highly specific and coherent semantic topic in a sense directly analogous to the discourse topics learned from text corpora in the probabilistic model described in Griffiths et al. (2007).

lowing Gibbs sampler: Arbitrarily choose initial parameter values  $\tilde{\phi}^0, \tilde{\pi}^0$  and then for  $\tau \geq 0$  iteratively sample

$$\tilde{x}_{1:T}^\tau \sim P(x_{1:T} | \mathcal{D}, \tilde{\phi}^\tau, \tilde{\pi}^\tau), \quad (7)$$

$$\tilde{\phi}^{\tau+1}, \tilde{\pi}^{\tau+1} \sim P(\phi, \pi | \mathcal{D}, \tilde{x}_{1:T}^\tau). \quad (8)$$

Using a noninformative Dirichlet prior on the parameters, both of these steps are tractable to compute and upon convergence, will lead to samples from the joint posterior over the parameters and latent variables given the training data. From this it is then a trivial matter to obtain samples from the posterior described in Equation (3).

## Analysis of the Inferred Model

For what follows, we exclusively use the Nelson norms as the data from which models are inferred. This set is large, publicly available and widely studied and an obvious choice as an initial training set. In addition, we restricted our vocabulary to words that occurred at least once as both a stimulus and a response word. With this restriction the vocabulary size was 4864, and the total number of stimulus-response training pairing  $T$  was 561,236 (In the original Nelson set there are 5018 stimulus word types, and a total of 612,627 stimulus-response training pairs). This restriction reduced the computational burden in simulations, but lead to no discernible differences in the results.

A preliminary issue of both practical and theoretical interest is to determine the number of component latent distributions  $K$  for the model that is fit to the Nelson data. It is of practical importance as different numbers of mixture components may dramatically affect the nature of the inferred model: Too few components and

the model will not adequately capture the underlying structure in the data (i.e. underfitting); too many and the model will capture spurious or idiosyncratic features of the data and will consequently generalize poorly (i.e. overfitting). It is of theoretical importance as the number of components that best fit the data will, to a certain extent, indicate the level of statistical redundancy in the data. This can be potentially informative when comparing similar models using different training data sets.

Determining the number of mixture components is essentially a matter of model selection. There are many potential solutions to this, but by far the simplest is to use cross-validation, i.e. withhold a fraction of the training data, and then test the inferred models of different sizes on their generalization to this withheld set. In Figure 1, we show the average log-likelihood of the reserved fraction of the training set (averaged over a set of parameter samples drawn from the posterior distribution) for a range of different model sizes. As is evident, anything less than  $K = 250$  will poorly generalize, the range  $300 \leq K \leq 500$  lead to roughly comparable performance (suggesting that model averaging over this range may be advisable), and as  $K$  becomes larger than 500, model generalization will again deteriorate.

In Table 1, we provide 10 randomly sampled components from a  $K = 300$  component model. As previously described, each component defines a distribution over the  $V$  words in the vocabulary, and in general will reserve most of its mass for a small subset of words. Shown in the table are (in descending order) the 20 words in each of the chosen components that have the highest probability mass. It is clear that each component defines a specific and coherent semantic topic. As such, the entire set of  $K = 300$  components define a core repertoire of semantic

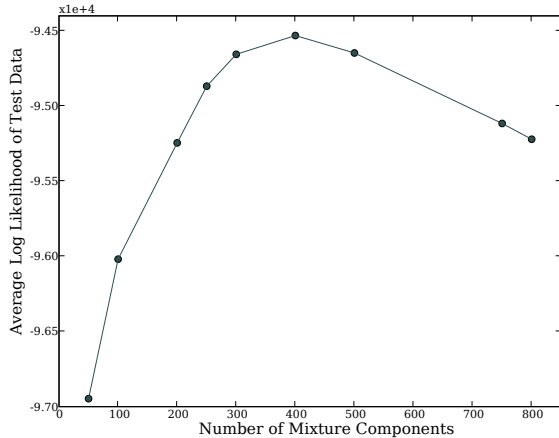


Figure 1: Cross-validation results used to determine the optimal number of model components  $K$ . The range  $300 \leq K \leq 500$  provides the best generalization performance.

patterns that can be regarded as the model’s semantic knowledge. These patterns arise from the web of inter-correlations in the association norms. As a single example, consider the second component (from the left) in Table 1. This component identifies quantification terms for low integers, a highly specific and coherent topic in itself. This topic arises as words such as *one*, *two*, *three*, etc. occur as either responses or eliciting stimuli to one another but also words such as *once*, *twice*, etc. These, in turn, will be paired with other words, such as *single*, *double*, *duplicate*, etc. This process of links effectively identifies a coherent web of interrelated concepts. Hence we can understand the components distributions of the model as identifying the core set of intercorrelated webs that underlie the association norms.

As mentioned, each word is represented as a distribution over the latent components. This distribution, which is inferred using Equation (2), can be seen as the word’s latent semantic representation, indicating the extent to which each component exemplifies the semantic character of the word. From this distribution, following Equation (1), it is a straightforward matter to simulate the production of word associations. Table 2 provides examples of simulated word associations for a set of 5 randomly chosen words. For the purposes of comparison, we also show the actual word association of these words as they occur in the Nelson data-set. From this, we see that the model extrapolates beyond the data, inferring associations that are not strictly given by the data but are reasonable inferences given the pattern of inter-correlations underlying the data. Thus, we see that the words *electric*, *shock*, *energy* and *volt*, etc., are highly predicted associates of the word *outlet* despite not being mentioned directly as such in the training data.

## Predicting False Recall Rates

A common application of word association norms is the prediction of so called false-recall rates. The false-recall experimental paradigm, and the use of word association norm for both its design and analysis, has been extensively employed by Deese, e.g. (Deese, 1965) and more recently by Roediger and McDermott (1995). In this paradigm, participants are tested on their recall of multiple word lists, each of approximately 15 words. These word list are specially constructed to relate to a critical word that is not explicitly listed. For example, a list including the words *cigarette*, *puff*, *blaze*, *fire*, etc. is constructed to relate to the unlisted critical word *smoke*. Typically, upon recall of the list items, the critical item is also, erroneously, recalled as being present on the list. It has been observed, however, that the rates of false-recall can differ substantially from item to item, with some critical words being falsely recalled as infrequently as 1% of the time, while others are recalled as often as 65% of the time (Roediger, Watson, McDermott, & Gallo, 2001).

The inferred model can simulate the extent to which any list of words will predict or imply a critical word. This proceeds, essentially, as an extension of the simulation of pairwise word associations, with the exception that instead of one word predicting an associate, we predict association based on an entire set of words. Thus, the extent to which the set  $w_0, w_1 \dots w_i \dots w_n$  predicts  $w_j$  is given by

$$\begin{aligned}
 &P(w_j | w_0, w_1 \dots w_i \dots w_n) \\
 &= \sum_{k=1}^K P(w_j | x = k) P(x = k | w_0, w_1 \dots w_i \dots w_n), \quad (9)
 \end{aligned}$$

where

$$P(x = k | w_0, w_1 \dots w_i \dots w_n) \propto \prod_{i=1}^n P(w_i | x = k) P(x = k). \quad (10)$$

Using this method, we can apply it to the prediction of false recall rates. Roediger et al. (2001) provide 55 lists along with the recorded rates at which they induce the false recall of their critical words. In Figure (2), we provide a scatter-plot of the probability of the critical word according to the model ( $x$ -axis) against the observed rate of false recall ( $y$ -axis). Superimposed on the scatter-plot is the line of best-fit. This correlation coefficient is  $r = .747$ , which is highly significant ( $p = 5.76 \times 10^{-11}$ ).

**Digression** While the accuracy of the model’s prediction is encouraging, it must be mentioned that using the raw data alone and calculating the mean associative strength of the list items to the critical word will also lead to a high correlation with the observed false-recall rates. Using the raw data upon which the model was trained (i.e. a somewhat restricted subset of the Nelson norms), and calculating the mean associative strength between each list and its critical word, we observe a correlation

1. **announce:**
  - (a) tell, speak, say, speaker, yell, engagement, microphone, inform, loud, marriage, wedding, state, message, public, radio, talk
  - (b) talk, tell, speak, answer, yell, say, speech, question, loud, marriage, conversation, scream, inform, communicate, notify, listen, communication, discuss, discussion, noise, hear, ear, announce, comment, lecture
2. **medium:**
  - (a) large, middle, small, average, size, mediocre, half, rare, regular, well, mean, paint
  - (b) small, big, large, little, normal, huge, average, tiny, regular, medium, usual, ordinary, size, common, norm, boring, same, shrink, middle, weird, strange, abnormal, grand, routine, giant
3. **outlet:**
  - (a) plug, electricity, inlet, map, socket, store, light, exit, power, shopping, cord, wall
  - (b) electricity, plug, light, outlet, electric, shock, energy, socket, volt, stopper, battery, inlet, inside, watt, television, cord, power, bulb, outside, wire, kinetic, electrician, indoors, outdoors, amp
4. **star:**
  - (a) sky, moon, sun, bright, shine, light, night, wish, galaxy, twinkle, astronomy, planet, space, dream
  - (b) star, space, moon, stars, astronomy, astronaut, astrology, sky, shuttle, telescope, planet, meteor, twinkle, asteroid, rocket, meteorite, comet, galaxy, forever, universe, crescent, crater, off, sun, launch
5. **wear:**
  - (a) clothes, tear, cloth, dress, remove, sweater, use, wash, worn,
  - (b) clothes, shirt, pants, dress, jeans, coat, jacket, blouse, skirt, wear, wardrobe, attire, trousers, hanger, outfit, fashion, denim, closet, shorts, cloth, sleeve, suit, zipper, button, designer

Table 2: A simulation of word-association. For each of a set of 5 randomly chosen words from a model with  $K = 300$  components, we list (a) its associates as given by Nelson norms (in decreasing order of associative strength), and (b) the 25 top inferred associates by the model (again, in decreasing order). The model clearly extrapolates well, and infers reasonable associates other than those strictly given by the data.

of  $r = .699$  with the false recall rates. Although this is lower than the model's performance, it nonetheless implies that the raw data themselves are not substantially worse than the inferred model. Moreover, Roediger et al. (2001) report a correlation of  $r = .73$  between the false recall rates and the mean associative strength calculated using an augmented version of the Nelson norms.

However, it is reasonable to argue that the mean associative strength works well as a predictor of false-recall only in special cases. More generally, the mean associative strength of a list of word can be highly misleading. We illustrate this in Table (3). The two lists that are shown have identical mean associative strengths with the critical word *whiskey*. The list on the left is obtained from Roediger et al. (2001), while the list on the right was randomly generated from the Nelson norms to have an identical mean associative strength to *whiskey* as the list on the left. It is reasonable to presume that the list on the left will be more likely to induce a false recall of *whiskey* than the list to the right. Indeed, using an inferred model with  $K = 500$  components, we see that *whiskey* is the most highly ranked prediction given the left list, yet not highly ranked by the right list.

## Discussion

In this paper, we introduced a method for the analysis of word association norms, the central feature of which is the inference of a probabilistic semantic memory model that is taken to be the generative model of the word associations. This method can be related to previously employed methods of analysis such as factor analysis,

singular value decomposition of the matrix of word associations, or multidimensional scaling of dissimilarity matrices derived from the association matrix. However, it also has notable points of departure from this methods and characteristics that independently motivate its use. In particular,

- This method treats the data appropriately as counts of discrete values. Word associations are naturally viewed as samples from a discrete conditional probability distribution. This is starting point for the derivation of the model that we advocate. By contrast, methods such as factor analysis, and principal components analysis are not ideally suited to dealing with multinomial data, and are based upon the assumption that the data can be characterized as multidimensional Gaussian random variables.
- The component distributions of the inferred model represent its semantic knowledge. These distributions are readily interpretable as coherent semantic topics. Inspection of the examples provided by Table (1) shows that it is intuitively obvious what information these components capture. Such intuitive interpretations are not as forthcoming with methods such as, for example, factor analysis.
- The Bayesian approach to parameter inference in the model, i.e. posterior inference of parameters, ensures that all information available in the data is used to infer parameters. By contrast, methods that rely on maximum likelihood estimation do not take into ac-

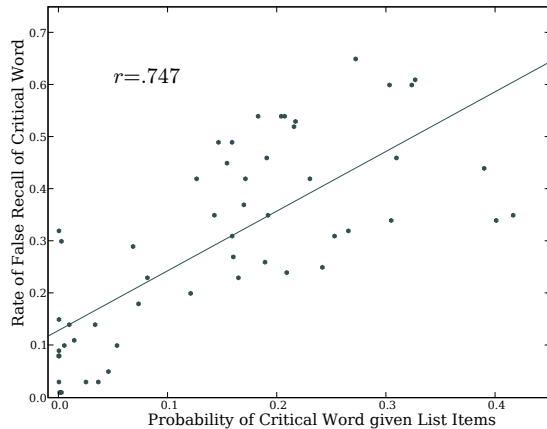


Figure 2: Predicting false recall rates. For each of 55 sets of stimuli for a false-recall experiment, we show the model’s inferred probability of the critical word given the list of associates plotted against the rates of false-recall of this critical word by human participants. The model is highly accurate at predicting the occurrence of false recalls.

count uncertainty in parameter estimation and may routinely overfit the data.

- The probabilistic nature of the model greatly facilitates the analysis of the model. This is particularly the case for the simulation of new word associations, or the simulation of the false recall. The manner in which these simulations are performed is generally straightforward to derive given the probabilistic relationships between the variables relevant to the model.
- A final notable characteristic of this approach is that it is directly comparable to a recently introduced probabilistic model of semantic memory known as the Topics-model Griffiths et al. (2007). The topics model is a Bayesian model of semantic structures that are inferred from the distributional characteristics of words in a large text corpus. The inter-word relationships derived from this model are often externally validated using word association norms. By contrast, in the approach we pursue here, we infer an (in important respects) identical Bayesian model *directly* from association norms. By characterizing the semantic structures revealed by the association norms, this approach may be particularly useful to thoroughly analyze the nature and validity of the semantic structures that are learned from texts by the topic models. We intend this to be the primary area of application for future work.

## References

Deese, J. (1962). On the structure of associative meaning. *Psychological Review*, 69(2), 161-175.

drink	edge
drunk	viola
beer	strict
liquor	care
gin	gin
alcohol	architecture
rye	cloak
glass	delay
wine	camera
rum	rum
bourbon	bourbon
evil	male
bar	genius
scotch	scotch

Table 3: An example of how mean associative strength in the raw data of the Nelson norms is inadequate as a general measure of how sets of words will induce false recalls. The two lists shown here have identical mean associative strengths with the word *whiskey*. It is reasonable to presume that the list on the left will be more likely to induce a false recall of *whiskey* than the list to the right.

Deese, J. (1965). *The structure of associations in language and thought*. Baltimore, MD: Johns Hopkins Press.

Galton, F. (1879). Psychometric experiments. *Brain*, 2, 149-162.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114(2), 211-244.

Nelson, D., McEvoy, C., & Schreiber, T. (1999). *Word association norms*, <http://w3.usf.edu/FreeAssociation/>.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21(4), 803-814.

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. a. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, 8(3), 385-407.

Steyvers, M., Shiffrin, R., & Nelson, D. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications*. Washington,DC: American Psychological Association.