Interpreting semantic clustering effects in free recall

Jeremy R. Manning1,2 and Michael J. Kahana3

1Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA
2Department of Computer Science, Princeton University, Princeton, NJ, USA
3Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA

The order in which participants choose to recall words from a studied list of randomly selected words provides insights into how memories of the words are represented, organised, and retrieved. One pervasive finding is that when a pair of semantically related words (e.g., “cat” and “dog”) is embedded in the studied list, the related words are often recalled successively. This tendency to successively recall semantically related words is termed semantic clustering (Bousfield, 1953; Bousfield & Sedgewick, 1944; Cofer, Bruce, & Reicher, 1966). Measuring semantic clustering effects requires making assumptions about which words participants consider to be similar in meaning. However, it is often difficult to gain insights into individual participants’ internal semantic models, and for this reason researchers typically rely on standardised semantic similarity metrics. Here we use simulations to gain insights into the expected magnitudes of semantic clustering effects given systematic differences between participants’ internal similarity models and the similarity metric used to quantify the degree of semantic clustering. Our results provide a number of useful insights into the interpretation of semantic clustering effects in free recall.

Keywords: Free recall; Clustering; Semantic; Simulation; Latent semantic analysis.

The free recall paradigm has participants study lists of items—typically words—and subsequently recall the studied items in the order they come to mind. Because the participants are instructed to recall the items in the order they come to mind, the recall sequence reflects how the items are stored and retrieved from memory. By analysing recall sequences during free recall researchers have uncovered a number of trends that many participants exhibit. For example, the recency and primacy effects refer to the well-established tendency of participants to show superior recall of items from the ends, and to a lesser extent from the beginnings, of the studied lists (Deese & Kaufman, 1957; Murdock, 1962). Another well-studied phenomenon, termed temporal clustering, refers to participants’ tendencies to successively recall items that occupied neighbouring positions in the studied lists (Kahana, 1996). In addition to ordering recalls by the study positions of the items, participants also exhibit striking effects of semantic clustering (Bousfield, 1953; Bousfield & Sedgewick, 1944; Cofer et al., 1966; Jenkins & Russell, 1952; Romney, Brewer, & Batchelder, 1993), whereby recall of a given item is more likely to be followed by recall of a similar or related item than a dissimilar or unrelated one.

The primacy, recency, and temporal clustering effects may be measured objectively by examining the relative probabilities of recalling or...
transitioning between items that appeared at each serial position on a studied list. By contrast, measuring semantic clustering requires making assumptions about what each word means to each participant. For example, by one metric, successively recalling the words “dog” and “collar” might serve as evidence of semantic clustering, since the two words might be expected to appear in similar contexts: the two words might be described as falling under the general category of “things related to common household pets”. However, by another metric, successively recalling “dog” and “collar” might serve as evidence against semantic clustering, because dogs should fall under the category of “mammals” whereas collars should fall under the category of “inanimate objects”. These issues are further complicated if one considers that the associations an individual forms between words, and the meanings ascribed to the words, likely depend on that individual’s subjective experiences.

Over the past decade a number of techniques have been developed for systematically quantifying the relative meanings of words. **Latent semantic analysis** (LSA; Landauer & Dumais, 1997) derives a set of pairwise similarity values by examining the co-occurrences of words in a large text corpus. Another measure of semantic similarity, termed the **Google similarity distance** (Calibrasi & Vitanyi, 2005), uses the Google search engine to compute the number of web pages containing both word x and y, relative to the total number of pages containing each word individually; a similar metric relies on Wikipedia links to measure the similarities between topics (Milne & Witten, 2008). A fourth technique, **Word association spaces** (WAS; Nelson, McEvoy, & Schreiber, 2004; Steyvers, Shifrin, & Nelson, 2004) derives its similarity values from a series of free association experiments in which participants were given a cue item and responded with the first word that came to mind. The goal of each of these techniques is to compute a set of pairwise similarities between the words that bears some resemblance to the similarities ascribed by a “typical” person.

Although the similarity values produced by each of these myriad similarity metrics are somewhat related, the pairwise correlations between the measures tend to be surprisingly low. For example, for the set of 308 highly imageable nouns listed in Table 1, the Pearson’s correlation between the LSA- and WAS-derived pairwise semantic similarity values is $r = 0.23$ (Spearman’s $\rho = 0.18$). The full distributions of similarity values derived from the two metrics are shown in Figure 1, Panels A and B. The relation between LSA and WAS similarity is illustrated in Panels C and D. If these seemingly objective semantic similarity metrics based on huge text corpora and experimental datasets fail to agree on a set of pairwise semantic similarities, how could one possibly expect to study effects of semantic organisation in individual participants? In particular, how should the magnitudes of semantic clustering effects be interpreted? In the present manuscript we use simulations to study these questions.

**METHODS**

**Analysis**

Our simulations are intended to estimate the maximum expected magnitude of semantic clustering effects in free recall. Our approach is motivated by the notion that, although we may measure the degree of semantic clustering using semantic similarity metric $f()$, the semantic similarity metric that would have best described a given participant’s “true” internal semantic similarity model is a different metric, $g_p()$. Assuming that the participant exhibits perfect semantic clustering according to metric $g_p()$, should we expect that the participant would also exhibit reliable semantic clustering effects according to metric $f()$?

In most free recall studies $g_p()$ is unknown. In theory one could estimate $g_p()$ for a given participant by having the participant make judgements about the semantic similarities between each pair of studied words. Then one could use $g_p()$ to quantify semantic clustering in that participant’s recall sequences. However, this method becomes impractical as the number of study items grows, since the number of pairwise comparisons grows with the square of the number of study items.

Rather than derive each participant’s $g_p()$ empirically, we instead construct a pool of 100,000 simulated participants, whose $g_p()$’s are known. Each simulated participant encounters many word lists, and we simulate a sequence of

---

1 Here the functions $f()$ and $g_p()$ are mappings from two words, $a$ and $b$, onto scalar similarity values. Note that $f()$ and $g_p()$ need not produce the same mapping. The subscript $p$ serves as a reminder that participants’ true internal similarity models may differ across individuals.
recalls after each studied list. As described below, the recall sequences are constructed to maximise semantic clustering—according to \( g_p() \)—for each participant. We then measure the degree of semantic clustering according to a different similarity metric, \( f() \). We quantify the degree of semantic clustering using the \textit{semantic clustering score} (Polyn, Norman, & Kahana, 2009), described in the next section. The distribution of semantic clustering scores according to \( f() \) tell us about the range of semantic clustering scores we should expect to observe in real participants, given that we use the “wrong” semantic similarity model to measure semantic clustering.

\textbf{Semantic clustering score}

The semantic clustering score, developed by Polyn et al. (2009), is intended to quantify the extent to which a given recall sequence shows evidence for semantic clustering (according to metric \( f() \)), taking into account the set of words that appeared on the studied list. For each recall transition we create a distribution of semantic similarity values—using \( f() \)—between the just-recalled word and the set of studied words that have not yet been recalled. We next generate a percentile score by comparing the semantic similarity value corresponding to the next item in the recall sequence with the rest of the distribution. Specifically, we calculate the proportion of the possible similarity values that the observed value is greater than, since strong semantic clustering will cause the observed similarity values to be larger than average. When there is a tie we score this as the percentile falling halfway between the two items. In this way if a participant always chose the closest semantic associate, then their semantic clustering score would be 1. A semantic clustering score of 0.5 indicates no effect of semantic clustering. The semantic clustering score must be computed independently for each studied list. We obtain a single semantic clustering score for each simulated participant by averaging the semantic clustering scores across all lists that the participant encountered.

\textbf{Generating recall sequences that maximise the semantic clustering score}

As defined above, the semantic clustering score according to metric \( g_p() \) is maximised (i.e., equal to 1) if the participant always chooses to next recall the closest semantic associate to the just-recalled word. Suppose the simulated participant has just studied a list of \( n \) words. We would now like to generate a \( k \)-item recall sequence (where \( k \leq n \)) that maximises the semantic clustering score according to \( g_p() \).

We begin by selecting the first recalled word, \( i_1 \), at random from the set of \( n \) studied words. We then create a pool of the \( n - 1 \) remaining words from the studied list. We order the words in the pool by their semantic similarity—according to \( g_p() \)—to \( i_1 \). We select the word with the highest semantic similarity as the next recall, \( i_2 \), and remove \( i_2 \) from the pool. We then re-order the \( n - 2 \) remaining words in the pool by their semantic similarities to \( i_2 \) and select the word most similar to \( i_2 \) to be recalled next. This process continues until the \( k^{th} \) word is recalled. Because this procedure ensures that each recall will be followed by the most similar word that is yet to be recalled, by definition it will maximise the semantic clustering score according to \( g_p() \).

\textbf{RESULTS}

We ran two batches of simulations. In the first batch we constructed \( g_p() \)s for each of 100,000 simulated participants according to the LSA-derived similarities between each pair of words in Table 1. We computed each word’s LSA vector by applying the LSA algorithm (Landauer & Dumais, 1997) to the Touchstone Applied Science Associates, Inc. (TASA) corpus. We then computed the similarity between each pair of words by measuring the cosine of the angle between the corresponding LSA vectors. In our simulations, all of the \( g_p() \)s were identical, and \( g_p(x,y) \) corresponded to the cosine of the angle between the LSA vectors for \( x \) and \( y \). For each participant we also constructed 50 lists of 15 unique items each, drawn from the word pool. We then generated five-item recall sequences after each list that maximised each participant’s semantic clustering scores according to LSA (see \textit{Generating recall sequences that maximise the semantic clustering score}). Finally, we computed each participant’s mean semantic clustering score using WAS similarity (Nelson et al., 2004; Steyvers et al., 2004).

---

\footnote{For all of the simulations reported in this manuscript, we used \( n = 15 \) and \( k = 5 \). However, the techniques developed here are equally applicable to arbitrary choices of \( n \) and \( k \).}
The distribution of simulated WAS-derived semantic clustering scores is shown in Figure 1E. We found that the mean semantic clustering score was 0.636. The analysis yields the distribution of maximum expected semantic clustering scores (computed using WAS similarity), given that participants’ “true” internal models of semantic similarity are perfectly described by LSA.

The second batch of simulations used the identical set of 15-item lists, presented to the same simulated participants. However, for the second batch of simulations we generated recall sequences that maximised the semantic clustering scores according to WAS-derived similarity. We then measured each participant’s mean semantic clustering score using LSA-derived similarity.
The distribution of simulated LSA-derived semantic clustering scores is shown in Figure 1F. We found that the mean semantic clustering score was 0.662. This second batch of simulations yields the distribution of maximum expected semantic clustering scores (computed using LSA similarity), given that participants’ “true” internal models of semantic similarity are perfectly described by WAS.

We also found that the semantic clustering scores computed using LSA were slightly (but reliably) higher than those computed using WAS—paired t-test: t(99,999) = 270.65, p < 10^{-6}; mean difference: 0.026. This indicates that different semantic similarity metrics used in analyses of semantic clustering may introduce slight biases. We expect that these biases are related to the form of the semantic similarity distributions derived from each measure (see Figure 1) and to the particulars of how each measure is derived. Given that the clustering scores obtained using any given model of semantic similarity are likely to be only noisy reflections of any true patterns in the data, one should use multiple models of semantic similarity whenever possible. If one observes (or fails to observe) a similar pattern of clustering scores across experimental conditions when using multiple semantic similarity models (e.g., LSA and WAS), then it is less likely that the observed patterns simply reflect the mismatches between participants’ internal similarity models and the similarities assumed by the scoring model. Note that our analysis makes no attempt to distinguish whether the LSA- or WAS-derived similarities more accurately reflect participants’ internal similarity models.

Across the 200,000 simulated recall sequences, and combining across the two semantic similarity measures, the observed semantic clustering scores ranged from 0.522 to 0.757. These scores reflect the range of maximum clustering scores one would expect, given that participants’ internal semantic similarity models differed systematically from the similarity measure used to quantify the...
degree of semantic clustering. This shows that even participants who exhibit strong semantic clustering may still show clustering scores near 0.5. Similarly, it is exceedingly unlikely that one would observe semantic clustering scores near 1 when aggregating over many lists, as this would suggest a near-perfect match between the participant’s internal similarity model and the (arbitrarily chosen) scoring model.

DISCUSSION

Our simulations yield four valuable insights into the interpretation of semantic clustering during free recall. First, it is important to use multiple measures of semantic similarity if one is to obtain an accurate estimate of whether participants are semantically clustering their recalls. Second, an observed near-chance clustering score does not necessarily indicate a true lack of semantic clustering, but may instead indicate a mismatch between a participant’s internal similarity model and the scoring model. For this reason the precise clustering score one observes is difficult to interpret, and one would be better served by instead comparing distributions of clustering scores obtained across conditions in an experiment or across participants. Third, an observed near-ceiling clustering score (> 0.757 by our simulations) must be interpreted with caution, as it is unlikely for the scoring model to precisely match participants’ internal model of semantic similarity. Rather, a near-ceiling clustering score may reflect the specific sequence of words presented to the participant, or the specific structure of the experiment. In such cases one might use simulations analogous to those we present here to gain insights into the range of clustering scores one might expect under various models (e.g., high semantic clustering vs low semantic clustering). Fourth, when fitting computational models that aim to predict semantic clustering, it is important to take the potential mismatch between participants’ internal similarity models and the scoring model into account. One might accomplish this by using, for example, an LSA-derived scoring model while using a WAS-derived internal similarity model in their simulation (or vice versa) as we have done here.

We have focused on a single semantic clustering metric, the semantic clustering score (Polyn et al., 2009), and two semantic similarity metrics, LSA (Landauer & Dumais, 1997) and WAS (Nelson et al., 2004; Steyvers et al., 2004). Our use of these metrics is not intended to imply that they are the only, or even necessarily the best, such measures. Rather, we simply found the semantic clustering score to provide a convenient means of quantifying semantic clustering. We chose the two semantic similarity metrics as representative examples from the broader range of metrics discussed in the introduction. LSA represents one technique for deriving similarity values via automated text processing. By contrast, WAS derives similarity values using experimental data from psychological experiments. The specific choice of clustering and similarity metrics used in analyses of experimental data should reflect the goals of the experiment and/or analyses.

In addition to measuring participants’ tendencies to semantically cluster their recalls, a number of recent studies have begun to examine how individual words are represented by measuring the patterns of neural activity evoked when a word or image is viewed (e.g., Mitchell et al., 2008; Just, Cherkassky, Aryal, & Mitchell, 2010; Shinkareva et al., 2008). There is some evidence that similarities in the neural patterns evoked by thinking about a given pair of words predict the tendencies of participants to successively recall the words, given that both appeared on the studied lists (Manning, Sperling, Sharan, Rosenberg, & Kahana, in press). In this way one might objectively infer each participant’s internal semantic similarity model by measuring their neural activity as they studied and recalled list items.

Studying semantic clustering effects requires making assumptions about participants’ internal semantic similarity models. In the absence of pairwise judgements or neural data, researchers must rely on measures that attempt to capture the semantic relations between words without knowing the specifics of participants’ subjective experiences, or about the way their brains represent the words. Our simulations demonstrate the degree of semantic clustering that can be expected, given that the semantic model used to measure the clustering effects is not a perfect match for participants’ internal models.
REFERENCES


