Strong structure in weak semantic similarity: A graph based account

Simon De Deyne (simon.dedeyne@psy.kuleuven.be)[a, b]
Daniel J. Navarro (daniel.navarro@adelaide.edu.au)[b]
Amy Perfors (amy.perfors@adelaide.edu.au)[b]
Gert Storms (gert.storms@psy.kuleuven.be)[a]

[a] University of Leuven, Department of Psychology, Tiensestraat 102, 3000 Leuven, Belgium
[b] University of Adelaide, School of Psychology, 5005 Adelaide, Australia

Abstract

Research into word meaning and similarity structure typically focus on highly related entities like CATS and MICE. However, most items in the world are only weakly related. Does our representation of the world encode any information about these weak relationships? Using a three-alternative forced-choice similarity task, we investigate to what extent people agree on the relationships underlying words that are only weakly related. These experiments show systematic preferences about which items are perceived as most similar. A similarity measure based on semantic network graphs gives a good account for human ratings of weak similarity.
Keywords: similarity; semantic networks; word associations.

Although similarity is a fundamental concept in cognitive science, it is still not yet well understood. Any two entities have a potentially infinite number of features or predicates in common, making it always possible to construct *post hoc* explanations for why any items are similar to each other (Goodman, 1972; Medin, Goldstone, & Gentner, 1993). Even if similarity is logically vacuous, of course, it is not necessarily psychologically vacuous: there may indeed be a small or at least finite number of shared *represented* predicates (Medin & Ortony, 1989). However, while shared representations may well explain why people share clear intuitions about the similarity of strongly related items like CATS and MICE, the notion of shared representations may not apply when the items are only weakly related. After all, the only predicates that apply to such disparate items as RAINBOW and TUNAFISH are so vague and generic that appealing to them to explain similarity begins to make it nearly as underconstrained psychologically as Goodman first showed it was in a logical sense.

Despite the questions that weak similarity raises about the nature of our underlying mental representations, it is almost entirely unstudied. Almost all investigations into stimulus similarity have focused on items that tend to be quite similar to one another – we ask people to compare the similarity of CATS to MICE, or of MICE and MEN. Rarely if ever do we ask people questions about weak similarities. We can get a sense of how extreme this bias is by examining the empirical data for a set of 372 concepts belonging to 15 natural categories (e.g., fruit, tools, sports), as in Ruts et al. (2004). We used numerical methods to calculate theoretical values for the similarities between all pairs of words in a database of 12,000 word associations. Comparing the two, we found that the *weakest* similarities for which we have empirical data were *stronger* than 97% of the similarities that were predicted according to the word association data base. This suggests that research into similarity has focused almost exclusively on similarities between only the most related items.

From a methodological point of view, this is not surprising: if asked to rate how similar HAIL and TEACHER are to each other, most people would struggle to know how to answer. Yet this struggle does not necessarily imply that no underlying representation of similarity exists. As Goodman (1972) and others have pointed out, it is always possible to find some basis for saying that HAIL and TEACHER are similar. The real question is which of these bases form part of human mental representations, and whether there exist any systematic regularities in how people spontaneously assess these weak relationships. The goal in this paper is to investigate (a) whether these regularities exist, and (b) whether they can be accommodated by existing theories of semantic representation.

Viewed as a problem of rating the stimuli between two entities that are only weakly related, the challenge seems intractable. Intuitively it feels like the the similarity between HAIL and TEACHER is zero, and there is little underlying structure to be found. However, suppose the task were framed as a three-alternative forced-choice problem (e.g., Navarro & Lee, 2002). Which of the following three concepts is the odd one out: CUP, TEACHER and HAIL? Framed in this fashion, the problem seems less intractable, and many people have very strong intuitions about what the answer should be. Sometimes the intuition can be so strong that it may be difficult to see why the answer to the question is not obvious.

As an illustration, in our discussions of this specific triple, one author strongly felt that TEACHER was obviously the odd one out because teachers are people and the other two are not (an "animate vs inanimate" distinction). Another strongly felt that HAIL is the odd one out because it is a mass noun and the other two are count nouns (a "things vs stuff" distinction). In both cases the choice also invokes quite abstract ontological categories, and relies on very broad general knowledge about the world. Obviously the decision to rely on a particular category to guide the decision making is the result of "on the fly" reasoning about the items. Although nobody felt that CUP was the odd one out, it is interesting that for both authors the intuitions were quite strong, so much so that they were somewhat surprised to discover that the supposedly "obvious" choice was not, in fact, so obvious.

This leaves us with an open question: how deep does the structure in our mental representations go? One possibility is that there is significant agreement and constraint in our mental representations only when considering the relationship between entities that are strongly related to each other.

In other words, the Medin and Ortony (1989) argument about shared predicates may only apply between items that are already highly related. If this is the case, then one might expect Goodman's problem to arise when we try to measure weak similarities, causing each person's judgment to be essentially arbitrary and there to be few stable preferences across people. The other possibility is that there is enough shared structure in our mental representations that there is a strong agreement even for such strange pairings as RAINBOW and TUNAFISH, HAIL and TEACHER and so on.

In the first half of this paper we present two experiments exploring weak similarity structure in humans. We show that similarity ratings of weakly related items are nevertheless surprisingly regular across people, and moreover that similarity judgments can be manipulated in sensible ways. In the second half, we investigate the nature of the underlying representations that might give rise to these similarity judgments. Computational modelling demonstrates that weak similarities like those found in our experiments can be at least partially captured by semantic network models constructed from word association data.

Experiment 1

Our main goal in this experiment was to investigate whether people reliably agree in their similarity judgments even between weakly related entities. In order to avoid the difficulties inherent in asking for similarity ratings between two very different items, we had participants choose which pair out of three possibles pairs in a triple was the most similar one.

Method

Participants Sixty-nine native Dutch speaking psychology students participated in exchange for course credit.

Stimuli and Materials The stimuli were 300 nouns taken from a set of 12,000 Dutch words used as cues in the word association task described in De Deyne and Storms (2008b) and De Deyne, Voorspoels, Verheyen, Navarro, and Storms (2011).[1] These items were used to produce triples, which were sampled at random given two constraints. Each item in a triple was required to have approximately the same frequency and imageability rating, in order to ensure that participant responses reflected underlying semantic relatedness rather than superficial similarities in concreteness or familiarity. Word frequency was calculated based on the log-transformed lemma frequencies taken from the CELEX database (Baayen, Piepenbrock, & van Rijn, 1993), while imageability was derived from judgments on a seven-point scale found in De Deyne and Storms (2008a). Within any triple, the maximum standard deviation was $SD_{max} = 0.52$ for lemma frequency and $SD_{max} = 0.84$ for imageability.

Procedure On each trial three words were presented at the corners of a triangle, as shown in Figure 1. Participants were instructed to click on the circle corresponding to the side of the triangle that connected the most related pairs. We stressed in these instructions that we were interested in the meaning of words rather than their orthographic similarity or phonological relatedness. To illustrate what we meant, we gave participants two example triples: in the first one (COLD - HOT - SQUARE) the first two words are related, and in the second one (MOIST - COLD - COOL) the last two are. Participants were asked to do their best even if the task seemed difficult, and not to dwell too long on a single trial but to complete the task in a spontaneous manner. The task was presented as a web questionnaire during a collective testing session.

[1]The complete list of stimuli including English translations is available at `http://ppw.kuleuven.be/concat/simon/`

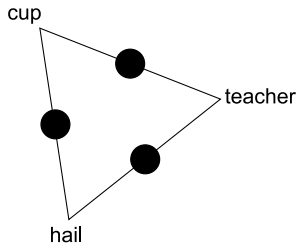

Figure 1: Example triple stimulus used in Experiment 1. The black circles indicate the controls used to select a pair with the mouse.

Results

Our key question was to what extent people tended to select the same pairs. If weak similarities do not exist or are not reliably shared by different people, we would expect all three possible pairs from every triple to be selected equally frequently. We test this in two different ways.

The first test of inter-rater reliability is to measure how often the most frequent pair from every triple is chosen. Since there are three possible pairs in any given triple, chance responding is 0.33. However, the median value was 0.67 – well above what one might expect by chance. Moreover, as Figure 2 illustrates, for 97 of the 100 triples the most commonly chosen response was selected significantly more frequently than would be expected by chance.[2]

Instead of just looking at the most frequent pair of any triple, we can also measure how much people's weak similarity judgments agree with one another in a more conventional way. We therefore ran χ^2 goodness-of-fit tests comparing the observed frequencies across the three responses to a null hypothesis that all three responses are equally likely for each triple separately. Taking this approach, the frequencies of 89 out of the 100 triples were significantly different from the null hypothesis, $\chi^2(2)$, $p < 0.05$.

The results so far suggest that people encode weak regularities from the environment and do this in a systematic and measurable way. How robust is this finding? We consider this question in the next experiment.

Experiment 2

The goal of this experiment is to investigate how robust the results from the first experiment are. If weak similarities are not "hard coded" in some way, then they must be derived or constructed somehow. Perhaps people are deriving them by searching a semantic network for the proximity of the two

[2]Note that the hypothesis tests here were conducted using a numerically simulated null distribution, since the sampling distribution of the maximum frequency is an extreme-value statistic and is not correctly described by a binomial distribution; it is, however, trivial to simulate numerically. Using this sampling distribution, the critical value was 0.39, corresponding to the cutoff shown in Figure 2.

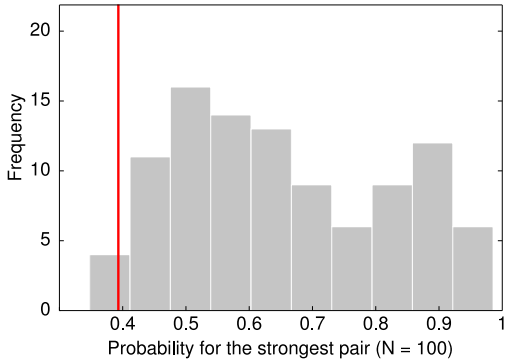

Figure 2: Distribution of the most frequently chosen pairs in Experiment 1. The vertical line indicates the 95% confidence boundary for the frequency one would expect if pairs were chosen randomly. Participants agreed with each other in selecting the pairs for almost all of the stimuli in the experiment.

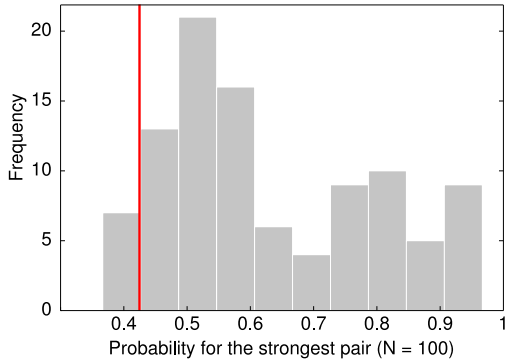

Figure 3: Distribution of the most frequently chosen pairs in Experiment 2. The vertical line indicates the 95% confidence boundary for the frequency one would expect if pairs were chosen randomly. As before, participants agreed with each other in selecting the pairs for most of the stimuli in the experiment. However, agreement was somewhat lower, suggesting that the time pressure made them unable to fully access their semantic representations, adding noise to their responses.

items to each other, or constructing them on the fly based on some other underlying representation. In either case, we would expect that time pressure would cause less accurate estimations and more disagreement between individuals, resulting in more uniform choice probabilities than were found in Experiment 1. We therefore repeated the experiment, with the variation that this time we put participants under time pressure by asking them to decide which pair is more related as quickly as possible.

Method

Participants Thirty native Dutch speaking students participated in exchange for course credits.

Stimuli and Materials The stimuli and materials were identical to those presented in Experiment 1.

Procedure The procedure was based on Experiment 1, but a few changes were made to allow for the accurate measurement of reaction times. Instead of using the mouse, participants were asked to use the keyboard, and to decide as quickly as possible which pair of words was related. At the beginning of each trial, the triple triangle was presented without the words until the participant pressed the space bar, which displayed the words at each corner. Unlike in Experiment 1, the black circles in Figure 1 were now labeled with either *J*, *K*, or *L*, and participants indicated which pair was most related by pressing the corresponding *J*, *K* or *L* key.

In order to make sure participants understood the task and were answering as quickly as possible, the main test was preceded by 20 practice items that had the words *word1*, *word2* and *word3* as labels at randomized locations. The participants were asked to click on the letter connecting word1 and word2 as quickly as possible. During this time a warning was shown when reaction times were slower than 3600ms, and participants were asked to try to make a faster response.

Results

Before evaluating what effect the time pressure manipulation had, we first needed to clean up the reaction time data. For each individual we therefore removed any responses with reaction times higher than three standard deviations above their average, as well as reaction times faster than 300ms. The average reaction time was 3771ms ($SD = 2131$). A log-transformation was used to reduce the skew in the reaction times. Next, for each participants the reaction times were transformed to *z*-scores, resulting in a Spearman-Brown split reliability (*z*RT) of 0.83. Since we did not record reaction times in the first experiment, it is not certain that the participants actually payed attention to the instructions and responding faster, as asked. We investigated if this was the case by running 18 new participants in Experiment 1, this time measuring their reaction times registered by keyboard response. The resulting reaction times had a mean of 4705ms ($SD = 2864$), about one full second slower than the speeded judgments in Experiment 2.

We can now explore the answer to our central question: what effect did time pressure have on the reliability of weak similarity judgments? As before, we can measure how often the most frequent pair from every triple was chosen. Remembering that chance responding would be reflected in a value of 0.33, we find a median value of 0.57. As predicted, this is lower than the 0.67 of Experiment 1, but higher than what one would find if responses were random. We also found that for the vast majority of triples (93 out of 100), the most common response was selected significantly more often than would be expected by chance. Figure 3 shows the distribution of responses. It is evident that, while putting people under time pressure increases the uniformity of the distribution of responses, there is still substantial agreement. This intuition is supported by the χ^2 goodness-of-fit tests done for each triple, which finds that the frequencies within 89 out of the 100 triplets were significantly different from a null hypothesis under which all three pairs would be equally likely ($\chi^2(2)$, $p < 0.05$).

The results of these two experiments demonstrate that there is consistent and reliable agreement on similarity judgments, even when the entities involved are only weakly related, like CUP and TEACHER. On some level, this agreement is surprising, because such items only share features if so many

features are represented that we begin to run afoul of Goodman's problem. Even then, it is unclear that the items on which there is agreement are the items with more shared features. A more likely explanation of this finding may be that people show strong agreement because they share the kind of semantic representation that is at least partially captured by a semantic network. In the next section, we explore this possibility by modelling the similarities from the two previous experiments based on semantic graphs.

Graph based models for weak similarity

In this section we investigate the hypothesis that at least some of the agreement between people about weak similarities arises due to shared semantic network representations. Network-based models for similarity have been proposed in related domains (e.g., the NETSCAL model by Hutchinson (1989) or the feature centrality model of Sloman and Rips (1998)), but the most similar models in psychology are the spreading activation models which accounted for a number of interesting semantic effects (e.g. Quillian, 1968).

Why might we expect semantic networks to capture some of the representation with which weak similarities are generated? Such networks probably reflect something about the way words are combined and used in the real world. For instance, the average American is exposed to about 100,500 words every day (Bohn & Short, 2009). The numerous ways that this vast amount of information can be combined may lead to an immense amount of mostly weak contingencies between items. Indeed, in recent years, the increasing availability of co-occurrence information to researchers has led to the development of models that derive representations of meaning from that co-occurrence. These models, which include latent semantic analysis (LSA, Landauer & Dumais, 1997) and topic models (Griffiths, Steyvers, & Tenenbaum, 2007), use statistical methods to extract the regularities underpinning the co-occurrence data. They thereby produce structured, meaningful representations that can be used to capture and explain human behavior and performance.

The goal behind network models is similar, though the approach is different. The network itself is derived from word associations which presumably reflect the patterns of co-occurrence in the world. We can then use the network as the core representation from which similarity measurements are derived. We theorize that although associations between individual entities may be too sparse to account for people's judgments about triples like CUP-TEACHER-HAIL, the network may capture broader relationships that *can* account for such judgments. If broad ontological distinctions like animacy or the count/mass noun distinction are reflected in the structure of the semantic network, we might expect a suitably chosen measure of network-based similarity to be able to capture, at least in part, the manner in which humans resolve the weak similarity questions that we asked in our experiments. How, though, can we measure similarity within a network? We address this problem in the next section.

Similarity in semantic networks

Large-scale semantic models are typically extremely sparse. In the case of networks derived based on word associations, this means that the number of edges connecting any two nodes (words) is very low. This is less of a problem when dealing with very similar concepts, because they are likely to share some edges despite the overall sparsity of the network. However, sparsity is a serious problem for other concepts. The same problem arises for non-network-based representations like feature overlap, because the number of features shared by weakly related items is very low, if not zero.

Given the problems imposed by sparsity, how can we measure similarity in a semantic network in a sensible way? We consider two different approaches here. The first is the widely used cosine measure of similarity (e.g., Landauer & Dumais, 1997; Steyvers, Shiffrin, & Nelson, 2004), which measures the extent to which two nodes in the graph share the same immediate neighbors. Two nodes that share no neighbors have a similarity of 0, and nodes that are linked to the exact same set of neighbors have similarity 1. Formally, it is defined as follows. Let $\mathbf{A}$ denote a weighted adjacency matrix, whose ij-th element a_{ij} contains a count of the number of times word j is given as an associate of word i in a word association task. Each row in $\mathbf{A}$ is therefore a vector containing the associate frequencies for word i. The cosine measure of similarity is the angle between these vectors, calculated as follows: because some words can have more associates than others, we normalize each row so that all of these vectors are of length 1. This gives us a new matrix $\mathbf{G}$, where $g_{ij} = a_{ij}/(\sum_j {a_{ij}}^2)^{1/2}$, and the matrix of all pairwise similarities is now

$$\mathbf{S} = \mathbf{G}\mathbf{G}^T \tag{1}$$

The key thing to recognize about the cosine measure is that it depends solely on the *local* structure of the graph: the similarities between two words is assessed by looking only at the words to which they are immediately linked.

Our second approach to similarity aims to take into account the overall structure of the entire network graph, and thus to reflect a broader view of the relationship between two nodes. This measure, similar to Leicht, Holme, and Newman (2006), is an example of a "random walk" approach to assessing similarity (see Kemeny & Snell, 1976; Van Dongen, 2000; Griffiths, Steyvers, & Firl, 2007, for related measures). In general terms, the idea is quite similar to the classical notion of spreading activation (e.g. Quillian, 1968). Similarity is thought to be related to the the number and length of the paths through the network that connect two nodes. If there are a lot of short paths that connect two nodes, then it is easy for a random walk through the graph to start at one node and end at the other; these nodes are therefore more similar. Formally, the measure is specified by beginning with the weighted adjacency matrix $\mathbf{A}$. This time, however, we normalize the rows so that each one expresses a probability distribution over words. That is, we use the matrix $\mathbf{P}$ where $p_{ij} = a_{ij}/\sum_j a_{ij}$, and then calculate

$$\mathbf{S} = (\mathbf{I} - \alpha\mathbf{P}^{-1}) \tag{2}$$

where $\mathbf{I}$ is a diagonal identity matrix and the α parameter governs the extent to which similarity scores are dominated by short paths or by longer paths. A path of length r is assigned a weight of α^r, so when $\alpha < 1$, longer paths get less weight than shorter ones.[3] Note that under this approach the

[3] As noted by Minkov (2008), this kind of mechanism can help avoid one of the major criticisms of the spreading activation mech-

similarities can be asymmetric (i.e., $s_{ij} = s_{ji}$). Since our experimental design forces the empirical similarities to be symmetric we use the average of $\mathbf{S}$ and $\mathbf{S}^T$ in our evaluations. Interestingly, our approach is very similar to the PageRank measure: $\mathbf{X} = (\mathbf{I} - \alpha\mathbf{P}^{-1})\mathbf{1}$. For PageRank it is standard practice to set α to a fixed value of 0.85 (Page, Brin, Motwani, & Winograd, 1998), where α is bounded between 0 and 1. Our choice of α was 0.6 and represents a reasonable trade-off between some degree of decay and a a non-trivial contribution of longer paths.[4]

For both measures the similarity indices for each triplet are normalized to sum to 1. This allows the model predictions to be directly comparable to the empirical choice probabilities, which also sum to 1.

Evaluating the similarity measures

In order to assess whether the semantic network based measures of similarity are capable of capturing the pattern of weak similarities we observed in our experiments, it is first necessary to construct a semantic network. In other words, we must determine the weighted adjacency matrix $\mathbf{A}$ from which our measures are derived. We constructed this network from a large dataset of word associations consisting of 12,571 cues and over 3 million responses. The data come from a task in which participants were given a short list of cue words and asked to generate three different responses to each single cue (see De Deyne & Storms, 2008b; De Deyne et al., 2011). From this data set we constructed two different weighted directed adjacency matrices. The graph $\mathbf{A}_1$ only counts the *first* response given by the participant, whereas $\mathbf{A}_3$ counts all three responses. The graph based on $\mathbf{A}_1$ is the more conventional approach, and its sparsity is comparable with previous word association studies (Nelson, McEvoy, & Schreiber, 2004). Because it is based on more responses $\mathbf{A}_3$ is somewhat denser, but in both cases the graphs were quite sparse. The graph $\mathbf{A}_1$ included 11,969 nodes and only 0.416% of the possible links, whereas $\mathbf{A}_3$ included 12,420 nodes and 1.176% of possible links.

To evaluate how well the weak similarities from our experiments can be captured using the semantic network models, we calculated the Spearman rank order correlations between the network-derived similarities and the empirical data. The results, summarized in Table 1, demonstrate that both measures of similarity are significantly correlated with the empirical data. As one might expect, the more global measure of similarity (the random walk measure) performs considerably better than the local cosine measure; and the richer network ($\mathbf{A}_3$) tended to produce higher correlations than the network based on less data. Taken together, the general finding is that the more data one uses to define the network, and the more that the similarity measure takes account of the structure in that network, the better one is able to capture human intuitions about weak semantic similarity.[5]

Table 1: Spearman rank order correlations (ρ) between the graph-derived similarities and the empirical similarities from both experiments. All correlations are significant at $p < .001$, indicated by the double stars. The more global measure of similarity (random walk) consistently outperforms the more local measure (cosine), and that the correlations are stronger for the denser network (i.e., $\mathbf{A}_3$).

	Cosine		Random walk	
Graph	Exp 1	Exp 2	Exp 1	Exp 2
$\mathbf{A}_1$	.19**	.22**	.48**	.49**
$\mathbf{A}_3$	.38**	.37**	.55**	.57**

For Experiment 2, we can extend the analysis to see if the network measures can account for decision latencies as well. In general, one would expect that more difficult pairs should result in longer decision latencies. For each pair, we calculated the absolute similarity of the strongest pair and compared it with the decision latency of that pair. Restricting our results to the random walk measure of similarity, we found a significant correlation between network-based similarities and decision latencies ($\rho = -.22$ for the $\mathbf{A}_1$ network, and $\rho = -.24$ for the $\mathbf{A}_3$ network, $p < .05$ in both cases). This is again consistent with the hypothesis that the semantic network encodes at least some information used to derive weak similarity.

Discussion

The work in this paper demonstrates that there is substantial agreement between people about the similarity structure of even weakly related items, like HAIL and TEACHER or RAINBOW and TUNAFISH. Moreover, at least some of this agreement can be accounted for by semantic networks constructed from word association data.

The most striking thing about this finding is that there is any agreement about weak similarity at all. In the abstract, there appears to be very little in common between any three items that are randomly thrown together, and it is not an obvious conclusion that people would agree on how they are related. In practice, many people have strong intuitions about any given triplet, just as two of the authors of this paper had strong intuitions about CUP, TEACHER, and HAIL. Two aspects of this are most intriguing. First, there isn't always agreement about these intuitions (just as one author thought TEACHER was the obvious odd one out, and one thought it should be HAIL). Second, as the data from our two experiments show, there is nevertheless substantial agreement (nobody thought CUP should be the odd one out).

The main question we are left with is *why* people should agree on something like this. There is almost certainly no external pressure in the environment to do so; it is difficult to think of any situations in which random unrelated things are thrown together or used, and people must agree with each other about them without communicating explicitly. Rather,

anism, namely the fact that the entire network is quickly activated (e.g. Ratcliff & McKoon, 1994).

[4]Other values of α were tried as well, but did not substantially change the pattern of results of our experiments

[5]Within the human data from Experiment 1, there are 28 triplets that did not share a single first association in our semantic network, and 72 that did. Because we were concerned that these results might simply be capturing this difference, we re-calculated the correlations separately for these two subsets of the data. The results did not differ in any substantive way from those reported in Table 1. Interestingly, 27 out of the 28 strongest pairs from these zero-overlap triplets were agreed upon by the human observers more than one would expect by chance. This amount of agreement was similar in Experiment 2, in which 25 of the strongest pairs from 28 triplets were agreed upon more than chance would predict.

such agreement probably stems from commonalities in the shared representations underlying the concepts. But what are those shared representations, and why should they exist at all? It is clear why it would be useful to represent similarities between entities that commonly co-occur, or that are often thought about together – but what benefit is there to building a representation that will probably never be used, and why do people seem to build similar ones?

Part of the answer to these questions may come from our analyses showing that semantic networks built from word associations can account for at least some of the agreement between people. This suggests that perhaps the shared representations measured in our weak similarity task don't occur because they offer some benefit, but rather occur as a by-product of the fact that the mind represents other things. In this case, it is interesting that networks formed from word associations capture some of those other things. We can be somewhat assured that the agreement accounted for by the networks is not the result of trivial or superficial similarities, since denser networks did better and things like frequency and imageability of the words was controlled for. Rather, it may be that these networks capture, at least to some extent, the kind of deep ontological similarities and abstract relationships that drove our intuitions about triples like CUP, TEACHER, and HAIL.

In light of this possibility, there are a number of areas that would be interesting to explore in future work. While our networks did account for a significant amount of the variance in people's weak similarity ratings, a substantial amount remains without explanation. One possibility for this is that our networks, despite being constructed from 12,000 associations, are still almost certainly much sparser and underspecified than people's actual semantic networks. Indeed, we found that the denser network constructed from more associations accounted for the data better. How much improvement is possible with increasingly dense networks and more items and associations? That is, to what extent is a large part of the variance in weak similarity ratings due to the same thing underlying the associations people make in word association tasks? How would this compare to networks constructed in other ways, like co-occurrence in language? How would this change if the networks were constructed in a more robust way, for instance, addressing the sparsity problem by inferring missing links, as in Miller, Griffiths, and Jordan (2009)? Is performance better or worse for different kinds of words, like abstract vs concrete? Work on all of these questions will help us to address the fundamental issue of what kind of semantic representation humans have – and how that representation underlies people's ability to estimate weak similarity.

Acknowledgments

This work was supported by a research grant funded by the Research Foundation - Flanders (FWO) to the first author and by the interdisciplinary research project IDO/07/002 awarded to Dirk Speelman, Dirk Geeraerts, and Gert Storms. Special thanks to Dinis Gökaydin and Steven Verheyen for helpful comments.

References

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database [CD-ROM]*. Philadelphia: University of Pennsylvania, Linguistic Data Consortium. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Bohn, R. E., & Short, J. E. (2009). *How much information? 2009. Report on American Consumers* (Tech. Rep.). Global Information Industry Center. University of California, San Diego.

De Deyne, S., & Storms, G. (2008a). Word Associations: Network and Semantic properties. *Behavior Research Methods, 40,* 213-231.

De Deyne, S., & Storms, G. (2008b). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods, 40,* 198-205.

De Deyne, S., Voorspoels, W., Verheyen, S., Navarro, D., & Storms, G. (2011). Graded structure in adjective categories. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (p. 1834-1839). Austin, TX: Cognitive Science Society.

Goodman, N. (1972). Problems and projects. In N. Goodman (Ed.), (p. 437-447). New York: Bobbs-Merrill.

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the Mind. *Psychological Science, 18,* 1069–1076.

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

Hutchinson, J. (1989). NETSCAL: A network scaling algorithm for nonsymmetric proximity data. *Psychometrika, 54,* 25-51.

Kemeny, J., & Snell, J. (1976). *Finite markov chains.* Springer-verlag.

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,* 211-240.

Leicht, E., Holme, P., & Newman, M. (2006). Vertex similarity in networks. *Psychical Review E, 73,* 026120.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100,* 254-278.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (p. 179-195). New York: Cambridge University Press.

Miller, K. T., Griffiths, T. L., & Jordan, M. I. (2009). Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems* (Vol. 22, p. 1276-1284).

Minkov, E. (2008). *Adaptive graph walk based similarity measures in entity-relation graphs*. Unpublished doctoral dissertation, School of Computer Science Carnegie Melon University, Pittsburgh, PA 15213.

Navarro, D. J., & Lee, M. D. (2002). Commonalities and distinctions in featural stimulus representations. In W. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (Vol. 24, p. 685-690). Mahwah, NJ: Lawrence Erlbaum.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, and Computers, 36,* 402–407.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The pagerank citation ranking: Bringing order to the web.* (Tech. Rep.). Computer Science Department, Stanford University.

Quillian, M. (1968). Semantic information processing. In M. Minsky (Ed.), (p. 227-270). Cambridge, MA: MIT Press.

Ratcliff, R., & McKoon, G. (1994). Retrieving information from memory: Spreading-activation theories versus compound-cue theories. *Psychological Review, 101,* 177-184.

Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behaviour Research Methods, Instruments, and Computers, 36,* 506–515.

Sloman, S. A., & Rips, L. J. (1998). Similarity as an explanatory construct. *Cognition, 65,* 87–101.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Experimental Cognitive Psychology and its Applications. In A. Healy (Ed.), (chap. Word association Spaces for Predicting Semantic Similarity Effects in Episodic Memory.). Washington, DC: American Psychological Association.

Van Dongen, S. (2000). *Graph clustering by flow simulation.* Unpublished doctoral dissertation, University of Utrecht.