All parts are not created equal: SIAM-LSA

Peter Wiemer-Hastings
peterwh@cti.depaul.edu
DePaul University
School of Computer Science,
Telecommunications, and Information Systems
243 S. Wabash
Chicago IL 60604

Abstract

Previous research has shown the inadequacy of models for computing similarity that rely on any type of simple combination of features. Human similarity judgments are sensitive to the structure of the items being compared. For visual stimuli, the spatial arrangement of the items provides an obvious structure. For textual stimuli, however, the structure of the items must be inferred. Prior research on textual similarity has shown the dominant effect of relational features. We extend that research by looking at human judgments of the similarity of sentence pairs within the framework set out by Goldstone's (1994) SIAM model, which calculates correspondences between objects and their features. We show that although the simple SIAM-based model fails to account well for the human judgments, a modified version which gives different weights to different semantic roles provides a strong match with human ratings.

Introduction

The ability to assess the similarity of objects in the world is fundamentally important to our survival. A variety of theories have been proposed for modeling human similarity judgments. Most of these theories involve comparing the sets of features of the compared items to determine the overlap between them. Many of the theories also ignore the structure of the objects and the relationships between the parts. Goldstone (Goldstone, 1994) showed that such systems fail to account for human similarity ratings of structured data. His SIAM system used a (non-learning) connectionist architecture to create correspondences between objects and their features in different scenes. Excitatory connections reinforced coherent mappings between objects (e.g. ObjectA to ObjectC and ObjectB to ObjectD). Inhibitory connections fought against redundant or contradictory mappings. Likewise, connections between the features of objects either supported or inhibited each other and the corresponding object–object connections. SIAM's connectionist architecture allowed it to take into account the structure of the scenes and the objects as well as the similarity of the features.

While we are interested in comparison of texts, Goldstone examined similarity ratings of visual scenes. His approach represented a scene as a spatially related set of objects (for example, pairs of schematic butterflies). Each object has a set of parts each of which has some value. For example, one of Goldstone's butterflies could be represented as:

```
(object1 (head square)
         (tail zig-zag)
         (body-shading white)
         (wing-shading checkered))
```

To map this approach to sentences, we broke the inputs into subject, verb, object, and indirect object parts. Thus, a simple representation of the sentence "The dog bit a man" as an object would be:

```
(object1 (verb "bit")
         (subject "The dog")
         (object "a man"))
```

Goldstone compared his approach to other models which used simpler combinations of features in the scenes, such as multidimensional scaling (Shepard, 1962) and Tversky's (1977) Contrast Model, and showed that SIAM accounted better for human judgments which were affected by the structure of the inputs. Bassok and Medin (1997) showed that in some cases, humans will make thematic inferences when rating the similarity of contextually related sentence pairs. While the stimuli that we use here are related, they are unlikely to be viewed as thematically connected.

The analyses describe here use Latent Semantic Analysis (LSA) as a basic technique for computing the similarity of texts. LSA was originally developed for the task of information retrieval (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), but has recently been shown to provide interesting correspondences with human language understanding on a variety of tasks (Landauer & Dumais, 1997; Foltz, 1996; Wiemer-Hastings, Wiemer-Hastings, & Graesser, 1999, for example). LSA often matches human similarity judgments well, but does not do well on single-sentence comparisons (Wiemer-Hastings & Zipitria, 2001). A likely culprit is its complete ignorance of syntax. This paper describes our application of Goldstone's SIAM model to attempt to account for human similarity judgments on pairs of sentences. We are interested in the extent to which similarity between textual components affects overall similarity judgments. More broadly, we want to explore the processes by which we derive sentence meanings compositionally from the meanings of the words.

Human sentence similarity ratings

In support of our research on Intelligent Tutoring Systems, we have developed LSA spaces for the domains of

Computer Literacy and Psychological Research Methods. Previous research has shown that human raters with expert domain knowledge achieved an inter-rater correlation of up to $r = 0.78$ when rating sentence similarities (Wiemer-Hastings et al., 1999) using real student answers in Computer Literacy. Raters with intermediate domain knowledge (the same level that many real-life tutors have) had correlations of up to $r = 0.52$. LSA cosine measures of sentence similarity correlated with the human judgments $r = 0.48$. Because the LSA representation ignores word order altogether, an obvious potential direction for improvement is to add structural knowledge to the technique. In the current experiment, we collected human ratings of sentence similarities where the structural and semantic overlap between the sentences was controlled. This section describes how these materials were created, the ratings that were collected, and what they suggest about the nature of similarity processes for texts.

Materials

To measure the structural overlap of the test items, we used Goldstone's (1994) approach which measures the Matches In Place (MIPs) and Matches Out of Place (MOPs) between the compared scenes. The basic idea is that when comparing scenes (or other complex, structured items), humans determine correspondences between objects and their component items. Their overall similarity judgments are affected more by feature similarities on corresponding components (MIPs) than on non-corresponding components (MOPs). Goldstone's research confirmed this hypothesis and showed how other models of similarity including multi-dimensional scaling (Shepard, 1962) and the Contrast Model (Tversky, 1977) failed to account for these effects.

Following Goldstone's approach, we generated pairs of items for comparison that had differing numbers of MIPs and MOPs. Goldstone's test items were schematic drawings with features that either matched or did not. When dealing with textual comparison items, we treat the main phrases within a sentence (subject, verb, object, and indirect object[1]) as features, and these can match *to some extent*, based on the similarity of those phrases. Thus, for a pair of sentences, MIPs measure the extent to which corresponding component phrases are similar. MOPs measure the extent to which non-corresponding component phrases (e.g. the subject of one sentence and the object of the other) are similar. We generated a test set of 50 sentence pairs (25 from each domain) using the following procedure.

Two sentence cores were created in both test domains. The sentence cores started with simple sentences broken down into subject, verb, object and (optionally) indirect object parts. These original features were augmented with additional noun phrases, verbs, and prepositional phrases that had similar or related meanings. For example, some of the candidate subjects in one template were: "the researcher," "the experimenter," and "a participant."

Next, 20 sentences were generated with each sentence core using this process:

1. Randomly select a subject, verb, object, and optionally an indirect object from the feature candidates. Also randomly choose if the sentence should be active or passive. If it is passive, randomly choose whether or not the subject should be dropped.

2. Set nMOPs (the number of MOPs) to a random number from 0 to 3. Set nMIPs to a random number between 0 and (4 - nMOPs).

3. Generate the second sentence as follows:

   (a) To make a MIP, choose any one of the values from the same slot in the same item.

   (b) To make a MOP, choose one of the values in the source slot, and put it in another slot in the second sentence (while maintaining syntactic coherence).

   (c) For the rest of the unfilled slots in the second sentence, choose some value from the corresponding slot of the *other item in that topic*. Presumably they will not match significantly, although they may to some extent.

4. Filter out the sentences that are obviously really uninterpretable (due to major violations of selectional constraints, for example).

5. Generate two booklets with 25 items chosen randomly from each topic. The order of the sentences was randomized, and the items were the same in the second booklet, but the order of the topics was switched and the order of the items within each topic was randomized.

One example sentence pair from Computer Literacy was:

Bits can be written.
Random access devices can be written.

These items are both passive and have their subjects dropped, and they have a MIP on the verb phrase.

Although some of the resulting sentences were quite sensible (e.g. "The CPU writes information to the peripherals."), it must be noted that others were a bit bizarre, e.g. "Causality retains causality," and "Correlation is caused." Some participants also reported difficulty with trying to compare the meanings of the sentences because they couldn't figure out what one or the other or both meant. Nonetheless, the similarity ratings showed some striking patterns, as described below.

Procedure

The participants were 14 Computer Science and Human-Computer Interaction graduate students studying Cognitive Science. The ratings task was given out during a

[1] We use these terms to describe the semantic roles within the sentence instead of more technical linguistic terms like Agent and Patient because we are interested in semantic information that can be derived from syntax alone.

break in class and was completely voluntary. The participants rated each of the 50 sentence pairs on a Likert scale from 1 (completely dissimilar) to 6 (completely similar). The participants were instructed to rate the similarity of the meanings of the sentence pairs, but were not given further instruction on how to determine the similarity.

Results

The average of the 14 participants' ratings were used as the gold standard for comparison of the other measures. There were no significant differences between the mean ratings for the two topics (3.171 and 3.170) or between the two booklets.

We calculated the Pearson's correlation between each participant's ratings and the target (average) rating. The range was $r = 0.08$ to $r = 0.678$, with a mean of $r = 0.463$. This is slightly below the level of our previous results where participants compared real-world sentences collected in a tutoring task. The inter-rater reliability here is also lower than for many similiarity-rating tasks. The task of comparing sentence meanings is a difficult one.

In Goldstone's SIAM model, the match between features was boolean — no partial credit was given. (The SIAM model includes a salience factor between each pair of features which was set by default to 1. More about this later.) To determine the effects of exact matches on human similarity judgments, we analyzed the similarity ratings for sentence pairs with exact MIPs[2], and performed a t-test to determine if there were differences. The similarity scores for sentence pairs with verb MIPs ($\bar{x} = 4.12$) was significantly higher than for subject and object MIPs ($t(100) = 2.71, p < 0.01, \bar{x} = 3.39$ and $t(28) = 2.52, p < 0.01, \bar{x} = 3.29$ respectively), but there was no significant difference between subject MIPs and object MIPs.

To look for interactive effects of exact MIPs and MOPs on similarity ratings, we did an ANOVA where the dependent variable was the average human rating, and the independent variables were the number of exact MIPs, the number of exact MOPs, and ActiveMatch, defined as 1 if both sentences were active or passive, and 0 otherwise. There was a main effect of exact MIPs $F(2, 10) = 13.12, p < 0.001, \eta^2 = .40$, but no other significant effects.

To account for partial matches between phrases, we used the LSA cosine metric.[3] For each sentence pair, we computed the sum of the partial MIPs (ΣMIPs), and the sum of the partial MOPs (ΣMOPs), and used a multiple regression analysis to determine the effects of these on the similarity ratings. ΣMIPs was a significant indicator of the similarity ratings $p < 0.005, \beta = 0.436$. ΣMOPs was not a significant predictor of the similarity ratings.

We also computed the individual effects of MIPs and MOPs for the different semantic slots, actor, object, direct object, and indirect object with a multiple regression analysis. (This is analogous to the feature correspondence aspect of SIAM, but in Goldstone's analyses, the feature values were randomly interchanged, so there was no differentiating effect of the features.) There were significant effects of Verb MIPs ($p < 0.005, \beta = 0.433$), Object MIPs ($p < 0.05, \beta = 0.347$), and indirect object MIPs ($p < 0.05, \beta = 0.324$). None of the other MIP measures or the MOPs had a significant effect. Table 1 shows the β weights and the significance level for all of the MIP and MOP variables.

Table 1: MIP and MOP β weights.

Model	β	**Sig.**
VMIP	0.433	0.002
OMIP	0.347	0.016
IOMIP	0.324	0.028
VMOP	0.212	0.116
SMOP	0.168	0.667
IOMOP	0.047	0.773
SMIP	-0.067	0.649
OMOP	-0.372	0.348

So far in this paper, we have been assuming that the *Place* in Matches In Place and Matches Out of Place is based on the semantic role of the phrase with respect to the head verb in the sentence. This assumption implies that the participants process the sentences enough to reach some sort of deep-level representation of the sentences which they use to make their similarity judgments. As mentioned above, the similarity judgments were not affected by the presence or lack of a match on the active/passive dimension.

It could still be the case, however, that the participants based their judgments on the surface syntactic features of the sentences which were spatially presented one over the other.[4] For example, the sentence pair, <"ROM stores data." "ROM is erased by the CPU"> could be rated as more similar because the syntactic subjects match, despite the fact that in the second sentence, "ROM" is the semantic object.

We calculated surface MIPs by measuring the LSA match values between the corresponding surface syntactic components of the sentences and did a regression analysis as before. The only surface MIPs which were significant indicators of the similarity rating were the verb and indirect object variables. These two are not affected by the passive – active transformation, and thus their surface role is the same as their deep role.

[2]Although the materials were generated with a specific manipulated target number of MIPs and MOPs, the random combination of items allowed for different numbers of MIPs and MOPs in the resulting sentence pairs.

[3]A number of other text similarity metrics are available, for example, Resnik's information theoretic measure which uses WordNet as its underlying ontological basis (Resnik & Diab, 2000). Analysis with other metrics is left for future research.

[4]This analysis is also consistent with two other hypotheses: the participants are affected by the order in which the words are presented, or by the order of the phrases.

Discussion

As mentioned above, the overall agreement between participants on the similarity ratings for these materials was not especially high, indicating that this is quite a difficult task. Nevertheless, clear effects of structure were evident in the ratings. MIPs were shown to be good predictors of the similarity ratings. In particular, verb MIPs had the strongest effect. This is consistent with the large body of prior research on the effects of structure on similarity ratings which shows that similarity of relational elements has a stronger influence than similarity of objects (Falkenhainer, Forbus, & Gentner, 1989; Medin, Goldstone, & Gentner, 1993; Forbus, Gentner, & Law, 1995; Bassok & Medin, 1997, for example).

In contrast with Goldstone's results, however, MOPs were not predictive of the overall similarity ratings. It is possible that this is primarily due to the difference in the stimulus set. Goldstone's schematic butterfly images imposed a spatial structure, but the features did not have functional significance. For example, the raters would probably not think that wavy-patterned wings would help the butterflies fly better or worse than solid pattern wings. When determining the similarity of texts, human raters apparently tend to ignore similarities between segments with different functional roles.

An alternative hypothesis (offered by a participant/student) is that MOPs were viewed by participants as *dis*similarities between the sentences. In other words, the participants may have been searching for positive or negative evidence about similarity, and interpreted a match on a different role as evidence that the sentences had different meanings.

Another significant implication of these results is the fact that similarity ratings are sensitive to the "deep structure" semantic roles as opposed to the order of phrases or the shallow syntactic structure of the sentences. This limits the power of bag-of-words techniques like LSA for comparing single-sentence items.

A more puzzling result is the lack of predictive power of subject MIPs. One possible interpretation is that we tend to focus on the predicative aspects of a sentence rather than the object that the predicate applies to. In previous research, we showed that averaging across subject, verb, and object similarities of sentence pairs led to better similarity judgments than by just averaging subject and predicate (verb and object combined) segments (Wiemer-Hastings & Zipitria, 2001). As we will explore below, there may be an alternative explanation for this difference.

SIAM-LSA

We originally re-implemented the SIAM model as a web-based demonstration for a class in Cognitive Science (written in Allegro Common Lisp and available by request). Our implementation follows the specifications of the SIAM model, and allows users to specify features of the butterfly images and see the results of running SIAM to compare them. This section describes how we extended that model to work with textual stimuli, and our evaluations of the model which we call SIAM-LSA.

Implementation

SIAM-LSA is a fairly trivial extension of SIAM. As mentioned above, the SIAM model allowed real-valued matches between features, but was primarily evaluated using binary match values. We extended our SIAM implementation to determine the match value between two text segments by simply calculating the LSA cosine between the segments. Everything else in the computation of the activations and overall similarity ratings was the same as for the image stimuli. The conceptually challenging task was to map the objects, features and relationships of the original stimuli to the textual stimuli.

Our first thought was to treat sentences as objects (analogous to Goldstone's butterflies), with subject, verb, object, and indirect object features, the values of which were the associated text segments. This is a straightforward mapping, but does not work well with the SIAM approach for two reasons. First, each scene has only one object, so there is no need for using the relations and determining correspondences. Second, SIAM never compares values of different features. For example, it would pay no attention to the fact that one butterfly's wings are checked and another's body is checked.

Another approach is to treat sentences as scenes. The verbs can be thought of as supplying relational information, so they could map to the spatial relations that were used in SIAM. The subject, object, and indirect objects were mapped to the SIAM objects, each of which had one feature that was the corresponding text segment.

The third approach was to treat the verb as an object as well, but to use relationships between the verb and the other phrases as the relational information in SIAM. Pilot testing on the model showed that this third approach performed best (although not especially well) in matching human similarity judgments, so that is the model that we evaluate here. The mappings are summarized in Table 2.

Table 2: SIAM to SIAM-LSA mappings.

SIAM	**SIAM-LSA**
Scene	Sentence
Object	Phrase
Feature/Dimension	Head (constant)
Value	Text of phrase
(spatial) relationship	Slot-filler roles, e.g. (subject-of <verb> <subj>)

Results

The overall similarity for SIAM-LSA was calculated using the same formula as the SIAM model: the normalized ratio between the sum of the products of the match values and the activations of the nodes over the sum of the activations of the nodes. For each of the sentence pairs, we created the SIAM network and ran the network for 20 cycles as in (Goldstone, 1994), recording the overall similarity result after each cycle. The Pearson correlation between SIAM-LSA and the average of the human

ratings was $r = 0.327$ after the first cycle (where there is no effect whatsoever of spreading activation), and decreased gradually but monotonically to $r = 0.273$ after 20 cycles. Thus, not only did SIAM-LSA fail to reliably predict the human similarity rating, the interactions between the nodes which allow it to establish correspondences actually made the similarity ratings worse with respect to human ratings.

SIAM-LSA-β

As described above, our analyses of the human similarity judgments showed that different features have different effects. Specifically, the similarity of the sentences' main verbs plays a large role, followed by that of their objects. Somewhat surprisingly, the similarities of the actors in the sentences played almost no role in this experiment.

The SIAM model allows for a Salience parameter between each two dimensions (or features). Ignoring the normally unused term which allows a non-normalized component, the formula for computing similarity in SIAM is:

$$similarity = \frac{\sum_{i=1}^{n} Matchvalue_i A_i S_i}{\sum_{i=1}^{n} A_i S_i}$$

where i ranges over the feature nodes and $Matchvalue_i$ is 1 (by default) if the corresponding features match and 0 otherwise. The A_i are the activation levels of the feature correspondence nodes. The activations are initially set to 0.5 and are modified each cycle as they receive excitation from coherent correspondence nodes and inhibition from conflicting correspondence nodes. The S_i parameters are defined as "$S_i = S_{ia} + S_{ib}, S_{ia}$ is the salience of Scene A's dimension i, and S_{ib} is the salience of the other scene's dimension i." (Goldstone, 1994, p. 15) By default, the S_i are all set to 1.

SIAM-LSA networks use only one feature, essentially a dummy feature whose value is the text segment of the semantic role. The functional semantic distinctions occur between the subject, verb, object and indirect object roles, as we used in our analyses of MIPs and MOPs in the human data. In the Siam-LSA-β model, we assigned values to the S_i values based on the β values in the multiple regression for MIPs and MOPs as shown in Table1. For matching semantic roles (e.g. verb–verb), we use the corresponding MIP β value. For non-matching roles, we took the average of the two corresponding MOP β values. For example, $S_{verb,object} = (\beta_{VMOP} + \beta_{OMOP})/2$.[5]

SIAM-LSA-β results

As above we ran the network for 20 cycles, recorded the overall similarity rating after each cycle, and compared it the human ratings. The Pearson correlation after cycle 1 was $r = 0.589$, increased slightly to $r = 0.591$ after 3 cycles, and then decreased to $r = 0.543$ after the 20^{th} cycle. This correlation is higher than any other LSA-based analysis on sentence pairs that we have seen, and is 30% higher than the average inter-rater correlation.

Discussion

Somewhat surprisingly, the SIAM-LSA model did not produce a strong correlation with the human similarity ratings on this task. As we will discuss in more detail below, this may have been due to the relative lack of structure in our textual stimuli. More surprising was the subsequent strong performance of the SIAM-LSA-β model. On second thought, however, this might have been expected, since the regression β weights give the linear contribution of each factor. On the other hand, Goldstone's analyses showed that simple combinations of feature matches could not account for structure-dependent aspects of human similarity judgments.

SIAM I am not

Because there was such a small difference between the correlations between the SIAM-LSA-β ratings throughout the different cycles, we created two models which use the SIAM similarity metric without the connectionist part of the SIAM model. The first, which we call the Non-interactive MIP and MOP Model (NM^3), is equivalent to the SIAM model on its first cycle. As mentioned above, the A_i start with the value of 0.5. Thus, they cancel each other out, and can be dropped from the formula. Using the default value of 1 for the S_i allows us to drop that from the formula as well, giving:

$$similarity = \frac{\sum_{i=1}^{n} Matchvalue_i}{\sum_{i=1}^{n} 1}$$

or simply: $similarity = average(Matchvalue_i)$. For the $Matchvalue_i$, we use the LSA cosine metric.

The Non-interactive MIP and MOP Model with β weights ($NM^3\beta$) calculates salience (S_i) values based on the β weights from the linear regression analysis of the human rating data using the same method as SIAM-LSA-β.[6]

We also computed similarities with three simple LSA-based models. The first (LSA) measured the cosine between the entire sentences. The second, SLSA, compared the corresponding phrases separately with LSA, and averaged the cosines as reported in (Wiemer-Hastings & Zipitria, 2001). In other words, the SLSA similarity value was the average of the subject–subject, verb–verb, object–object, and (if applicable) indirect object – indirect object LSA cosines for the two sentences. Finally, the weighted LSA model (WLSA) used the same approach as SLSA, but multiplied each component cosine by the corresponding MIP β weight above. The correlations between these 5 models and the human ratings are shown in Table 3.

[5] We also explored the transformation of the (non-learning) SIAM model into a hybrid connectionist model which would learn weights on connections from the feature nodes to a single output node which would give the similarity value. An analysis of this approach was left to future research.

[6] The use of the regression weights in the SIAM-LSA-β did not reek too strongly of circularity because the weights were fed into the network. Here, however, they are fed directly into a linear weighting formula. Additional testing is required to determine if these weights generalize to other texts.

Table 3: Non-interactive model correlations.

	Human	NM$^3\beta$	NM^3	LSA	SLSA
Human	1.000				
NM$^3\beta$	0.592	1.000			
NM^3	0.237	0.311	1.000		
LSA	0.109	0.055	0.807	1.000	
SLSA	0.445	0.692	0.467	0.209	1.000
WSLSA	0.408	0.624	0.422	0.179	0.825

Discussion

One moral of the modeling story here is that the interactive determination of the one-to-one correspondence mapping which distinguishes SIAM does not have much of an effect here. Although SIAM-LSA calculates the correspondences of MIPs and MOPs, it does not assign different weights to MIPs and MOPs on different features. When this is added to the model (albeit in an *ad hoc* manner), the model's ratings come much closer to those of the human raters.

These findings extend significantly beyond the simple SLSA model of averaging LSA matches between semantic roles. First, human similarity ratings are strongly affected by verb (relational) similarity and somewhat so by object and indirect object similarity. The similarity of the subjects has a non-significant effect. Second, as indicated by the WLSA results above, MOPs also have an effect. By including the MIP and MOP similarities, the SIAM-LSA-β and NM$^3\beta$ models match human judgments relatively well.

Conclusions

In this research, we have examined how matches between and within semantic roles affect human similarity judgments for textual stimuli. We have replicated previous research results that show the strong effect of relational similarity on overall similarity judgments. We have also found that different semantic roles affect these judgments to different extents. Using a computational model that gives differentiated sensitivity to structural matches and calculates text segment similarity using LSA, we were able to produce similarity judgments which correspond well with human ratings.

The simplicity of the structural analysis that we have used here will make it possible to use this technique for natural language understanding in constrained tasks like dialog-based intelligent tutoring systems. In such situations, the system has an expectation of what the student might answer to a given question. An off-the-shelf syntactic parser can segment a student's response, and then it can be compared to the expected answers using this technique. Because it is much easier to develop than a traditional natural language understanding mechanism, this technique can facilitate the delivery of such language-critical applications.

Acknowledgements

Many thanks to the students in CSC 587, Cognitive Science, in Winter 2004 at DePaul University who served (voluntarily) as the participants in the study and who were involved (later) in discussions of the results.

References

Bassok, M., & Medin, D. (1997). Birds of a feather flock together: Similarity judgments with semantically rich stimuli. *Journal of Memory and Language*, *36*, 311–336.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Falkenhainer, B., Forbus, K., & Gentner, D. (1989). The Structure-Mapping Engine. *Artificial Intelligence*, *41*, 1–63.

Foltz, P. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments, and Computers*, *28*, 197–202.

Forbus, K., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, *19*, 141–205.

Goldstone, R. (1994). Similarity, Interactive Activation, and Mapping. *Journal of Experimental Psychology*, *20*(1), 3–28.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

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

Resnik, P., & Diab, M. (2000). Measuring Verb Similarity. In *Proceedings of the* 22^{nd} *Annual Conference of the Cognitive Science Society* Mahwah, NJ. Erlbaum.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. Part I.. *Psychometrika*, *27*, 125–140.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (1999). How Latent is Latent Semantic Analysis?. In *Proceedings of the Sixteenth International Joint Congress on Artificial Intelligence*, pp. 932–937 San Francisco. Morgan Kaufmann.

Wiemer-Hastings, P., & Zipitria, I. (2001). Rules for Syntax, Vectors for Semantics. In *Proceedings of the* 23^{rd} *Annual Conference of the Cognitive Science Society* Mahwah, NJ. Erlbaum.