

Article

## The World Within Wikipedia: An Ecology of Mind

Andrew M. Olney <sup>1,\*</sup>, Rick Dale <sup>2</sup> and Sidney K. D'Mello <sup>3</sup>

<sup>1</sup> Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152, USA

<sup>2</sup> Cognitive and Information Sciences, University of California, Merced, CA 95343, USA;

E-Mail: rdale@ucmerced.edu

<sup>3</sup> Computer Science and Psychology, University of Notre Dame, Notre Dame, IN 46556, USA;

E-Mail: sdmello@nd.edu

\* Author to whom correspondence should be addressed; E-Mail: aolney@memphis.edu;

Tel.: +1-901-678-5008; Fax: +1-901-678-1336.

Received: 22 May 2012; in revised form: 11 June 2012 / Accepted: 12 June 2012 /

Published: 18 June 2012

---

**Abstract:** Human beings inherit an informational culture transmitted through spoken and written language. A growing body of empirical work supports the mutual influence between language and categorization, suggesting that our cognitive-linguistic environment both reflects and shapes our understanding. By implication, artifacts that manifest this cognitive-linguistic environment, such as Wikipedia, should represent language structure and conceptual categorization in a way consistent with human behavior. We use this intuition to guide the construction of a computational cognitive model, situated in Wikipedia, that generates semantic association judgments. Our unsupervised model combines information at the language structure and conceptual categorization levels to achieve state of the art correlation with human ratings on semantic association tasks including WordSimilarity-353, semantic feature production norms, word association, and false memory.

**Keywords:** association; comparison; semantic; feature; false memory; gist; activation; Wikipedia; W3C3

---

## 1. Introduction

Miller [1] offered the term *informavore* to capture our tendencies as cognitive agents to devour the information that we encounter in our environment. Miller's notion places emphasis on the agent as recipient of this information. Yet human beings are also actively constructing this information. In the past few millennia, the cultural institutions of human beings have produced a vast wealth of artifacts, and since the onset of written language, much of this creation has been linguistic in nature. Humans both devour, and construct, the informational culture around them, thus producing a dynamic that some have termed *niche construction* in biological systems [2]. In niche construction, the behavior of an organism transforms its environment in a manner that can then facilitate the survival of the organism itself, thus producing a feedback loop. Just as beavers' dam construction modifies its immediate ecology in which it lives, or particular species of trees can alter the nutrient content of a forest floor around them, human cognition and its external linguistic products affect each other through mutual influence: Over a short time scale, a single human extracts information from and adds information to this environment (e.g., linguistic input, reading, *etc.*), and over a longer time scale, the cumulative impact of the linguistic behavior of many humans produces change in that environment itself [3], creating an inherited linguistic and cognitive ecosystem.

To some cognitive scientists who focus solely on internal mechanisms, this may seem like a strange theoretical agenda. It may be useful to note, however, that there are a multitude of explanatory goals in the cognitive sciences, and these goals lie at different timescales [4]. If a cognitive scientist is interested in the immediate influences on language behavior—communicative goals, lexical knowledge, and so on—it may make sense to focus on cognitive theories that best explain the rapid deployment of such knowledge and processes. However, if a cognitive scientist is interested in understanding longer-timescale phenomena—such as cultural or linguistic change, or language evolution and origins—then a broader set of variables becomes relevant. Many theorists have argued that an understanding of longer-timescale biological phenomena is incomplete without attending to the ecological conditions in which these phenomena function (e.g., recently [5]). This notion of the cognitive agent and its environment suggests a mutuality that defies the conventional internalist/externalist dichotomy sometimes framed in the philosophy of mind and cognitive sciences. The linguistic environment and the cognitive agent are, from this perspective, parts of the same system, mutually constraining and shaping each other over a range of time scales [6].

Some effects of our linguistic environment are enormous and unfold gradually. For example, a child raised in an English speaking community will learn to speak English natively and will not spontaneously start speaking Chinese. However, many effects of our linguistic environment are quite subtle. It is well known that exposure to certain kinds of sentence structure will temporarily facilitate participants to produce language with the same structure, a phenomena called structural priming [7]. In addition, prior exposure to particular sentence structures can also cause grammaticality judgments of greater acceptability for new sentences with the same structure [8]. This effect persists at least seven days after exposure and increases when participants read for comprehension. These research efforts indicate that subtle sentence structure effects are long enough in duration to be an influential part of our cognitive-linguistic environment.

Perhaps the most striking evidence of a cognitive-linguistic ecosystem comes from developmental studies. The common experimental paradigm in these studies is to situate a child and adult in a play session with a new toy. The adult then produces a novel name for the toy multiple times in the session, and some time later the adult uses the novel name to ask the child to get the toy. Children at 13 months of age will correctly respond to a paired novel non-word as well as a paired novel word, but at 20 months children lose this ability and can only correctly respond when labels are novel words [9]. Similarly 20–26 month old children respond to words as labels but only when they are produced by the mouth, rather than by a tape recorder held by the adult [10]. During development, attention is increasingly focused on words and child-directed words as a cue to naming objects.

Related work in named category learning builds on these effects. In this paradigm, multiple objects/toys belonging to the same category are presented with a word label. When 17 month old children are presented with a label for two toys that are different in all respects except shape, not only do they correctly learn that the label corresponds to shape and generalize it to new objects, but when presented with a new label and new objects with a novel shape, children are able to correctly generalize that the new label refers to the novel shape in a single trial [11]. In addition, children who participated in the 8 week experiment showed a roughly 250% increase in object name vocabulary growth during this time compared to a control group that was exposed to the same objects without corresponding word labels. Only children exposed to categories and word labels were able to generalize the property of shape to new objects in a single trial. In a related study with 13 month olds, not only were word labels found to increase attention to novel objects of the same category, but word labels were also found to increase attention to the superordinate category (cow–animal), relative to a non-word-label condition [12]. These studies demonstrate the mutual influence between language and cognition during development: Word labeling focuses attention on category features, attention to discriminating features improves category structure, and improved category structure facilitates the learning of more word labels.

Although the growing body of empirical work above indicates that our cognitive-linguistic environment affects language structure and categorization, it also highlights the difficulty of long duration experiments with human participants. An alternative approach is to provide a comparable cognitive-linguistic environment to a computational cognitive model and observe the similarities between that model's behavior and human behavior. There is an extensive literature using this approach to model human semantic behavior. One popular approach, known as latent semantic analysis (LSA), represents text meaning as the spatial relationships between words in a vector space [13,14]. LSA has been used to model a variety of semantic effects including approximating vocabulary acquisition in children [13], cohesion detection [15], grading essays [16], understanding student contributions in tutorial dialogue [17,18], entailment detection [19], and dialogue segmentation [20], amongst many others. LSA is part of a larger family of distributional models. The underlying assumption of distributional models is that the context of use determines the meaning of a word [21]. Thus *doctor* and *nurse* would have a similar meaning, because these words (as well as their referents) typically occur in the same context. In the example of LSA, the contexts associated with a word are represented as vector components, such that the  $j$ th component of the word vector is the number of times that word appeared in the  $j$ th document in the text collection. Other distributional models vary according to how they define, represent, and learn contexts [22].

However, traditional models such as LSA are based solely in language structure, and so they do not model the mutual influence between cognition and language. This is partly because the available environments for such models have been entirely linguistic, e.g., text-dumps of books, newspapers, and other abundant sources of text. In contrast, the advance of the Internet has given rise to data sets that are created and organized in novel ways that reflect human conceptual/categorical organization. Wikipedia is the prototypical example of this new breed of cognitive-linguistic environment. It is read and edited daily by millions of users [23]. As an online encyclopedia, Wikipedia is structured around articles pertaining to concept-specific entries. Additionally, Wikipedia's structure is augmented by hyperlinks between articles and other kinds of pages such as category pages, which provide loose hierarchical structure, and disambiguation pages, which disambiguate entries with exact or highly similar names. Using Wikipedia as a cognitive-linguistic environment, a computational model that incorporates both the mutual influences of conceptual/categorical organization and the structure of language should produce behavior closer to human behavior than a model without such mutual influence.

Several researchers have already used Wikipedia's structure in models that emulate human semantic comparisons [24–26]. In this paper we extend their work in two significant ways. First, rather than focus on a single type of structure, e.g., link structure or concept structure, we present a model that utilizes three levels of structure: Word-word, word-concept, and concept-concept (W3C3) to more fully represent the cognitive-linguistic environment of Wikipedia. As we will show in the following sections, each of these levels independently contributes to an explanation of human semantic behavior. Secondly, in addition to the common dataset considered by previous researchers using Wikipedia, the WordSimilarity-353 [27] dataset, we apply the W3C3 model to a wider array of behavioral data, including word association norms [28], semantic feature production norms [29], and false memory formation [30]. Studies 1 to 4 examine how the W3C3 model manifests language structure and categorization effects across this wide array of behavioral data. Our analysis suggests that, at multiple levels of structure, Wikipedia reflects the aspects of meaning that drive semantic associations. More specifically, meaning is reflected in the structure of language, the organization of concepts/categories, and the linkages between them. Our results inform the internalist/externalist debate by showing just how much internal cognitive-linguistic structure used in these tasks is preserved externally in Wikipedia.

## 2. Semantic Models

In the following sections we present three approaches that when applied to Wikipedia extract models of semantic association at three different levels. The first model, the Correlated Occurrence Analogue to Lexical Semantics [31], operates at a word-word level. The second model, Explicit Semantic Analysis [24,25], operates at a word-concept level. The third and final model, Wikipedia Link Measure [26], operates at a concept-concept level. We then describe a joint model (W3C3) that trivially combines these three models.

### 2.1. Correlated Occurrence Analogue to Lexical Semantics

The Correlated Occurrence Analogue to Lexical Semantics (COALS) model implements a sliding window strategy to build a word by word matrix of normalized co-occurrences [31]. Because the

meaning representation of each word is defined by its co-occurrence with other *words*, COALS can be considered to be a word-word level model. The COALS matrix is constructed using the following procedure. For each word in the corpus, the four words preceding and following that word are considered as context. The word at the center of the window is identified with a respective row of a matrix, and each of the eight context words in the co-occurrence window are identified with a respective column. Thus for a particular window of nine words, eight matrix cells can be identified corresponding to the row of the center word and the columns of the eight context words. These eight cells are incremented using a ramped window, such that the immediate neighbors of the center word are assigned a value of 4, the next neighbors a value of 3, and so on such that the outermost context words are assigned a value of 1. Thus the eight cells are incremented according to a weighted co-occurrence, where the weight is determined by the distance of the context word from the center word. After the matrix is updated with all the context windows for the words in the corpus, the entire matrix is normalized using Pearson correlation [31]. However, since the correlation is calculating the joint occurrence of row and column words (a binary variable), this procedure is equivalent to calculating the phi coefficient, which we present as a simpler description of the normalization process. Let  $v$  be the value of a cell in the co-occurrence matrix,  $c$  be the column sum of the column containing  $v$ ,  $r$  be the row sum of the row containing  $v$ , and  $T$  be the sum of all cells in the matrix. Table 1 summarizes the entries for calculating the phi coefficient.

**Table 1.** Per cell calculation of the phi coefficient.

		Word B		Total
		Present	Absent	
Word A	Present	$v$	$r - v$	$r$
	Absent	$c - v$	$T - r - c + v$	$T - r$
	Total	$c$	$T - c$	$T$

The corresponding phi coefficient [32] is

$$\phi = \frac{Tv - cr}{\sqrt{c(T - c)r(T - r)}}$$

In addition to transforming each cell value into its corresponding phi value, COALS “sparsifies” the matrix by replacing all non-positive cells with zero, such that for any cell value  $v$

$$v = \begin{cases} \phi & \text{if } \phi > 0 \\ 0 & \text{otherwise} \end{cases}$$

Thus the final representation for a given word is its associated row vector, whose only non-zero components are positive correlations between that word and context words. The semantic similarity between two such words may be compared by locating their corresponding row vectors and calculating the correlation between them.

It is worth noting that the original COALS article proposes several variants based around the above process. One such variation removes 157 stop words before processing the corpus; another restricts

the columns of context words to the most frequent words only. Although the usefulness of these variations has been disputed in other frameworks [33], we nevertheless follow the original process for replication purposes. A third variant of COALS transforms the co-occurrence matrix using singular value decomposition (SVD) [34]. SVD is the key step in LSA and is used in COALS in a similar way: To eliminate noise in the matrix. In this variant, called COALS-SVD, the matrix  $A$  is first reduced to its most common 15,000 rows and 14,000 columns, forming the submatrix  $B$ . Phi-normalization as described above is applied to  $B$ , and finally  $B$  is transformed using SVD into three matrices

$$B = U\Sigma V^T$$

where  $U$  and  $V$  are orthonormal matrices and  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$  and  $\sigma_1 \geq \dots \geq \sigma_n \geq 0$ . The  $\sigma_i$  are the singular values of the matrix  $B$ . The desired matrix for word-word comparisons is  $U$ , whose row vectors are the SVD-denoised versions of  $B$ 's row vectors. Observe that right multiplying  $B$  by  $V\Sigma^{-1}$  yields  $U$

$$BV\Sigma^{-1} = U$$

By this identity, the full vocabulary of the original  $A$  matrix may be projected into the SVD solution for  $B$ , as long as the column dimensions of  $A$  and  $B$  match (e.g., 14,000). To do this simply right multiply  $A$  by  $V\Sigma^{-1}$

$$AV\Sigma^{-1} = U_A$$

$U_A$ 's row vectors are SVD-denoised versions of  $A$ 's row vectors, defined by the SVD solution of  $B$ .

## 2.2. Explicit Semantic Analysis

Explicit Semantic Analysis (ESA) uses the article structure of Wikipedia without considering the link structure [24,25]. The intuition behind ESA is that while traditional corpora are arranged in paragraphs, which might contain a mixture of latent topics, the topics in Wikipedia are explicit: Each article is a topic, or correspondingly a *concept*. ESA defines term vectors in terms of their occurrence in Wikipedia articles. Because the meaning representation of each word is defined by its co-occurrence with article *concepts*, ESA can be considered as a word-concept level model. ESA vectors are based on frequency counts weighted by a variation of term frequency-inverse document frequency (tf-idf) [35]:

$$tf_{ij} = \begin{cases} 1 + \log(v_{ij}) & \text{if } v > 0 \\ 0 & \text{otherwise} \end{cases}$$

where  $v_{ij}$  is the number of occurrences of a term  $i$  in an article  $j$ . Correspondingly:

$$idf_i = \log \frac{|A|}{|a_j : t_i \in a_j|}$$

where  $|A|$  is the total number of Wikipedia articles and the denominator is the number of articles in Wikipedia that contain a given term. An ESA vector for term  $i$  is defined by:

$$v_{ij} = \frac{tf_{ij} * idf_i}{\|tf_{ij} * idf_i\|}$$

where  $v_{ij}$  is the  $j$ th component of vector  $v_i$ , normalized by the length of the tf-idf vector. As a result, all ESA vectors have a length of 1. Similarity between terms is computed as the vector cosine between term vectors.

Because Wikipedia is a rather large corpus with many unique terms, ESA approaches have used pruning heuristics to reduce the space to a more manageable size [24,25]. Pruning heuristics include removing articles with less than 100 words, removing articles with fewer than five total links (inlinks and outlinks), removing high frequency words (stop words), removing rare words, and transforming the remaining words into their approximate uninflected root form (a process called stemming).

### 2.3. Wikipedia Link Measure

Previous work has used the link structure of Wikipedia to derive a semantic similarity measure using the Wikipedia Miner toolkit [26]. The basic premise of this approach is that every Wikipedia page has pages that link to it (*inlinks*) as well as pages it links to (*outlinks*). The links are inline with the text, meaning that a word or phrase in the text has been hyperlinked. The words/phrases themselves are called *anchors*, and they can be viewed as synonyms for the target page, e.g., *motor car*, *car*, and *automobile* link to the Wikipedia page *Automobile*. The Wikipedia Link Measure (WLM) uses anchors to map input text to Wikipedia articles and then uses the inlinks and outlinks of these articles to derive the similarity between words [26]. Because the meaning representation of each word is defined by the links between its associated concept and other concepts, WLM matches our definition of a concept-concept level model.

Consider two articles, *Football* and *Sport*. For a particular link type, say inlinks, and these two articles, we can place all other articles in Wikipedia into one of four categories as shown in Table 2. The frequencies in Table 2 are hypothetical, but they serve to illustrate a common structure. First, the number of links shared by two articles are likely to be relatively small compared to the number of links they possess individually. Secondly, the number of links that neither has is likely to be very large and relatively close to the total number of articles.

**Table 2.** Hypothetical inlink overlap for *Football* and *Sport*.

		Sport	
		Has Link	No Link
Football	Has Link	4	3450
	No Link	563	3 million

Intuitively, two articles that share inlinks and outlinks are likely to be similar; however, to the extent that some links may be common across many articles they should be weighted less. This intuition is captured in the WLM outlink metric, which weights each outlink  $o$  by  $\log(|A|/|O|)$ , the log of the number of total articles in Wikipedia  $|A|$  divided by the number of articles that link to that page  $|O|$ . Weighted outlink vectors are constructed based on the union of outlinks for two articles, and the outlink

similarity is the cosine between these two vectors. The inlink metric is modeled after Normalized Google Distance [36] and measures the extent to which the inlinks  $X$  of article  $x$  intersect the inlinks  $Y$  of article  $y$ . If the intersection is inclusive,  $X = Y$ , the metric is zero:

$$\text{inlink}(x, y) = \frac{\log(\max(|X|, |Y|)) - \log(|X \cap Y|)}{\log(|A|) - \log(\min(|X|, |Y|))}$$

Inlink and outlink metrics are averaged to produce a composite score. Since each anchor defines a set of possible articles, the computations above produce a list of scored pairs of articles for a given pair of anchors. For example, the anchor *bill* links to bill-law and bill-beak, and the anchor *board* links to board-directors and board-game, leading to four similarity scores for each possible pairing. WLM selects a particular pair by applying the following heuristics. First, only articles that receive at least 1% of the anchor's links are considered. Secondly, WLM accumulates the most related pairs (within 40% of the maximum related pair) and selects from this list the most related pair. It's not clear from the discussion in Milne & Witten [26] whether this efficiency heuristic differs from simply selecting the most probable pair except in total search time.

#### 2.4. W3C3: Combined Model

In this section we present our combined model using implementations of the models described above. We call this model W3C3 because it combines information at the word-word, word-concept, and concept-concept levels. For each model except COALS, reference implementations were chosen that are freely available on the web.

To implement Wikipedia Miner's WLM, we downloaded version 1.1 from Sourceforge [37] and an xml dump of Wikipedia from October 2010 [38]. ESA does not have a reference implementation provided by its creators. However, Gabrilovich recommends another implementation with specific settings to reproduce his results [39]. Following these instructions, we installed a specific build of Wikiprep-ESA [40] and used a preprocessed Wikipedia dump made available by Gabrilovich [41]. We created our own implementation of COALS and created a COALS-SVD-500 matrix using the same xml dump of Wikipedia from October 2010 as was used for WLM above.

One intuition that motivates combining all three techniques into a single model is that each represents a different kind of meaning at a different level: Word-word, word-concept, and concept-concept. This intuition was the basis for our simplistic unsupervised W3C3 model, which is simply to average the relatedness scores given by these three techniques. Two relevant properties of the W3C3 model are worth noting. First, it has not been trained on any part of the data. Secondly, it has no parameters for combining the three constituent models; rather their three outputs are simply averaged to yield a single output score.

### 3. Study 1: WordSimilarity-353

The WordSimilarity-353 [27,42] collection is a standard dataset widely used in semantic relatedness research [24–26,43–47]. It was developed as a means of assessing similarity metrics by comparing their output to human ratings. WordSimilarity-353 contains 353 pairs of nouns and their corresponding judgments of semantic association. The nouns range in frequency from low (*Arafat*) to high (*love*)

and from concrete (*car*) to abstract (*psychology*). Judgments are made on a scale of 0 to 10, with 0 representing no relatedness and 10 representing maximal relatedness. The data is divided into two sets. The first set contains ratings by thirteen judges on 153 word pairs. The second set consists of ratings by sixteen judges on 200 word pairs. We assessed inter-rater reliability for both sets using Cronbach's  $\alpha$  and found a high level of agreement,  $\alpha = 0.97$ . The following analyses present results on all 353 pairs using the average rating across judges on each pair. Previous work on this dataset has reported results in terms of a non-parametric Spearman correlation; this metric of performance is also adopted here.

COALS has been previously shown to exhibit impressive performance on semantic tasks [31,48], including the WordSimilarity-353 task. The previous best result of COALS on the WordSimilarity-353 task,  $r(351) = 0.67$ , was achieved using COALS-SVD with 500 singular values on a 1.2 billion word Usenet corpus [31]. Using Wikipedia, our implementation yielded a stronger correlation with WordSimilarity-353 than previously reported,  $r(351) = 0.72$ ,  $p < 0.001$ , but this difference is not significant,  $z = -1.28$ ,  $p = 0.10$ .

The correlation between ESA and the WordSimilarity-353 data set is the highest previously reported,  $r(351) = 0.75$ . It should be noted that an ESA model using additional link information has been attempted, but yielded no improvements over this basic model [25]. The ESA implementation we used also correlated with the human ratings,  $r(351) = 0.67$ ,  $p < 0.001$ , which is significantly lower than the original reported correlation,  $p = 0.02$ . A plausible explanation of this difference is that some tweaks crucial for high performance, such as inverted index pruning [25], which are not implemented in the reference implementation we used, are necessary to achieve the originally reported correlation.

Milne & Witten [26] found that WLM was highly correlated with human judgments in WordSimilarity-353,  $r(351) = 0.69$ . The WLM reference implementation we used also correlated with human ratings in the data set,  $r(351) = 0.66$ ,  $p < 0.001$ , but was lower than the original reported correlation  $r(351) = 0.69$ . The difference in correlations was not significant,  $z = -0.73$ ,  $p = 0.23$ . The reason for this discrepancy is unclear, but it may be attributed to the differences in versions of Wikipedia used here and in the initial reported research.

The W3C3 model has state of the art correlation with the WordSimilarity-353 data set,  $r(351) = 0.78$ ,  $p < 0.001$ . Correlations for all models are presented in Table 3. The W3C3 model's correlation is significantly higher than all correlations in the replicated results,  $p \leq 0.03$ , but not significantly higher than the best previously published ESA result,  $p = 0.17$ .

**Table 3.** Current and previous correlations with WordSimilarity-353.

Model	Current	Previous
W3C3	0.78	
COALS	0.72	0.67
ESA	0.67	0.75
WLM	0.66	0.69

Previous work has found that distributional models and graphical (WordNet-based) models have differing performance on WordSimilarity-353 depending on whether the word pairs in question have

a similarity relationship or a more general relatedness relationship [43]. To test this hypothesis with the COALS, ESA, WLM, and W3C3 models, we used the same partitioning of the dataset into similarity and relatedness pairs. The similar pairs are synonyms, antonyms, identical, or hyponym-hyperonym, and the related pairs are meronym-holonym or other relations. Inter-rater agreement in the coding of the pairs was high, Cohen's kappa = 0.77. The similarity and relatedness subsets contained the similar and related pairs described above and shared the same unrelated pairs, yielding 203 pairs for similarity and 252 pairs for relatedness. Correlations for all models on these subsets are presented in Table 4. The difference in correlations between the W3C3 model and the previous best Agirre model [43] is significant for both similarity,  $p = 0.0465$  and relatedness,  $p = 0.02$ .

**Table 4.** Correlations with WordSimilarity-353 similarity and relatedness subsets.

<b>Model</b>	<b>Similarity</b>	<b>Relatedness</b>
Agirre <i>et al.</i> 2009	0.77	0.62
W3C3	0.83	0.72
COALS	0.78	0.64
ESA	0.72	0.64
WLM	0.70	0.58

For both similarity and relatedness subsets, the W3C3 model performed significantly better than its constituent models, and each model performed significantly better on the similarity set than on the relatedness set,  $p < 0.05$ , except for ESA,  $p = 0.06$ . However, these are rather coarse sets: As mentioned above, the similarity set is an aggregation of common semantic relationships. In order to better understand the relative performance of each model on these subtypes, we used the labeled semantic categories for each WordSimilarity-353 pair provided in the similarity/relatedness subsets. These are antonym, hypernym (first word is hypernym of second word or vice versa), identical, part-of (first word is a part of the second or vice versa), siblings (share a parent category, e.g., *dalmatian* and *collie* are children of *dog*), synonyms, and topical (some relationship other than previous relationships, e.g., *ladder* and *lightbulb*). Grouping pairs by semantic category, we calculated the average distance between predicted rank and the human rank for each model. The results are shown in Table 5. Since the ideal distance to the human ranking is zero, lower scores are better. The lowest score in each row is in boldface, and the second lowest score is italicized.

The most striking pattern in Table 5 is that two-thirds of the best scores per category belong to the W3C3 model. Moreover, for every category save one, the W3C3 model either has the best score or the second best score. Thus breaking down the WordSimilarity-353 pairs by semantic categories is producing the same pattern of results seen in Tables 3 and 4: The three constituent models are providing different kinds of information, and averaging their outputs is creating a more human-like measure of semantic comparison than any of them individually. A linear regression was conducted to explore this possibility. The scores given by COALS, ESA, WLM, and human judges were converted to ranks, and then a linear regression on the ranks was performed [49], using COALS, ESA, and WLM ranks to predict the human judgment ranks. The results of the linear regression are presented in Table 6. Tolerance

analyses were conducted to test for multicollinearity of COALS, ESA, and WLM by regressing each on the other two. The obtained tolerances, all between 0.49 and 0.60, suggest that the three models are not collinear. The explanation that each model is contributing substantially and equitably to the prediction is further supported by the similar magnitudes of  $\beta$  in Table 6.

**Table 5.** Average distance by WordSimilarity-353 semantic category.

Relation Types	COALS Distance	ESA Distance	WLM Distance	W3C3 Distance
Antonym	158	80	104	<b>75</b>
First hypernym	40	<b>27</b>	40	36
Second hypernym	53	62	56	<b>48</b>
Identical	<b>0</b>	<b>0</b>	8	<b>0</b>
Second is part	57	<b>56</b>	79	65
First is part	56	76	<b>45</b>	53
Siblings	62	<b>55</b>	59	<b>55</b>
Synonyms	43	92	55	<b>29</b>
Topical	64	65	67	<b>53</b>
All	60	64	64	<b>51</b>

**Table 6.** Regression on ranks of COALS, ESA, and WLM, for human judgment ranks (N = 353).

Feature	B	SE(B)	$\beta$
COALS	0.358	0.046	0.358 *
ESA	0.270	0.045	0.269 *
WLM	0.304	0.042	0.303 *

Notes:  $R = 0.80$ ,  $*p < 0.0001$ .

To address the question of the maximum potential of the COALS, ESA, WLM, and W3C3 models for correlation with human ratings, an oracle analysis was undertaken [43]. The oracle first converts the output of each model to ranks. Then for each word pair, the oracle selects the output of the model whose rank most closely matches the rank of the human rating. This procedure generates the best possible correlation with the human ratings, based on the assumption that the oracle will choose the closest model output every time. Using this methodology with all four models, the oracle correlation is  $r(351) = 0.93$ . Using only the three constituent models, the oracle correlation is  $r(351) = 0.92$ , which is equivalent to the previous best reported oracle correlation that used roughly an order of magnitude more data than the present study [43]. So the maximum potential correlation of the three constituent models matches the previous best result, with a minor improvement due to including the W3C3 model in the oracle.

The preceding analyses provide fairly strong evidence for reason behind the W3C3 model's efficacy. The W3C3 model has significantly higher correlations than the constituent models on the entire dataset,

this improvement is preserved across most semantic categories, the regression shows equal contribution by the constituent models, and the oracle shows the upside potential of these models is consistent with or perhaps slightly better than the previous best. All of these results support the conclusion that each constituent model's semantic level, *i.e.*, word-word, word-concept, and concept-concept, contributes positively to increasing the correlation with human semantic comparisons.

Our final analysis tests whether an artifact of the similarity scores might be responsible for this difference. It has been previously noted that models can perform better on WordSimilarity-353 when word pairs that lack a semantic representation are removed from the model [43]. Because of missing representations, these defective word pairs always have a zero similarity score (the default score). By averaging the three constituent model scores, the W3C3 model removes this deficiency: At least one of the models is likely to have a representation or otherwise produce a non-zero score. For example, due to missing or extremely sparse semantic representations for WordSimilarity-353 words, WLM yielded 73 zero relatedness scores, ESA yielded 81, and COALS yielded 0. Thus one explanation for the improved correlation of the W3C3 model over the individual models is that the W3C3 model minimizes the effect of missing/sparse word representations.

To explore the effect of zero relatedness scores on the performance of the individual and W3C3 models, we created a subset of WordSimilarity-353 for which none of the three models had a zero relatedness score, consisting of 226 word pairs. For the three individual models, higher correlations on this subset than on the whole set would support the missing/sparse representation explanation. Similarly, for the W3C3 model, a similar correlation to the other three models (with zeros removed) would also support the missing/sparse representation explanation. Correlations for all models on this subset are presented in Table 7.

**Table 7.** Correlations with WordSimilarity-353 without missing word representations.

<b>Model</b>	<b>Correlation</b>
W3C3	0.72
COALS	0.64
ESA	0.56
WLM	0.60

The pattern of correlations in Table 7 do not support the missing/sparse representation explanation. First, each model's correlation is lower than its counterpart on the whole data set given in Table 3, indicating that eliminating pairs with zero scores does not improve the performance of the individual models. Secondly, the W3C3 model in Table 7 has a higher correlation than any of the individual models by 0.08 or more, which is similar to the pattern on the whole dataset in Table 3, though all correlations were lower on this subset of data. Thus, the W3C3 model yields an improvement in correlation regardless of whether the words with missing/sparse representations are removed.

#### 4. Study 2: Semantic Feature Production Norms

Although the WordSimilarity-353 data set has been widely used, it does have one notable weakness as a measure of human semantic behavior, namely its size. Given its limited size, it is possible that the results from Study 1 might not generalize. In order to assess the ability of the W3C3 model to generalize to new data sets, we applied the constituent models and W3C3 model from Study 1 to a large set of semantic feature production norms [29]. For the sake of exposition we will refer to these norms using the first initial of the last name of the authors presenting the norms, MCSM.

The MCSM norms consist of 541 nouns, or concepts. Participants were asked to list features for a concept such as the physical, functional, or encyclopedic aspects of the concept. The generated features were regularized, e.g., *usually has a tail* and *tail* would be coded as *has tail*. After regularization, 2,526 features remained. Thus the data can be represented as a  $541 \times 2,526$  matrix whose cell values  $v_{ij}$  are the number of production occurrences for a feature  $j$  given a concept  $i$ . From this matrix, a  $541 \times 541$  matrix was created in which the value at each cell is the cosine of two 2,526 dimension row vectors associated with their respective concepts. The  $541 \times 541$  matrix has 95,437 non-zero pairs representing similarities between concepts. Although the collection methodology for the feature norms is an associative production task, this  $541 \times 541$  matrix represents the feature overlap between concepts, which is more comparative in nature.

Table 8 presents the Pearson correlations between the similarities from the  $541 \times 541$  similarity matrix and both the predictions from the constituent models and the W3C3 model. The overall pattern of correlations in Table 8 has a striking similarity to those of Table 3. First, the correlations of ESA and WLM are almost identical in both cases. Secondly, the correlation for COALS is significantly greater than that of both ESA and WLM. And finally, the correlation of the W3C3 model is significantly greater than all of the constituent models. That the pattern of correlations is the same in both cases, especially when the MCSM set is so large, suggests that the properties of the models observed in Study 1 are generalizing in a systematic way.

**Table 8.** Pearson correlations with MCSM (N = 95,437).

Model	Correlation
W3C3	0.67
COALS	0.63
ESA	0.52
WLM	0.53

In order to assess the relative contributions of each constituent model to MCSM performance, a linear regression was conducted. The regression used COALS, ESA, and WLM raw scores to predict the MCSM similarity scores. The results of the linear regression are presented in Table 9. Tolerance analyses were conducted to test for multicollinearity of COALS, ESA, and WLM by regressing each on the other two. The obtained tolerances, all between 0.55 and 0.68, suggest that the three models are not collinear. In contrast to the WordSimilarity-353 regression presented in Table 6, the constituent models

are not equally weighted. The magnitudes of  $\beta$  in Table 9 show that COALS, ESA, and WLM may be rank ordered in terms of their contribution to the overall model. However, the difference between the correlation produced by the W3C3 model in Table 8 and the correlation from the regression equation in Table 9 is extremely small (0.01), suggesting that equal weighting of the three constituent models is fairly robust.

**Table 9.** Regression of COALS, ESA, and WLM raw scores on MCSM similarity scores (N = 95,437).

Feature	B	SE(B)	$\beta$
COALS	0.362	0.003	0.417 *
ESA	0.415	0.005	0.231 *
WLM	0.146	0.003	0.151 *

Notes:  $R = 0.68$ ,  $*p < 0.0001$ .

### 5. Study 3: Word Association Norms

Word association norms represent a qualitatively different type of task than WordSimilarity-353 or MCSM. Typically in word association tasks, a human participant is presented with a word and asked to produce the first word that comes to mind. This production task is quite different from the raw data of MCSM, where subjects have the task constraints of listing physical, functional, or encyclopedic features and are asked to list 10 such features for a given concept. Transforming the MCSM data into a similarity matrix in Study 2 further removes the data from a stimulus-response production task and more squarely situates it with a semantic comparison task.

The relationship between semantic and associative relations has been the subject of recent discussion in the literature [50–54]. In particular, some have argued that framing word association and semantic relatedness as separate and distinct is a false dichotomy [53], whereas others have argued that word association and semantic feature overlap measure different kinds of information [51]. Study 1 examined semantic relatedness quite generally; Study 2 examined similarity based on semantic feature overlap. The present study examines word association as a stimulus-response production task without the task constraints of explicitly comparing two concepts.

Perhaps the most widely known word association norms have been collected by Nelson and colleagues over several decades [28]. The data used in the present study consists of 5019 stimulus words and their associated 72,176 responses. Each stimulus response pair is annotated by the proportion of participants who produced it, which we refer to as forward associative strength. We refer to these 72,176 triples (stimulus word/response word/forward associative strength) as NMS after the last names of its authors.

Previous work using a type of distributional model called a topic model (also known as latent Dirichlet allocation) and the TASA corpus [13] to train it, used the NMS dataset on two tasks [55]. The first task examined the central tendency of the model predictions via the median rank of the first five predicted responses. Thus this task first ranks the human responses for a stimulus word by their associated production probabilities and then compares these to the model's predicted ranking. If the first five

predicted responses for all stimulus words are some ordering of the ranks 1–5, then the median rank will be 3. The second task is simply the probability that the model’s first response matches the human first response. The results from this previous work as well as the results from our three constituent and W3C3 models on this task are presented in Table 10.

**Table 10.** Median rank of each model’s first five associates compared to human associates and proportion of model first associates that are the human first associate (N = 5,019).

Model	Median Rank	First Associate
LSA *	31	0.12
Topics Model *	18	0.16
W3C3	5	0.24
COALS	6	0.22
ESA	6	0.19
WLM	6	0.15

Notes: \* from Griffiths *et al.* [55].

As in Studies 1 and 2, the W3C3 model has higher agreement with the human data than both its three constituent models or previous models. It should be noted that the previous results reported in Table 10 are based on models that used a corpus that is two orders of magnitude smaller than Wikipedia and also were tested against only about 90% of the NMS forward associative strength data. Thus the present study used more data but was also tested against 100% of the NMS forward strength data.

The results from Table 10 consider the NMS dataset as a collection of lists: Each stimulus word matched to list of response words ranked by forward associative strength. However, it is also informative to consider each triple (stimulus word/response word/forward associative strength) individually as was done in Study 2. Accordingly, Table 11 presents the correlations of model scores for each pair with the associated forward associative strength.

**Table 11.** Pearson correlations with NMS (N = 72,176).

Model	Correlation
W3C3	0.28
COALS	0.26
ESA	0.15
WLM	0.20

The W3C3 model has a significantly higher correlation with the human data than its three constituent models,  $p < 0.001$ . However, these correlations are less than half as high for the NMS dataset as they were for the MCSM dataset presented in Table 8. This difference might imply that the underlying assumptions of these models may not be well aligned with the constraints of the word association task.

These findings are consistent with previous work that suggests measures of word association (e.g., NMS), semantic feature overlap (e.g., MCSM), and text-based distributional similarity are in fact measuring something different in each case. Maki and Buchanan [51] conducted a factor analysis in which these three types of measure loaded onto separate factors that were coherently either associative, semantic, or distributional in nature. One interpretation of these findings is that the observed separation is due to three separate cognitive representations aligned with these three measures. Alternatively, it could be the case that task-adaptive processing is acting on the *same* representation yet manifesting three different measures (*cf.* [52,53]). In either case, the work of Maki and Buchanan [51] suggests that modeling both semantic relatedness and word association data with a single representation and procedure is unlikely to be successful. The low correlations in Table 11 lend additional evidence to this claim.

As in the previous study, a linear regression was conducted to assess the relative contributions of each constituent model to NMS performance. The regression used COALS, ESA, and WLM raw scores to predict the NMS forward associative strength. The results of the linear regression are presented in Table 12. Tolerance analyses were conducted to test for multicollinearity of COALS, ESA, and WLM by regressing each on the other two. The obtained tolerances, all between 0.78 and 0.82, suggest that the three models are not collinear. As was previously found in Study 2, the constituent models are not equally weighted. The magnitudes of  $\beta$  in Table 12 show that COALS, WLM, and ESA may be rank ordered in terms of their contribution to the overall model. Thus compared to Table 9, the relative order of WLM and ESA is reversed. Interestingly, the correlation produced by the W3C3 model and the correlation from the regression equation in Table 12 are identical, again supporting the robustness of equally weighting the three constituent models.

**Table 12.** Regression of COALS, ESA, and WLM raw scores on NMS similarity scores (N = 72,176).

Feature	B	SE(B)	$\beta$
COALS	0.105	0.002	0.196 *
ESA	0.066	0.006	0.040 *
WLM	0.047	0.002	0.117 *

Notes:  $R = 0.28$ ,  $*p < 0.0001$ .

## 6. Study 4: False Memory

Perhaps the most striking evidence of semantic relatedness' influence on cognitive processing can be found in the Deese–Roediger–McDermott (DRM) paradigm [56]. In this paradigm, participants are presented with a list of words highly associated with a target word in previous word association norms experiments. For example, a list containing *bed*, *rest* and *dream* will likely lead to false recall of *sleep*. Participants in the DRM paradigm are highly likely to recall the non-presented target word—in some cases even when they are warned about such false memory illusions [57]. These effects have lead Gallo *et al.* [57] to conclude that the influence of semantic relatedness on retrieval is intrinsic and beyond the participant's conscious control. Because word association norms are asymmetric,

e.g., *bed* may evoke *sleep* with high probability but not the reverse, Roediger *et al.* [30] conducted a multiple regression analysis to determine whether forward association strength (target word evoking a list member), backward association strength (list member evoking a target word), or other features were most predictive of false memory. That study found that backward association strength was strongly correlated with false recall,  $r(53) = 0.73$ . It is generally believed that properties of the word lists themselves are only part of the explanation: The major theories of false memory, activation/monitoring [30] and fuzzy trace theory [50], both allow for cognitive processes of monitoring or strategies that intervene in the process of rejecting false memories.

Several researchers have proposed computational models to implement gist-type semantic representations that are consistent, but not tightly integrated with, fuzzy-trace theory [13,55]. In general, any distributional model, by pooling over many contexts, creates a gist-type representation. LSA has been proposed to create a gist-type semantic representation [13]. The intuition is that LSA abstracts meaning across many different contexts and averages them together. In LSA, although a single word may have multiple senses, e.g., *river bank* and *bank deposit*, LSA has one vector representation for each word which is pooled across all the documents in which that word occurs. More recently, latent Dirichlet allocation (LDA, also known as a topic model) has been proposed as an alternative to LSA for representing gist [55]. One notable advantage of LDA is that the conditional probabilities used to compare two words are inherently asymmetric and therefore consistent with the asymmetries in human similarity judgments [58]. Another advantage of LDA is that words have probabilistic representations over latent variables corresponding roughly to word senses. Thus the two senses for *bank* above could be preserved and represented in probability distributions over two distinct latent variables.

In this study we investigated the relationship between the W3C3 model and constituent models on backward associative strength in the DRM paradigm. Following previous research, we chose to focus on backward associative strength because it is highly correlated with false recall and may be considered independently of the monitoring processes that mediate false recall. In order to investigate the W3C3 model in this context, several of the constituent models had to be amended to produce gist representations for DRM lists. To create COALS gist vectors, the raw unnormalized vectors for each word were summed, then normalized using correlation, and then projected into a 500 dimensional SVD solution as described in Section 2.1. This operation ensured that each element of the gist vector was a correlation, just as is the case in a normal COALS vector. ESA naturally creates gist vectors from multi-word strings, so no additional algorithm needed to be developed. For WLM, we create a synthetic article using inlinks/outlinks of the most likely sense for each word. The most likely sense for each word was determined by considering only the previous word in the DRM list. The most likely sense is the sense that has the highest similarity to the previous word. For example, if the previous word is *baseball* and the current word is *bat*, then the *club* sense is more similar than the *flying mammal* sense according to WLM. Then the inlinks/outlinks for these most likely senses were aggregated and used to create a synthetic gist article with the union of inlinks and union of outlinks. This gist article was then compared to the non-present target article in the standard way. As before, the W3C3 model was an unweighted average of these three scores.

We used the above methods for calculating gist and applied it to the standard set of 55 DRM lists [30]. For each list, we computed the gist representation and then compared it to the representation for the

non-presented target word, yielding a similarity score. Table 13 presents the correlation between the similarity scores for each model and backward associative strength, including correlations previously reported for the LSA and LDA models described above [55].

Although none of the comparisons in Table 13 are statistically significant, the W3C3 model has a higher correlation with the human data as reflected by backward associative strength than the constituent models. Since previous results with LDA or LSA used a different corpus (TASA), a direct comparison is not warranted. Nevertheless, the correlation of the W3C3 model is not as strong as has been previously reported for LDA. This result was surprising, particularly with regard to the low correlation for WLM. We undertook a qualitative analysis to determine if there is a better correspondence between Wikipedia’s link structure and the associative strength behind the DRM paradigm than is reflected by the WLM metric.

**Table 13.** Spearman rank correlations with backward associative strength for DRM lists (N = 55).

Model	Correlation
LSA *	0.30
LDA *	0.44
W3C3	0.34
COALS	0.27
ESA	0.30
WLM	0.24

Notes: \* From Griffiths *et al.* [55]; LSA and LDA results include only 52 of 55 lists.

Table 14 provides some suggestion that the raw link structure of Wikipedia might be more strongly related to backward associative strength than the gist-like WLM metric reveals. Each word in Table 14 is from the DRM list for *sleep* [30]. As shown in the table, most words (11/15) have *sleep* as an outlink or are used equivalently to mean *sleep*. In other words, this pattern of links is consistent with the backward association strength found in [30].

**Table 14.** DRM list for target *sleep* (outlink ○, inlink ●, redirect/anchor ☆).

bed ○ ●	wake ○	snore ○
rest ☆	snooze ☆	nap ☆
awake ○	blanket ○ ●	peace
tired	doze	yawn
dream ○	slumber ☆	drowsy ○

In order to more rigorously assess the possibility that raw Wikipedia link structure might better reflect backward associative strength, we recomputed the correlations from Table 13 with separate measures for Wikipedia inlinks and outlinks. Recall from Section 2.3 that WLM has two separate measures for inlinks

and outlinks, which are averaged together to compute the WLM metric. Table 15 presents correlations of these inlink/outlink measures separately, with the standard WLM measure, and with the corresponding W3C3 models.

By treating inlinks and outlinks separately, the correlation to backward associative strength increases markedly. What is perhaps most interesting about the pattern of correlations in Table 15 is that the traditional WLM metric performs worse than the two individual metrics, as though averaging somehow cancels them out. The implication is that list words and non-present target words may share inlinks (the same pages link to them), and they may share outlinks (they link to the same page), but they do not tend to share both at the same time. Thus there is an implicit asymmetry to the associative relationship that is lost if the gist-like representation considers both inlinks and outlinks. This finding is consistent with asymmetries in human similarity judgments [58] and may also explain why LDA performs so well at this task: It, unlike most distributional methods, is inherently asymmetric in the way it calculates gist.

**Table 15.** Spearman rank correlations with backward associative strength for DRM lists, after disaggregating WLM inlink/outlink measures (N = 55).

Model	Correlation
W3C3	0.34
W3C3 (inlink)	0.42
W3C3 (outlink)	0.42
WLM	0.24
WLM (inlink)	0.36
WLM (outlink)	0.34

We conducted a linear regression on ranks to evaluate the relative contributions of each constituent model. The regression used COALS, ESA, and WLM outlink scores converted to ranks to predict the DRM backward associative strength. In this first model COALS was not a significant predictor and so was removed. The results of the linear regression are presented in Table 16. Tolerance analyses were conducted to test for multicollinearity of ESA and WLM outlink by regressing each on the other. The tolerances were both 0.98, strongly indicating a lack of multicollinearity. Consistent with previous regressions, the fit of the model is very close, in this case identical, to the correlation of the W3C3 inlink/outlink models given in Table 15. Therefore it appears that even though COALS is not a significant predictor in this task, it does not detract from the overall performance of the W3C3 model.

**Table 16.** Regression of ESA and WLM outlink ranked scores on BAS ranks (N = 55).

Feature	B	SE(B)	$\beta$
ESA	0.262	0.127	0.262 *
WLM (outlink)	0.299	0.127	0.298 *

Notes:  $R = 0.42$ ,  $*p < 0.05$ .

Although the results in Table 15 are informative, they do not completely capture the qualitative data and intuition behind Table 14, the outlink structure for *sleep*. What that table describes is much further away from a gist-like representation and much closer to an activation based representation. Assuming that these articles are activated by list words, one can imagine activation flowing from them, through their outlinks and redirect links, to the article for *sleep*. Or put another way, instead of calculating the difference between entire vectors for list words and target word, only the vector element of the list word that corresponds to the target word is considered. We performed this analysis for Wikipedia outlinks on the DRM lists, using WLM’s model of outlink structure. For each word on the list, we collected all outlinks from all possible senses of that word and counted those pointing to the target article. Since the list words converge on a particular sense (the target word’s sense), taking the most frequent linked-to sense provided a strong predictor of backward associative strength,  $r(53) = 0.53$ .

We applied this same strategy to COALS and ESA. For COALS, we first identified the dimension associated with the non-presented target word. Then for each word on in the list, we retrieved the associated vector and added up the value at that target dimension. No normalization or SVD projection was used. Likewise, for ESA we found the target dimension and summed the word list vectors on that dimension. In the case of ESA, the standard normalization was used. The obtained correlations using these more activation-aligned models are presented in Table 17, along with the correlation for WLM outlink-based activation measure above.

**Table 17.** Spearman rank correlations with backward associative strength for DRM lists, using an activation-type metric (N = 55).

Model	Correlation
COALS Activation	0.23
ESA Activation	0.41
WLM Activation	0.53

As shown in Table 17, framing the model more in terms of activation rather than gist improves correlations considerably for ESA and WLM Activation models, such that the WLM Activation model has a non-significantly higher correlation with backward associative strength than the gist-type LDA model in Table 13 and W3C3 model presented in Table 15. For WLM this is perhaps not surprising because of how the outlinks on Wikipedia pages are generated in the first place: by people. Wikipedia’s guidelines on linking center on the likelihood that a reader will want to read the linked article [59]. It is up to the authors of Wikipedia pages to consider the association between one page’s topic and another before linking them. It seems only natural that some level of backward association strength would manifest in this process. By the same token, one might argue that WLM Activation only provides a circular definition of backward associative strength, since similar word associative processes are at work when people link to Wikipedia pages as are in word association tasks. While this is likely true, it also is evidence that the internal cognitive-linguistic processes involved in word association and DRM are externally represented in Wikipedia’s link structure.

As in previous work, these results seem to give some level of support to both activation/monitoring theory and fuzzy-trace theory. Activation/monitoring theory explains false memory largely in terms of the spreading activation in an associative network between words [56], which is consistent with the large predictive role of backward associative strength. Fuzzy-trace theory explains false memory in terms of gist, and gist traces, which are fuzzy semantic representations hypothesized to be more durable than verbatim traces, consistent with the finding that recall of the non-presented target words is greater than recall of list words and that the recall for list words decays more rapidly than the recall of target words [60]. The results in this section suggest that these two accounts might be different perspectives on the same underlying mental representation. Since a vector may be compared to another vector using all elements and therefore all contexts, a vector can be used to represent gist. Likewise, since the vector elements can be treated individually and the rest of the vector ignored, a vector element can be treated as an associative strength in a given dimension. However an important caveat is that in our models, an entire list of words is required to raise the activation level of the target dimension above the noise of the other dimensions. Thus this approach does not work for simple stimulus-response word association like the NMS task in Study 3.

## 7. Discussion

We believe that the results of Studies 1 through 4 substantiate the claim that humans and Wikipedia are part of the same cognitive-linguistic ecosystem. The literature described in Section 1 demonstrates how our cognitive-linguistic environment affects our language structure and categorization. If Wikipedia's structure is an externalization of internal cognitive and linguistic processes, then there is strong reason to believe in the cognitive-linguistic influence of Wikipedia's past authors on future readers. In other words, Wikipedia would appropriately be described by the process of niche construction.

There are some good reasons for taking this niche-construction concept seriously. It is perhaps trivially true that reading a book or similar work will have some effect on an individual, e.g., through learning. However, the argument being made here is stronger, that the influence of Wikipedia derives from both its language structure and its network of concepts/categories. Analogous to developmental studies [11,12], one prediction would be that reading Wikipedia would affect a participant's language structure and category structure. By creating a computational cognitive model based on Wikipedia and applying it to multiple semantic tasks, we indirectly tested this hypothesis and found support for it.

First, the unsupervised W3C3 model produced state of the art correlations with human data in Studies 1 to 3. We claim the model is unsupervised because in all cases the three constituent predictors were evenly weighted by taking their average, or equivalently, not weighted at all. Studies 1 and 2 are best characterized as semantic comparison tasks. Study 1's comparison task included a mixture of semantic relations, e.g., synonym, antonym, part-of, whereas Study 2's task involved overlap of semantic features. The high correlations in these studies, between 0.67 and 0.78, indicate that the information necessary make semantic comparisons is well represented in the structure of Wikipedia.

Second, the unsupervised W3C3 model had a higher correlation than any of its constituent models in Studies 1 to 4. Regressions conducted in these studies indicate that each constituent model explains a unique portion of the variance in the human data except COALS in Study 4, and that while their relative weights change slightly for each task, equal weighting is nearly identical to the

regression-derived weights in all cases. This finding directly supports our stronger claim: By allowing both the language structure and category structure of Wikipedia to guide the W3C3 model, we achieved a higher correspondence to human semantic behavior than if we had used either separately. Furthermore we allowed these two dimensions to influence each other by incorporating word-word, word-concept, and concept-concept levels into the model.

Our approach is consistent with a recently proposed new model visualization for language [61]. In this work traditional box-and-arrow models are characterized as modular, stage based, and rank ordered. In contrast, our W3C3 model does not have autonomous modules, but overlapping ones (word-word, word-concept, and concept-concept). These constituent models operate in parallel, without stages, and have no rank ordering or dominance. We differ from this new model visualization in that although our constituent models exist in multidimensional spaces (vector spaces), they do not occupy a single state space subject to dynamical state space processes.

We propose that by using vector spaces as a conceptualization, we have moved closer to unifying accounts of activation-based models and gist-based models, which are closely aligned with association and comparison-based tasks. Word association (Study 3) is a quite straightforward case of retrieval in the absence of a larger discourse context: Given a stimulus word, retrieve a response word. In contrast, semantic relatedness (Studies 1 and 2) is much more closely related to a comparison task. Intuitively, it asks the question: Given a pair of words, how does their meaning compare? We argue that distributional models by definition are much more aligned with semantic relatedness than with association. The reason is that semantic relatedness is a holistic judgment that considers many possible contexts. Distributional models are well aligned with holistic judgments because they are defined in terms of many contexts. Word association, on the other hand, can and does operate on a single dimension of a context. For example, in the NMS dataset, *tumor* has association with *kindergarten cop*. This stimulus-response pair is a clear reference to the film entitled, “Kindergarten Cop,” in which Arnold Schwarzenegger says in reference to his headache, “It’s not a tumor!” This line in this one film is quite possibly the only way in which the stimulus-response pair *tumor-kindergarten cop* is associated. This example illustrates that word associations need not be guided by many converging contexts but rather may be solely determined by a single context.

Study 4 in particular illustrates that gist and activation accounts are complementary views of the same underlying vector space structure. Both can have the same knowledge representation but differ in the operation performed on that structure.

Gist-like operations are inherently holistic, as in comparison tasks, and use the entire vector representation. When we applied the standard W3C3 model using gist-like measures, the correlation to backward associative strength was 0.34. In contrast, activation-like operations are inherently local, as in word association tasks, and can use only a single element of the vector. Incorporating activation into the gist-like W3C3 model by removing WLM inlink vectors increased correlation to 0.42. Further creating a completely activation-based measure using single WLM outlink vector elements increased correlation to 0.53. Since both gist-like and activation-like measures correlated positively with backward associative strength, these results explain why other studies may have evidence for either a semantic based or association based explanation of false memory [30,54]. However, rather than requiring different cognitive representations for word association or semantic comparison, we argue that word

association and semantic comparison are more productively viewed as task-driven operations on the same vector-based cognitive representation (*cf.* [51]). That the same underlying structure can be accessed differently according to different task demands is intuitive and matches observed behavior. Colunga and Smith [10] note that when children are asked to group a carrot, tomato, and rabbit, children will group rabbits and carrots together. However if children are told the carrot is a *dax* and are asked to find another *dax*, children will get the tomato. In the same way, comparison tasks evoke category structure and holistic judgments, whereas raw association as evoked by grouping or production tasks may be based on a single strong point of association, e.g., *rabbit–carrot*.

## 8. Conclusions

In summary, the internal cognitive-linguistic processes engaged in constructing Wikipedia has created a cognitive-linguistic environment that can be exploited by computational cognitive models. The crowd-sourcing process of creating, merging, and deleting article pages establishes a common view of shared concepts and topics for discussion. The words used within each page are a collaborative minimal summary of that concept, and the links between pages represent relevant associations between concepts. Wikipedia is perhaps unique in that it provides moderately clean structural relationships in natural language. As a product of the human mind, Wikipedia reflects aspects of human semantic memory in its structure. Our W3C3 model capitalizes on the cognitive-linguistic structure at different resolutions just as theories of memory purport access at different resolutions: COALS represents words in word contexts, ESA represents words in concepts, and WLM represents the links between concepts. The work we have presented in these studies suggests that these three resolutions may contribute to a more complete model of semantic association. Thus in creating Wikipedia to describe the world, we have created a resource that may reveal the subtleties of the human mind.

## Acknowledgements

The authors would like to thank Doug Rhode for his remarks on deriving the Pearson correlation of binary variables and the anonymous reviewers for their constructive feedback. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080594 and by the National Science Foundation, through Grant BCS0826825, to the University of Memphis. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education or the National Science Foundation.

## References

1. Miller, G.A. Informavores. In *The Study of Information: Interdisciplinary Messages*; Machlup, F., Mansfield, U., Eds.; Wiley: New York, NY, USA, 1983.
2. Clark, A. Language, embodiment, and the cognitive niche. *Trends Cogn. Sci.* **2006**, *10*, 370–374.
3. Michel, J.; Shen, Y.K.; Aiden, A.P.; Veres, A.; Gray, M.K.; Team, T.G.B.; Pickett, J.P.; Hoiberg, D.; Clancy, D.; Norvig, P.; *et al.* Quantitative analysis of culture using millions of digitized books. *Science* **2011**, *331*, 176–182.

4. Mitchell, S.D. *Biological Complexity and Integrative Pluralism*; Cambridge University Press: Cambridge, UK, 2003.
5. Chemero, A. *Radical Embodied Cognitive Science*; The MIT Press: Cambridge, MA, USA, 2009.
6. Christiansen, M.H.; Chater, N. Language as shaped by the brain. *Behav. Brain Sci.* **2008**, *31*, 489–509.
7. Pickering, M.; Ferreira, V. Structural priming: A critical review. *Psychol. Bull.* **2008**, *134*, 42–459.
8. Luka, B.J.; Choi, H. Dynamic grammar in adults: Incidental learning of natural syntactic structures extends over 48 h. *J. Mem. Lang.* **2012**, *66*, 345–360.
9. Woodward, A.L.; Hoyne, K.L. Infants' learning about words and sounds in relation to objects. *Child Dev.* **1999**, *70*, 65–77.
10. Colunga, E.; Smith, L.B. The emergence of abstract ideas: Evidence from networks and babies. *Philos. Trans. R. Soc. Lond. Series B Biol. Sci.* **2003**, *358*, 1205–1214.
11. Smith, L.B.; Jones, S.S.; Landau, B.; Gershkoff-Stowe, L.; Samuelson, L. Object name learning provides on-the-job training for attention. *Psychol. Sci.* **2002**, *13*, 13–19.
12. Waxman, S.R.; Markow, D.B. Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cogn. Psychol.* **1995**, *29*, 257–302.
13. Landauer, T.K.; Dumais, S.T. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychol. Rev.* **1997**, *104*, 211–240.
14. Landauer, T.K. LSA as a Theory of Meaning. In *Handbook of Latent Semantic Analysis*; Landauer, T., McNamara, D., Dennis, S., Kintsch, W., Eds.; Lawrence Erlbaum: Mahwah, NJ, USA, 2007; pp. 379–400.
15. Foltz, P.W.; Kintsch, W.; Landauer, T.K. The measurement of textual coherence with latent semantic analysis. *Discourse Process.* **1998**, *25*, 285–308.
16. Foltz, P.W.; Gilliam, S.; Kendall, S.A. Supporting content-based feedback in on-line writing evaluation with LSA. *Interact. Learn. Environ.* **2000**, *8*, 111–127.
17. Graesser, A.C.; Wiemer-Hastings, P.; Wiemer-Hastings, K.; Harter, D.; Tutoring Research Group; Person, N. Using latent semantic analysis to evaluate the contributions of students in autotutor. *Interact. Learn. Environ.* **2000**, *8*, 129–147.
18. Olde, B.A.; Franceschetti, D.; Karnavat, A.; Graesser, A.C. The Right Stuff: Do You Need to Sanitize Your Corpus When Using Latent Semantic Analysis? In *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*, Fairfax, USA, 7–10 August 2002; Erlbaum: Mahwah, NJ, USA, 2002; pp. 708–713.
19. Olney, A.M.; Cai, Z. An Orthonormal Basis for Topic Segmentation in Tutorial Dialogue. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics: Philadelphia, PA, USA, 2005; pp. 971–978.
20. Olney, A.M.; Cai, Z. An Orthonormal Basis for Entailment. In *Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference*, Clearwater Beach, FL, USA, 15–17 May 2005; AAAI Press: Menlo Park, CA, USA, 2005; pp. 554–559.
21. Harris, Z. Distributional structure. *Word* **1954**, *10*, 140–162.

22. McNamara, D.S. Computational methods to extract meaning from text and advance theories of human cognition. *Topics Cogn. Sci.* **2011**, 3, 3–17.
23. Wikipedia. Wikipedia: Statistics. 2007. Available online: <http://en.wikipedia.org/wiki/> (accessed on 8 February 2011).
24. Gabrilovich, E.; Markovitch, S. Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, 6–12 January 2007; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2007; pp. 1606–1611.
25. Gabrilovich, E.; Markovitch, S. Wikipedia-based semantic interpretation for natural language processing. *J. Artif. Int. Res.* **2009**, 34, 443–498.
26. Milne, D.; Witten, I.H. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, Chicago, IL, USA, 13–14 July 2008; AAAI Press: Chicago, IL, USA, 2008; pp. 25–30.
27. Finkelstein, L.; Gabrilovich, E.; Matias, Y.; Rivlin, E.; Solan, Z.; Wolfman, G.; Ruppin, E. Placing search in context: The concept revisited. *ACM Trans. Inf. Syst.* **2002**, 20, 116–131.
28. Nelson, D.L.; McEvoy, C.L.; Schreiber, T.A. The University of South Florida word association, rhyme, and word fragment norms, 1998. Available online: <http://www.usf.edu/FreeAssociation/> (accessed on 12 June 2011).
29. McRae, K.; Cree, G.S.; Seidenberg, M.S.; McNorgan, C. Semantic feature production norms for a large set of living and nonliving things. *Behav. Res. Methods* **2005**, 37, 547–559; PMID: 16629288.
30. Roediger, H.L.; Watson, J.M.; McDermott, K.B.; Gallo, D.A. Factors that determine false recall: A multiple regression analysis. *Psychon. Bull. Rev.* **2001**, 8, 385–407.
31. Rohde, D.; Gonnerman, L.; Plaut, D. An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence. Unpublished manuscript, 2005.
32. Hays, W. *Statistics for Psychologists*; Holdt, Rinehart, & Winston: New York, NY, USA, 1963.
33. Bullinaria, J.; Levy, J. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behav. Res. Methods* **2007**, 39, 510–526.
34. Trefethen, L.N.; Bau, II, D. *Numerical Linear Algebra*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 1997.
35. Manning, C.D.; Schütze, H. *Foundations of Statistical Natural Language Processing*; MIT Press: Cambridge, MA, USA, 1999.
36. Cilibrasi, R.L.; Vitanyi, P.M.B. The google similarity distance. *IEEE Trans. Knowl. Data Eng.* **2007**, 19, 370–383.
37. Milne, D. Wikipedia Miner. 2010 Available online: <http://wikipedia-miner.sourceforge.net> (accessed on 21 February 2011).
38. Wikipedia. Enwiki-20101011-pages-articles.xml. 2010. Available online: <http://download.wikimedia.org/enwiki/20101011/enwiki-20101011-pages-articles.xml.bz2> (accessed on 31 November 2010).

39. Gabrilovich, E. Explicit Semantic Analysis (ESA). 2011. Available online: <http://www.cs.technion.ac.il/~gabr/resources/code/esa/esa.html> (accessed on 5 December 2010).
40. Calli, C. Wikiprep-esa. 2011. Available online: <https://github.com/faraday/wikiprep-esa/archives/c36cb9481f46e9edabda1663b7a3be8c1b205bd5> (accessed on 15 December 2010).
41. Gabrilovich, E. Publicly available implementations. 2011. Available online: <http://www.cs.technion.ac.il/~gabr/resources/code/wikiprep/wikipedia-051105-preprocessed.tar.bz2> (accessed on 20 December 2010).
42. Gabrilovich, E. The WordSimilarity-353 Test Collection. 2011. Available online: <http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/> (accessed on 20 December 2010).
43. Agirre, E.; Alfonseca, E.; Hall, K.; Kravalova, J.; Paşca, M.; Soroa, A. A Study on Similarity and Relatedness Using Distributional and WordNet-Based Approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics ( NAACL '09)*, Association for Computational Linguistics: Stroudsburg, PA, USA, 2009; pp. 19–27.
44. Pirró, G.; Euzenat, J. A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. In *Proceedings of the 9th International Semantic Web Conference on the Semantic Web—Volume Part I*; Springer-Verlag: Berlin, Germany, 2010; ISWC'10, pp. 615–630.
45. Reisinger, J.; Mooney, R. A Mixture Model with Sharing for Lexical Semantics. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing ( EMNLP '10)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2010; pp. 1173–1182.
46. Tsatsaronis, G.; Varlamis, I.; Vazirgiannis, M. Text relatedness based on a word thesaurus. *J. Artif. Int. Res.* **2010**, *37*, 1–40.
47. Yeh, E.; Ramage, D.; Manning, C.D.; Agirre, E.; Soroa, A. WikiWalk: Random Walks on Wikipedia for Semantic Relatedness. In *Proceedings of the 2009 Workshop on Graph-Based Methods for Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2009; pp. 41–49.
48. Riordan, B.; Jones, M.N. Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics Cogn. Sci.* **2011**, *3*, 303–345.
49. Conover, W.J.; Iman, R.L. Rank transformations as a bridge between parametric and nonparametric statistics. *Am. Stat.* **1981**, *35*, 124–129.
50. Brainerd, C.J.; Yang, Y.; Reyna, V.F.; Howe, M.L.; Mills, B.A. Semantic processing in “associative” false memory. *Psychon. Bull. Rev.* **2008**, *15*, 1035–1053.
51. Maki, W.; Buchanan, E. Latent structure in measures of associative, semantic, and thematic knowledge. *Psychon. Bull. Rev.* **2008**, *15*, 598–603.
52. McRae, K.; Jones, M.N. Semantic Memory. In *The Oxford Handbook of Cognitive Psychology*; Reisberg, D., Ed.; Oxford University Press: Oxford, UK, 2012. In Press.
53. McRae, K.; Khalkhali, S.; Hare, M. Semantic and Associative Relations in Adolescents and Young Adults: Examining a Tenuous Dichotomy. In *The Adolescent Brain: Learning, Reasoning, and Decision Making*; Reyna, V.F., Chapman, S.B., Dougherty, M.R., Confrey, J., Eds.; American Psychological Association: Washington, DC, USA, 2011; pp. 39–66.

54. Cann, D.R.; McRae, K.; Katz, A.N. False recall in the Deese-Roediger-McDermott paradigm: The roles of gist and associative strength. *Q. J. Exp. Psychol.* **2011**, *64*, 1515–1542.
55. Griffiths, T.L.; Steyvers, M.; Tenenbaum, J.B. Topics in semantic representation. *Psychol. Rev.* **2007**, *114*, 211–244.
56. Roediger, H.L.; Gallo, D.A. Associative Memory Illusions. In *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*; Pohl, R., Ed.; Psychology Press: East Sussex, UK, 2004.
57. Gallo, D.A.; Roediger, H.L.; McDermott, K.B. Associative false recognition occurs without strategic criterion shifts. *Psychon. Bull. Rev.* **2001**, *8*, 579–586.
58. Tversky, A. Features of similarity. *Psychol. Rev.* **1977**, *84*, 327–352.
59. Wikipedia. Manual of Style (linking). 2011. Available online: [http://en.wikipedia.org/wiki/Wikipedia:Manual\\_of\\_Style\\_\(linking\)](http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style_(linking)) (accessed on 26 January 2011).
60. Seamon, J.; Luo, C.; Kopecky, J.; Price, C.; Rothschild, L.; Fung, N.; Schwartz, M. Are false memories more difficult to forget than accurate memories? The effect of retention interval on recall and recognition. *Mem. Cogn.* **2002**, *30*, 1054–1064.
61. Onnis, L.; Spivey, M.J. Toward a new scientific visualization for the language sciences. *Information* **2012**, *3*, 124–150.

© 2012 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).