# Small Worlds and Big Data: Examining the Simplification Assumption in Cognitive Modeling

Brendan T. Johns[1], D. J. K. Mewhort[2], & Michael N. Jones[3]

[1]University at Buffalo

[2]Queen's University

[3]Indiana University

Corresponding Address
Brendan Johns
Department of Communicative Disorders and Sciences
University at Buffalo
122 Carey Hall
Buffalo, NY 14214

## Abstract

The simplification assumption of cognitive modeling proposes that to understand a given cognitive system, one should focus on the key aspects of the system and allow other sources of complexity to be treated as noise. The assumption grants much power to a modeller, permits a clear and concise exposition of a model's operation, and allows the modeller to finesse the noisiness inherent in cognitive processes (e.g., McClelland, 2009; Shiffrin, 2010). The small-world (or "toy model" approach) allows a model to operate in a simple and highly controlled artificial environment. By contrast, big-data approaches to cognition (e.g. Landauer & Dumais, 1997; Jones & Mewhort, 2007) propose that the structure of a noisy environment dictates the operation of a cognitive system. The idea is that complexity is power; hence, by ignoring complexity in the environment, important information about the nature of cognition is lost. Using models of semantic memory as a guide, we examine the plausibility, and the necessity, of the simplification assumption in light of big-data approaches to cognitive modeling.

Approaches to modeling semantic memory fall into two main classes: those that construct a model from a small and well controlled artificial dataset (e.g. Collins & Quillian, 1969; Elman, 1990; Rogers & McClelland, 2004) and those that acquire semantic structure from a large text corpus of natural language (e.g., Landauer & Dumais, 1997; Griffiths, Steyvers, & Tenenbaum, 2007; Jones & Mewhort, 2007). We refer to the first class as the *small-world approach*[1] and the latter as the *big-data approach*. The two approaches differ on the method by which to study semantic memory and exemplify a fundamental point of theoretical divergence in subfields across the cognitive sciences.

The small-world approach to semantic memory is appealing because it makes ground-truth possible; all of the knowledge about a world is contained within the training materials, and the true generating model is known. For example, Figure 1 reconstructs the taxonomic tree from Collins & Quillian (1969, and is also used to generate the training materials in Rogers & McClelland, 2004). The reconstructed tree contains knowledge about a limited domain of living things, namely plants and animals. Sampling propositions from the tree (as Rogers & McClelland do), produces a well-structured and accurate knowledge for a limited domain of items. Thus, even though it is small in scale, the theorist knows the structure of the world under question (e.g., that a *canary* is a *bird*), and hence, is able to gauge how well the model acquires and processes information from the training set.

The big-data approach to the same problem is fundamentally different. Rather than assuming that the world is structured in a single way, it proposes that the lexical environment provides the ground-truth for the meaning of words, an approach that also has a rich history (e.g. Wittgenstein, 1953). Although different learning mechanisms have been proposed (see Jones, Willits, & Dennis, 2014 for a review), they all rely upon co-occurrence patterns of words across large text corpora to infer the meaning of a word. A model's performance is assessed on a behavioral task (e.g., a synonym test, ratings of semantic similarity, semantic priming, etc.) where the model's use of a word is compared against human performance. The model's match to the item-level behavior is used to justify the plausibility of the specific algorithm used by the model.

Big-data techniques underscores the amount of information that is available in the linguistic environment about a word's meaning and the kind of mechanisms that are needed to learn from it (e.g. Recchia & Jones, 2009; Bullinaria & Levy, 2012). Although big-data approaches include more noise and uncertainty than the small-world approaches, they have greater ecological validity. The big-data approach reminds us of the natural environment in which cognitive agents are embedded, ideas that have a long history in the cognitive sciences (Simon, 1969; Estes, 1975).

[1] What we refer to here as a small-world approach is also commonly referred to as a "toy model" approach.

As engines with which to understand semantic memory, both approaches have advantages and disadvantages. Small worlds support a more complete understanding of how a model functions and what type of data it can learn than the big-data approach; the big-data approach, by contrast, learns the structure of the natural linguistic environment, data that are more plausible than hand-constructed representations. Hence, there is a trade-off of clarity for plausibility in the different techniques.

Clarity is a central tenant of the simplification assumption in cognitive modeling (McClelland, 2009; Shiffrin, 2010). The assumption allows researchers to focus on specific aspects of the cognitive system, while assuming other complexities away. As McClelland (2009, p. 18) states "The more detail we incorporate, the harder the model is to understand." However, as McClelland (2009) also points out, there is a need to ensure that one's assumptions are plausible and accurate. Otherwise one may be misled by assuming away too many details.

In the study of semantic memory, the small-world approach exemplifies the simplification assumption (Rogers & McClelland, 2004). The approach assumes that the environment is structured in systematic ways and that the important issue to be understood is how this information is acquired and processed. The risk is, however, that a small word ignores the complexity of the problem—the acquisition of meaning from the natural environment. By using artificially structured data to train a model, one risks endorsing a learning mechanism that cannot be scaled to the complexity of the natural environment. That is, the approach may trade clarity for plausibility; such a trade-off forces the model to solve an artificial problem that is not representative of human experience.

Our goal is to explore the issues of simplification in cognitive modeling, using semantic memory as an example. The first part of the chapter examines multiple small worlds used in semantic memory modeling with the goal of understanding the power that using highly structured data sets provides. The second part examines the amount of natural language information that is needed to approximate the structure that is contained in the small world. We will use a standard co-occurrence model of semantics to examine both (Jones & Mewhort, 2007; Recchia, et al., 2015). By assessing the amount of data needed to form an approximation, it provides an insight into the relationship between the small world and the big data approaches.

## 1. Analysis of a Small World

In their book *Semantic Cognition*, Rogers and McClelland (2004) offer a simple and powerful model of semantic memory. This model accounts for a variety of semantic memory phenomena, such as the learning of different concepts, developmental trends, and the use of causal knowledge in semantics. The basic operation of the model is backpropagation of propositional information within a connectionist network, based on the previous work of

Rumelhart (1990) and Rumelhart and Todd (1993). The model is experience dependent and requires training materials, derived from a corpus of propositions, to account for behavior.

Although we accept many of the central tenets of the Rogers and McClelland semantic cognition theory, we question one aspect of their approach, namely the training set used. They used a hierarchical representation of the world, building off of Collins and Quillian (1969, see in Figure 1), to generate a corpus of training propositions. Only the item-level propositions are used in training (e.g. *canary isa living_thing*, *canary isa bird*, *canary can fly*, etc.). Two hierarchies were used, one with 8 items and one with 21 items.

Although the plausibility of using such propositional information has been questioned elsewhere (e.g. Barsalou, 1999), our question is the requisite amount of structure in the training materials. The small world, as outlined in Figure 1, is essentially noise free and contains a great deal of structure about the relationship between items. The natural environment is much noisier and ambiguous than this setup. Given such heavily structured training materials, we question whether the explanatory power of the model derives from the structure of the materials used in training or the learning mechanism used by the model.

The question echoes recent work in the implicit memory literature: there may be no need for an advanced learning mechanism to acquire the structure of an artificial grammar, but rather that the similarity structure of the items to which subjects are exposed contains sufficient information to model the relevant behavioral findings using a simple memory model (Jamieson & Mewhort, 2009a, b; 2010; 2011). That is, the structure of the training materials is sufficiently constrained that a simple encoding provides enough power to account for the behavior under question. Recent work in natural language sentence processing has strengthened the importance of item-level information in the modeling of language effects (Bannard, Lieven, & Tomasello, 2012; Johns & Jones, 2015).

To analyze the amount of structure in Rogers and McClelland's (2004) small world, we will employ the BEAGLE model of lexical semantics (Jones & Mewhort, 2007) to construct representations of items. BEAGLE has been used in several domains, including semantic priming (Jones, Kintsch, & Mewhort, 2006; Hare, et al., 2009), memory search (Hills, Jones, & Todd, 2012), and semantic deficits in cognitive impairment (Johns, et al., 2013). A key aspect of the model—one that makes it appealing for use here—is that its lexical representations are a direct reflection of the language environment; there is no advanced inference or abstraction procedure. Thus, its representation will provide insight into just how structured the small world is.

BEAGLE uses four vectors to represent a word: 1) a static environment vector (with elements sampled from a Gaussian distribution in the standard implementation), which is assumed to represent perceptual properties of a word, 2) a context vector, which represents pure

co-occurrence relationships, 3) an order vector, which records the position of a word relative to the other words in a sentence, and 4) a composite vector formed by summing the context and order vectors. A word's context vector is updated with the sum of the environmental vectors for the other words appearing in the same sentence. A word's order vector is formed by binding it with all ngram chunks in the sentence with directional circular convolution (see Jones & Mewhort, 2007 for additional detail).

The BEAGLE model offers a different view of semantic memory than Rogers & McClelland's (2004) approach. The most obvious difference is that BEAGLE does not depend on error-driven learning, unlike a backpropagation model. BEAGLE is a vector accumulation model; it builds a representation of the world through the direct encoding of experience without requiring an error signal. Hence, meaning is not based in the refinement of the predictions from the model, but from the gradual buildup of knowledge through episodic experience. The vector accumulation approach to semantics provides a simple framework to determine the power of experience in explaining structure in semantic memory.

### 1.1. Learning a Small World

In this section, we will examine the acquisition of hierarchical knowledge through the learning of propositional information. Rogers and McClelland demonstrated that, across hundreds or thousands (in the case of the large network) of learning epochs, the model gradually acquired a hierarchical representation of the items, similar to the semantic network displayed in Figure 1. They assumed that semantic memory is generally hierarchical so the acquisition of a hierarchical representation was of central importance to their modeling goals.

To determine how well BEAGLE learns this type of structure, we used the same set of propositions from Rogers and McClelland (2004). This was done for both the small network displayed in Figure 1, which includes 8 items and 84 item-level propositions (propositions with the items in them), as well as the larger semantic network contained in Appendix B.3 of Rogers and McClelland (2004). The larger semantic network consisted of 21 words and 220 item-level propositions.

To train BEAGLE, 25 propositions were iteratively sampled randomly with replacement, and the word vectors were updated with those propositions. Iteratively sampling with replacement means that a set of 25 propositions were selected at each training point, with a selected proposition not being removed from the search set if it was chosen (that is, it could be sampled multiple times across training). Each proposition was treated as a sentence, with both relations (e.g. *isa*, *has*, *can*, *is*) and features (*fly*, *feathers*, *wings*, *swim*, etc.), being represented as words with their own unique environmental vectors. Words were represented as the composite of the context and order vectors, similar to the approach taken to understanding artificial grammar acquisition in Jamieson & Mewhort (2011). Similarity values taken from Beagle were

averaged across 25 resamples of the environmental vectors, to ensure that the results were not due to the random initial configuration of similarity and that the emerging structure is from the learning of propositional information.

To test how well the model learned the structure contained in the semantic networks, a rank correlation of the vector cosine of the word's representation to the number of shared features between the words in the semantic network was taken. For example, in the network displayed in Figure 1, *robin* and *canary* share 10 features (*isa bird, has wings, can fly, has feathers, isa animal, can move, has skin, can grow, is living, isa living_thing*), while *robin* and *oak* contain 3 (*isa living_thing, can grow, is living*). We used a simple metric of learning by assessing how related the similarity values are to the feature overlap values. We use dendograms generated from a hierarchical clustering algorithm to display that the model is learning accurate hierarchical information, similar to how structure was illustrated by Rogers and McClelland (2004).

Figure 2 shows the increase in correlation of the BEAGLE similarity values to the amount of feature overlap of words in the two proposition corpora, as a function of the number of propositions studied for both the small and large networks. Figure 2 shows a simple trend: The model learned the structure of the small world rapidly, even with minimal exposure to the training materials. For the small network, performance hits asymptote at 75 propositions, close to the size of the actual corpus. However, even at only 25 propositions the model was capable of inferring a large amount of structure. As would be expected, it took longer for the model to learn the structure of the large semantic network, with the model hitting asymptote at around 150 propositions. The entire proposition set is not needed to acquire accurate knowledge of the small world, due to the highly structure nature of the learning materials. That is, the model was capable of acquiring an accurate representation of the small world with only 75% of the total number of propositions, in only a single training session.

To determine if the model reproduced the hierarchical structure of the propositions correctly, Figure 3 shows the dendogram of the hierarchical clustering of the similarity space for the small network, while Figure 4 contains the same information for the large network. The dendograms were generated from a BEAGLE model that was trained on all of the propositions from the different networks, with no repeats. A dendogram uses hierarchical clustering to determine the hierarchical relationships among items. Figures 3 and 4 show that the model learned the correct hierarchical generating structure of the environment, in that it is identical to the structure displayed in the semantic network in Figure 1 (for the small network). This demonstrates that the Beagle model was able to acquire hierarchical information (central to the proposals of the semantic cognition theory), even with no explicit goal to do so, and with a very limited amount of experience (as compared to the amount of training that a typical backpropagation algorithm would require). These simulations demonstrate that a simple encoding of the training materials provides enough power to learn both networks.

### 1.2. Discussion

Our goal in this section was twofold: first to determine the power that a well-formed description of a small world provides a learning mechanism (in the form of sets of propositions derived from semantic networks), and secondly to assess how easily this information is learned with a simple co-occurrence model of semantics. We did not intend to compare the vector accumulation techniques of Jones & Mewhort (2007) with the backpropagation techniques of Rogers & McClelland (2004). Both are perfectly capable of learning the required information, but the BEAGLE model provides a simple method of determining the power contained within the structure of the training materials. There is nothing complicated about the vector accumulation model; it provides a mechanism for efficiently recording the usage of a word. The model's explanatory power comes from the statistical structure of the environment.

BEAGLE was able to learn the structure of the small worlds efficiently, with very limited training, through exposure to propositional information. One difference between the vector accumulation approach and that of Rogers & McClelland (2004) is that we assumed propositional information was equivalent to a sentence in a natural language corpus. Rogers & McClelland, by contrast, assumed that the goal of the learning operation was to take a word (e.g. *canary*) and a relation (e.g. *can*) to construct a prediction about what the output pattern should be (e.g. *sing*). There is no a priori reason why either approach should be preferred over the other, as in both models the learning tasks are completely linguistic (that is, there is no formal basis for assuming that the training materials represent more than patterns to associate).

However, the Beagle model did not even require the full proposition set to learn the small world (that is, it required less than a single training epoch); backpropagation required hundreds or thousands of epochs of training to capture the correct hierarchical structure of the training materials (Rogers & McClelland, 2004), depending on the size of the training corpus. The differential amount of training required to learn the latent structure of the small worlds provides an interesting look into the motivation of the two theories. Rogers & McClelland used the evolution of knowledge gained across epochs to relate the model to developmental trends in the acquisition of knowledge. The BEAGLE model was not designed to explain small datasets, and, given the limited training materials required for the model to acquire the small world, it would not be possible to speak to developmental data with this analysis. Instead, the appealing aspect of the framework is that it can be easily scaled up to analyze massive amounts of text (Recchia, et al., 2015), on a scale that is consistent with the amount of language a typical adult would experience.

The ability to scale provides an opportunity to constrain the learning mechanisms used to explain semantic memory, as it should be possible to determine the amount of natural-language information that is necessary to learn than the structure of a proposed small world. Thus, it allows for a connection to be formed between the assumptions of the small world approach

(heavily structured, small-scale training materials) with the big data approach (noisy, large-scale data) to understanding semantic memory. This affords a formal relationship to be formed between the two approaches: Given the structure of a small world (where the simplification assumption makes the learning task much more straightforward), it should be possible to determine how much natural language information is necessary to approximate a small world structure.

## 2. Big Data Analysis of Small-World Structure

This section will assess the minimum amount of natural language information that would be necessary to learn the approximate structure of a proposed small world with a semantic space model. The aim is an understanding of the complexity of scaling from a small world to the natural environment. We will use a vector accumulation model to analyze a large, and unique, collection of high quality texts. A data fitting methodology will be used to determine the most informative set of texts to approximate the structure of a small world. The high quality of the texts and the use of a data fitting method will allow for confidence that a set of highly informative texts is being assembled.

### 2.1. Model

The model that will be used here is an approximation of BEAGLE that is based on sparse representations rather than Gaussian ones (Sahlgren, Host, & Kanerva, 2008; Recchia, et al., 2015). The advantage of the sparse-vectors approach is that it greatly reduces the computational complexity of the model and allows for a greater degree of scaling. A very large amount of text is going to be used in this analysis, so the computationally simpler model has obvious advantages. Only context vectors will be used, rather than order vectors, in order to simplify the learning task, as now the model is only using sentential context to form semantic representations. Similar to past studies (e.g. Recchia, et al., 2015), vectors will be large (5,000 dimensional) and the environmental vectors are very sparse (6 non-zero values randomly sampled), similar to binary spatter codes.

### 2.2. Training materials

The set of texts that will be used to train the model is drawn from five different sources: 1) Wikipedia, 2) Amazon product descriptions (from McAuley & Leskovec, 2013), 3) a set of 1,000 fiction books, 4) a set of 1,050 non-fiction books, and 5) a set of 1,500 young adult books. All book text was scraped from e-books, and all were published in the last 60 years by popular authors. To ensure that each text source would contribute equally, each source was trimmed to a set of six million sentences with random sampling, for a total of 30 million sentences across all texts (approximately 400 million words). The data fitting method will determine which set of texts are the most informative for generating the small worlds.

### 2.3. Data fitting methodology

Rather than training the model with all language materials, we used a new way to determine which sets of text offer the best fit to a proposed small world. Taking the different corpora described above, we split them into subsections of 10,000 sentences (approximately the size of an average fiction novel, a rather small amount of language). The result was 3,000 separate text subsections. The subsections were treated as individual corpora, and a set of vectors were learned for each subsection. That is, 3,000 different models were constructed. These representations were rather sparse in terms of the amount of knowledge that they contain, given the rather limited amount of language each part contains.

These 3,000 different vector sets were used to generate an overall representation that was maximized on the amount of knowledge contained about the small world. To do this, a hill-climbing algorithm was used to determine which parts provide the best fit to the proposed structure of a small world, by iteratively selecting the sections that provides the largest increase in fit to the structure of the small worlds. The first iteration of the method selects the section that provides the best fit. On subsequent iterations, vector sets are summed together to form an overall representation. Whichever section provides the largest increase in fit is selected and summed into the representation. All sets are combined with the overall representation at each iteration, meaning that the entire language space is being tested. Once a vector set has been selected, it is removed from the search set. In this way, the semantic representation that is constructed continuously increases its resolution, as is the amount of knowledge that they contain about a small world. The process ends when the addition of further vectors into the overall representation decreases the fit of the model to the data.

Although the problems with hill-climbing algorithms are well-known (e.g., getting stuck in local maxima), the method used here provides a simple means by which to determine how much linguistic information is necessary to form an approximation of the structure contained in a small world. One could think of the process as a form of parameter fitting, whose use is ubiquitous across cognitive modeling (see Shiffrin, Lee, Kim, & Wagenmakers, 2008), but instead of maximizing the internal parameters to explain a set of behavioral data, we instead maximized the structure of the external world (i.e. linguistic information).

### 2.4. Small worlds

The small worlds that will be approximated here are the same that were used in Part 1, both the small and large semantic networks. The only difference was that the word *sunfish* was replaced with *trout*, due its very low frequency across the different corpora. The first tests conducted examined the similarity values (assessed with a vector cosine) between the word-level items in the hierarchy. It was assessed with a rank correlation to the amount of feature overlap in

the semantic network, identical to the small world analysis described above. This analysis will provide an analogous examination to those used in part 1.

In order to further test the knowledge acquisition process, we used an additional test over both semantic networks. Specifically, we used an alternative forced choice (AFC) task, where the model has to determine which semantic feature (e.g. *plant* or *animal*) is more likely for a particular word (e.g. *canary*). This test was conducted at each level of the hierarchy (e.g. the model was asked to discriminate *plant*/*animal*, and then *tree*/*flower*/*bird*/*fish*/*mammal*, etc… until the bottom level was reached). The test involves 52 questions for the small network and 140 questions for the large network, derived from every level of the hierarchy for both of the networks. Not all levels contained the same number of features, so the number of alternatives ranges from 2 to 10. Obviously, discriminating among 10 alternatives is a difficult task for the model, but it does provide a strong test of the semantic representation that the model is constructing. Performance was assessed by determining the proportion of correct responses on the AFC test.

### 2.5. Results

Figure 5 shows the fit of the model when it is optimized to account for feature overlap values for both the small and large semantic networks. As shown in Figure 5, the model is capable of rapidly acquiring the semantic structure of a small world, as the correlation between the model's vector cosine and the feature overlap values from the network increases substantially as more linguistic information is learned. The complexity of the small world obviously plays a large role in the amount of linguistic experience that is necessary to account for the proposed structure. The small network maximized at 150,000 sentences, while the large network maximized at 880,000 sentences. However, the small network hit asymptote at around 50,000 sentences, however, while the large network did the same at approximately 200,000 sentences.

To demonstrate that the model acquired the same hierarchical information learned in the small-world modeling, the dendograms for both the small and large networks are displayed in Figure 6 and Figure 7, respectively. For the small network, the hierarchical clustering method accurately infers the clusters across the 8 words. For the large network, the model reproduces the overall cluster properties (clustered into *trees*/*flowers*/*fish*/*bird*/*mammals*), with only one error (*robin* was classified in its own cluster, closer to *mammals*). The error arose because the word *robin* had approximately equal similarity to both *birds* and *mammals*. Additionally, the clusters are not as well discriminated as the dendogram in Figures 3 and 4, a result expected given the differences in noise across the training sets. Natural language contains much more knowledge about the words under question (e.g., that a robin nests in trees), that shift the similarity space. However, the simulation is impressive because acquiring hierarchical information not an explicit goal for the model. Instead, the language environment was enough to

acquire such data, as the hierarchical structure emerged across training, similar to the findings of Rogers & McClelland (2004).

A stronger test of the model's ability to acquire the requisite information is given by the AFC test described above. In the AFC test, the model has to discriminate features associated with the word in the semantic networks. Figure 8 displays the increase in performance across the corpus sampling routine. Even though the learning task was quite difficult, the model reached an impressive 98% accuracy for the small network questions, and an 87% accuracy for the large network questions. The test is similar to a synonym test, such as the TOEFL test used in Landauer & Dumais (1997), where LSA achieved a performance level of 55%. That the model achieves such a high performance level demonstrates the power both of the training materials that were assembled and of the data fitting method that was used. Rather surprisingly, the model did not require as many sentences to learn the features of the semantic network as it did to learn the connection among words, as only 130,000 sentences were needed to reach maximum performance for the small network, and 560,000 sentences were needed for the large network. It is worth noting that the TASA corpus—a standard corpus used in co-occurrence models of semantics since Landauer & Dumais (1997)—contains approximately 760,000 sentences. So, the number of sentences that the optimization method required is not overly large when compared with standard corpora that are used in the field of semantic memory. The test demonstrates that a model trained directly with data from the linguistic environment not only learned the hierarchical structure of a small world but also learned the semantic features that define it.

The simulations demonstrate that a big-data analysis approximates the structure contained in small worlds quite readily. However, a different question lies in the comparative complexity of the learning task that models of semantic memory face. For the small network, the small world had 84 propositions. In the big-data analysis, the model maximized performance at 1,807 times more sentences than is contained in the proposition corpus. For the larger semantic network, the scale was that 3,911 times more sentences were required. However, the feature AFC test only needed 1,566 and 2,488 times more sentences than the proposition corpus for the small and large networks respectively, suggesting that different types of information can be learned more efficiently. Overall, our analysis suggests that the amount of information contained in a small world versus that contained in big data are on scales that differ by multiple orders of magnitude.

### 2.6. Discussion

This section attempted to understand of the connection between the proposals of small world analyses of semantic memory, compared to those that rely upon big data. We found that a model which relied upon large amounts of lexical experience to form a semantic representation of an item was able to acquire a high quality approximation to the structure of different small worlds, assessed with multiple tests, including the acquisition of hierarchical information and the

learning of semantic features. The corpus-sampling algorithm allowed the model to select a set of texts that provided the best fit to the small world structure. Even with this data fitting methodology, to obtain a maximal fit required large amounts of natural language materials.

## 3. General Discussion

The goal of this article was to examine the simplification assumption in cognitive modeling by comparing the small world versus big data approaches to semantic memory and knowledge acquisition. A central aspect of semantic memory is the learning of environmental structure through experience. The small-world approach proposes that in order to understand the mechanisms that underlie this ability, the structure of the environment must be constrained to lay bare the learning process itself. By constraining the complexity of the training materials, it is possible to study the operations of a model at a fine-grain level, as the theorist knows the generating model that produced the environmental experience. In contrast, a big-data approach proposes that ecologically valid training materials are necessary – natural language materials. Although the researcher loses control of the minutiae of what the model learns, it gains power through the plausibility of the approach, as such models readily scale up to the linguistic experience that humans receive.

To determine the relation between these two approaches, Part 1 examined how readily a vector-accumulation model of semantics (a standard big-data model) could acquire the structure of a small world. We found that the model rapidly acquired knowledge of the small world through a small corpus of propositions. By limiting the complexity of the training materials, the learning task became quite simple. The structure of the items used was sufficient to explain the behaviors under question, with no advanced learning or abstraction processes necessary. This echoes recent work on implicit memory and artificial grammar learning (e.g. Jamieson & Mewhort, 2011) and natural language sentence processing (Johns & Jones, 2015).

Given the ease with which the model was capable of learning the small world, Part 2 determined how much natural language information was necessary to approximate the structure proposed by the two semantic networks from Rogers & McClelland (2004). We used a very large amount of high quality language sources – large sets of fiction, non-fiction, and young adult books, along with Wikipedia, and Amazon product descriptions. The texts were split into smaller sections, and a hill-climbing algorithm was used to select the texts iteratively that allowed the model to attain the best fit to the proposed structure. Across multiple tests, the model was readily capable of learning the small worlds, but the amount of experience needed was orders of magnitude greater than the size of the proposition corpora. This analysis suggests that the learning tasks under the two approaches differ greatly in the complexity of the materials used.

This issue of informational complexity gets at the crux of the problems surrounding the simplification assumption in studying semantic memory: By reducing the complexity of the structure of the environment to a level that is tractable for a theorist to understand fully, the problem faced by a hypothesized learner is trivialized. The linguistic environment, although heavily structured in some ways, is still extremely noisy, and requires learning mechanisms that are capable of discriminating items contained in very large sets of data. Thus, the simplification assumption biases theory selection towards learning and processing mechanisms that resemble humans on a small and artificial scale. The emergence of big-data approaches to cognition suggests that artificial toy datasets are no longer necessary. Models can now be trained and tested on data that is on a similar scale to what people experience, increasing the plausibility of the model selection and development process. For models of semantic memory, the existence of high quality large corpora to train models with eliminates the necessity for oversimplification, and offers additional constraints as to how models should operate.

This is not to say that the small-world approach is without merits, a point clear in the history of cognitive science. The goal of small-world assumption is to provide an accurate understanding of the operation of a model with clear and concise examples, something that models that focus only on big-data techniques cannot achieve. Thus, as in the evolution of any new theoretical framework, past knowledge should be used to constrain new approaches. The big-data approach should strive to embody the ideals of the simplification assumption in cognitive modeling, that of clear and concise explanations, while continuing to expand the nature of cognitive theory.

## References


Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 17284 –17289.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral Brain Science*, 22, 577-660.

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods*, 44, 890–907.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8, 240-247.

Estes, W. K. (1975). Some targets for mathematical psychology. *Journal of Mathematical Psychology*, 12, 263-282.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in Semantic Representation. *Psychological Review, 114*, 211-244. DOI: 10.1037/0033-295X.114.2.211

Hare, M. Jones, M. N., Thomson, C., Kelley, S., & McRae, K. (2009). Activating event knowledge. *Cognition*, 111, 151-167.

Hills, T. T., Jones, M. N., & Todd, P. T. (2012). Optimal foraging in semantic memory. *Psychological Review*, 119, 431-440.

Jamieson, R. K., & Mewhort, D. J. K. (2009a). Applying an exemplar model to the artificial-grammar task: Inferring grammaticality from similarity. *Quarterly Journal of Experimental Psychology*, 62, 550-575.

Jamieson, R. K., & Mewhort, D. J. K.. (2009b). Applying an exemplar model to the serial reaction time task: Anticipating from experience. *Quarterly Journal of Experimental Psychology*, 62, 1757-1784.

Jamieson, R. K., & Mewhort, D. J. K. (2010). Applying an exemplar model to the artificial-grammar task: String-completion and performance for individual items. *Quarterly Journal of Experimental Psychology*, 63, 1014-1039.

Jamieson, R. K., & Mewhort, D. J. K. (2011). Grammaticality is inferred from global similarity: A reply to Kinder (2010). *Quarterly Journal of Experimental Psychology*, 64, 209-216.

Johns, B. T., Taler, V., Pisoni, D. B., Farlow, M. R., Hake, A. M., Kareken, D. A., Unverzagt, F. W., & Jones, M. N. (2013). Using cognitive models to investigate the temporal dynamics of semantic memory impairments in the development of Alzheimer's disease. In the *Proceedings of the 12th International Conference on Cognitive Modeling* (ICCM).

Johns, B. T., & Jones, M. N. (2015). Generating structure from experience: A retrieval-based model of sentence processing. *Canadian Journal of Experimental Psychology*, 69, 233-251.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, 55, 534-552.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114, 1-37.

Jones, M. N., Willits, J. A., & Dennis, S. (in press). Models of semantic memory. In J. R. Busemeyer & J. T. Townsend (Eds.) *Oxford Handbook of Mathematical and Computational Psychology*.

Landauer, T. K., & Dumais, S. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.

McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems, Rec-Sys '13*, pp. 165-172.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, 1, 11-38.

Recchia, G. L., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information to latent semantic analysis. *Behavior Research Methods*, 41, 657-663.

Recchia, G. L., Jones, M. N., Sahlgren, M., & Kanerva, P. (2015). Encoding sequential information in vector space models of semantics: Comparing holographic reduced representation and random permutation. *Computational Intelligence and Neuroscience*. http://dx.doi.org/10.1155/2015/986574

Rogers, T. T., & McClelland, J. L. (2004). *Semantic Cognition: A parallel distributed processing approach*. MIT Press.

Rumelhart, D. E. (1990). (1990) Brain style computation: Learning and generalization. In: An introduction to neural and electronic networks, ed. S. F. Zornetzer, J. L. Davis & C. Lau, pp. 405–20. Academic Press.

Rumelhart, D. E. & Todd, P. M. (1993) Learning and connectionist representations. In: *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience*, ed. D. E. Meyer & S. Kornblum, pp. 3–30. MIT Press.

Sahlgren, M., Holst, A., & Kanerva, P. (2008). Permutations as a means to encode order in word space. *Proceedings of the 30th Conference of the Cognitive Science Society*, pp. 1300–1305.

Shiffrin, R.M., Lee, M.D., Kim, W. & Wagenmakers, E. -J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, 32, 1248-1284.

Shiffrin, R. M. (2010). Perspectives on modeling in cognitive modeling. *Topics in Cognitive Science*, 2, 736-750.

Simon, H.A. (1969). *The Sciences of the Artificial*. MIT Press.

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell.

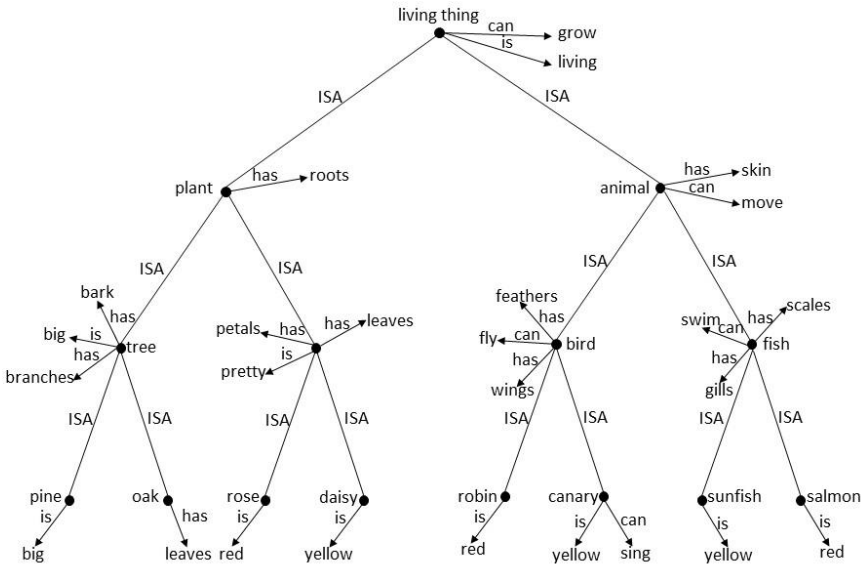


**Figure 1.** The semantic network used in Collins & Quillian (1969) and Rogers & McClelland (2004).

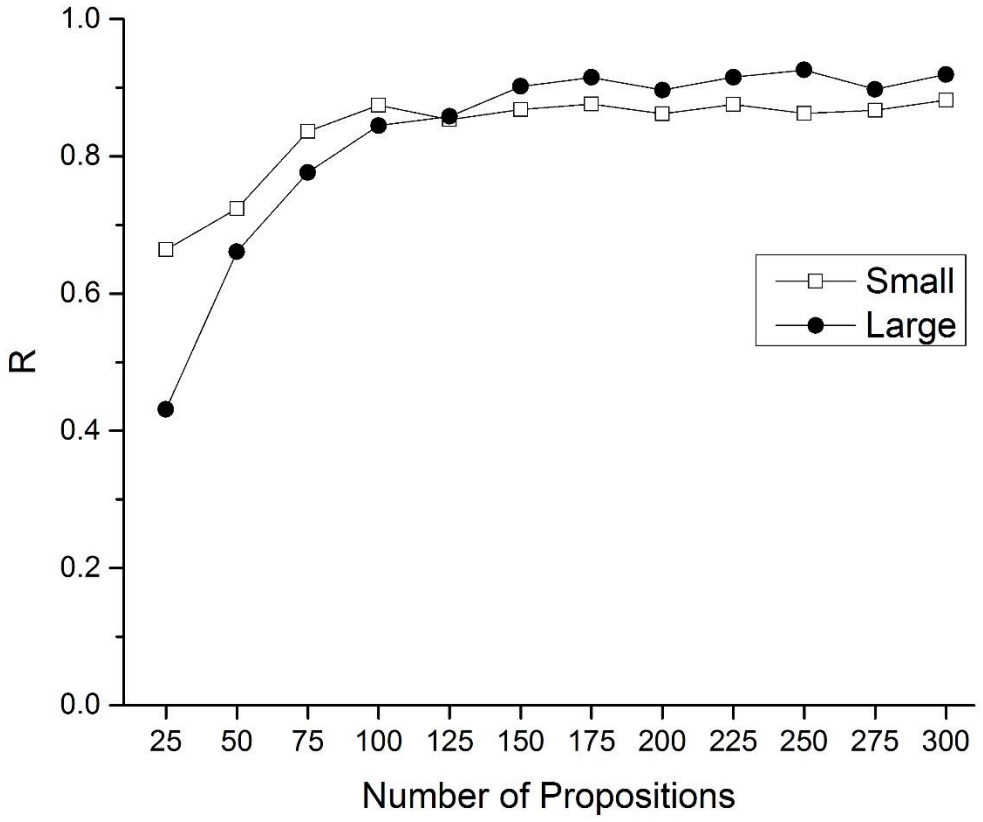


**Figure 2.** Increase in correlation between the cosine similarity of items and feature overlap values derived from the two semantic networks, as a function of the number of propositions studied.

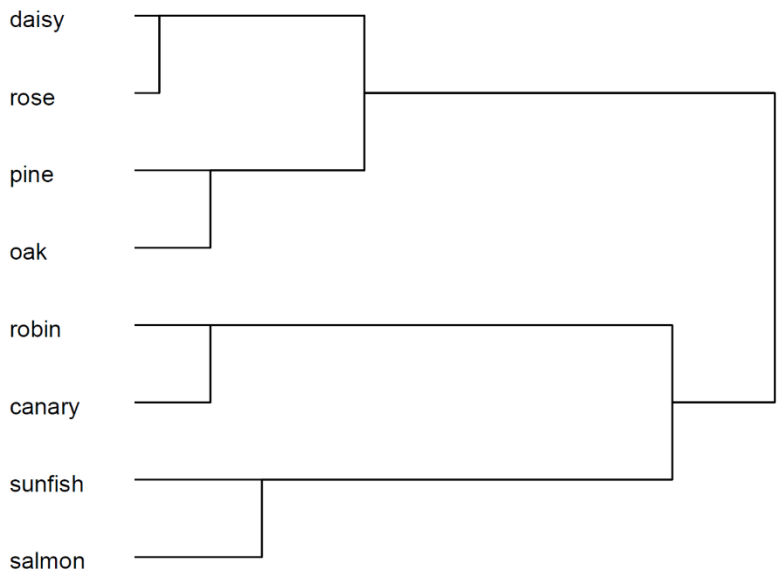


**Figure 3.** Dendogram of the hierarchical structure of Beagle trained on propositions derived from the small semantic network.

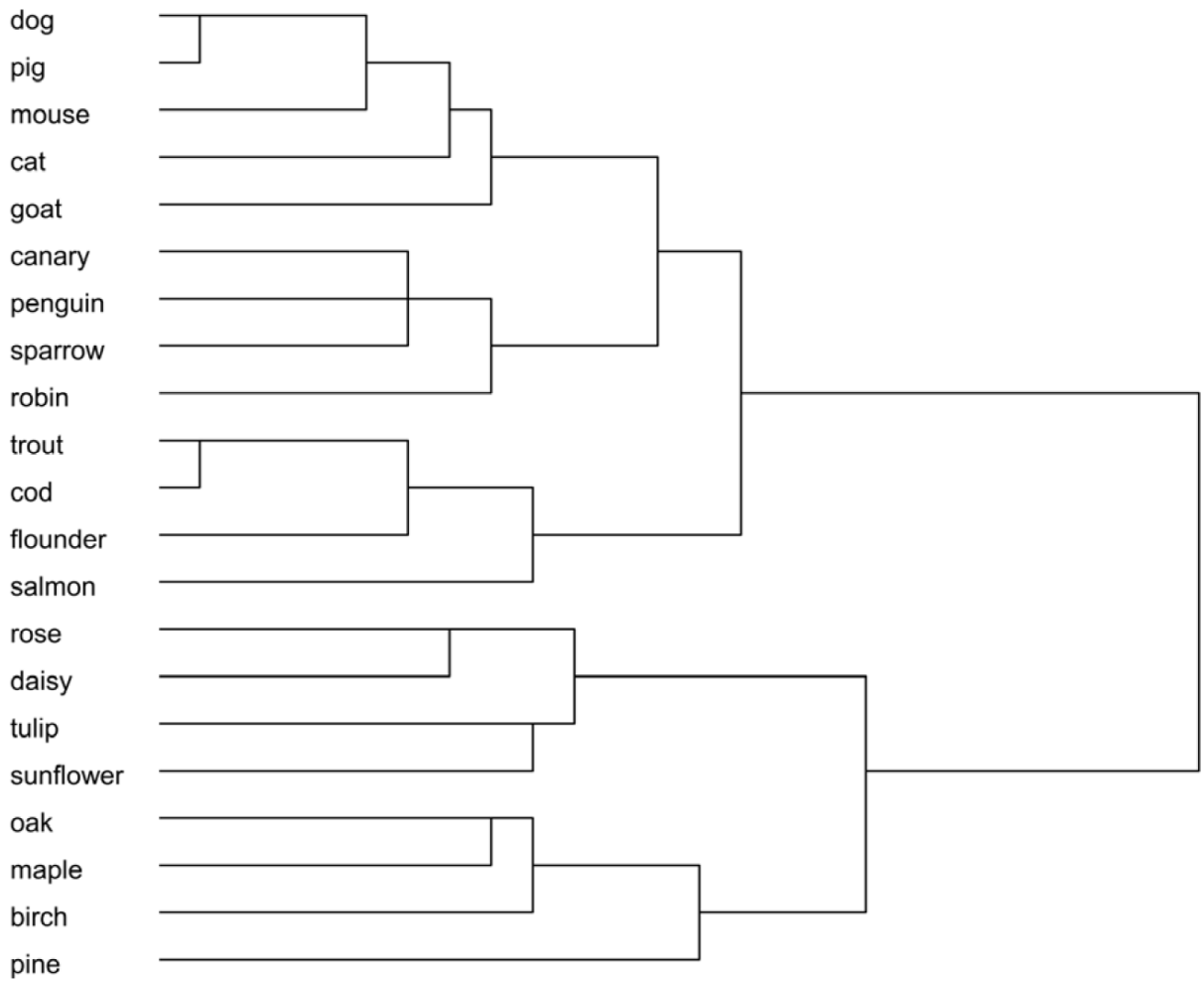


**Figure 4.** Dendogram of the Beagle model trained on propositions derived from the large semantic network.

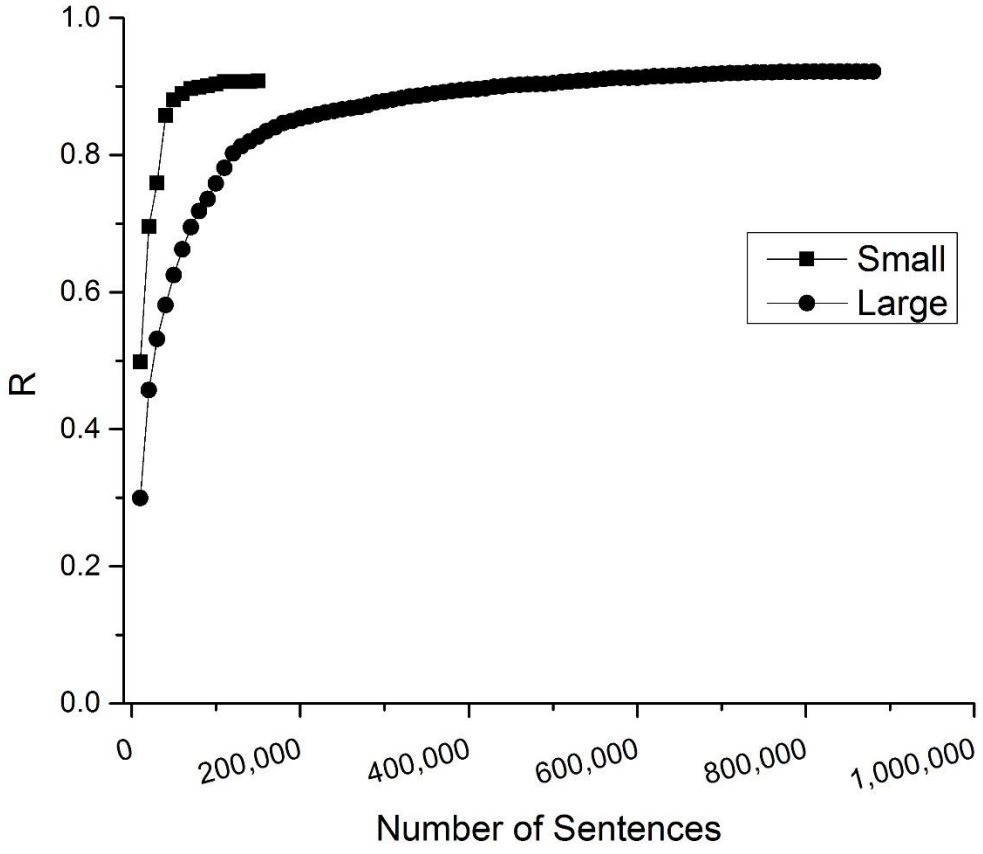


**Figure 5.** Correlation of the similarity between words and the feature overlap values from the two semantic networks, as a function of the number of sentences sampled.

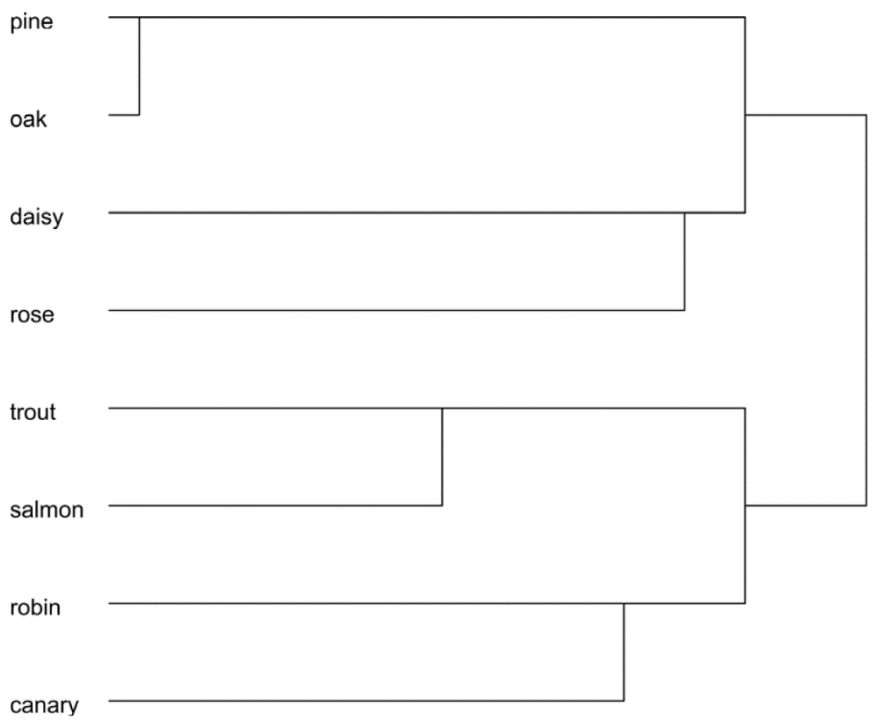


**Figure 6.** Dendogram of the hierarchical structure of the representation learned from the corpus sampling routine for words from the small network.

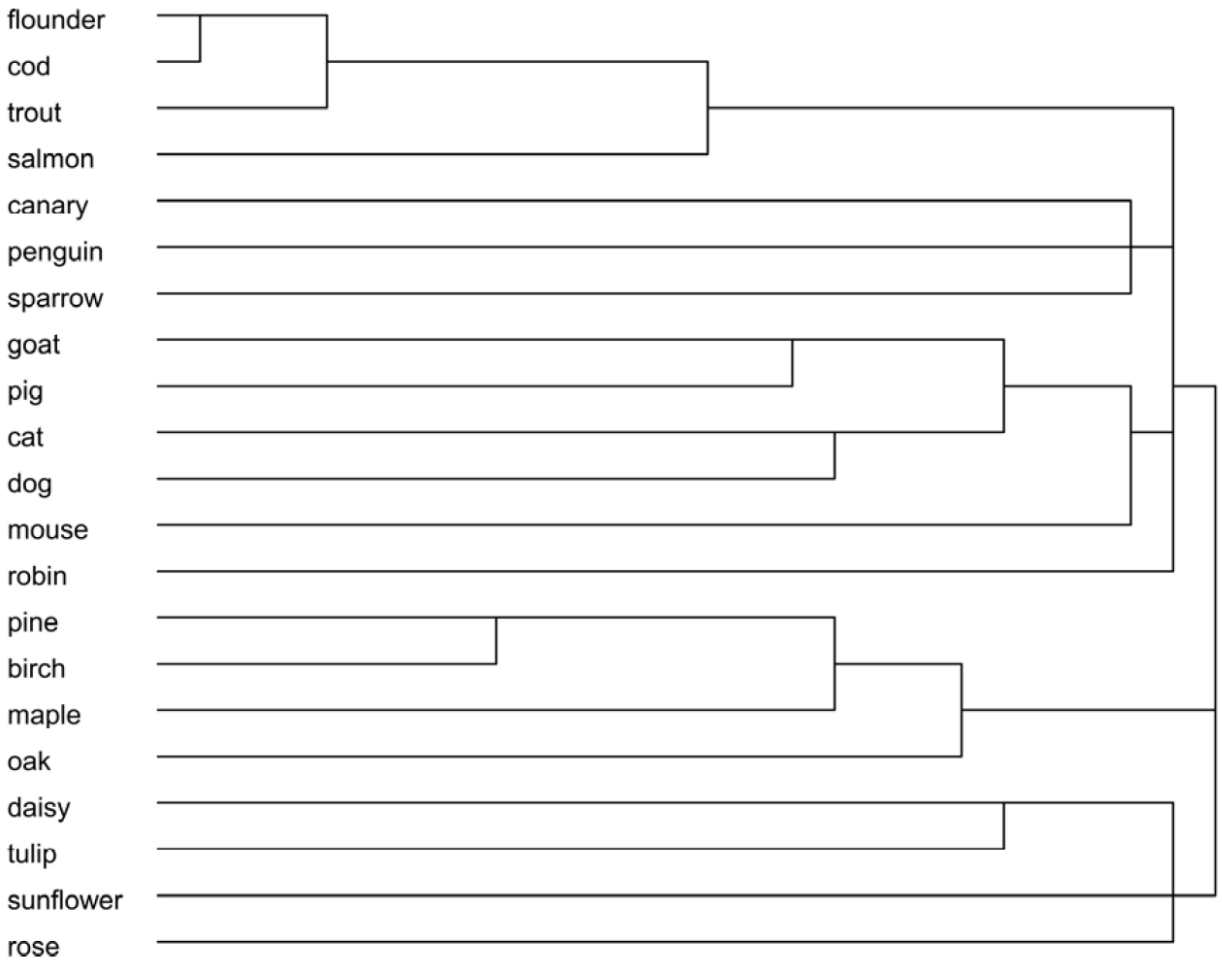


**Figure 7.** Dendogram of the hierarchical structure of the representation learned from the corpus sampling routine for words from the large network.

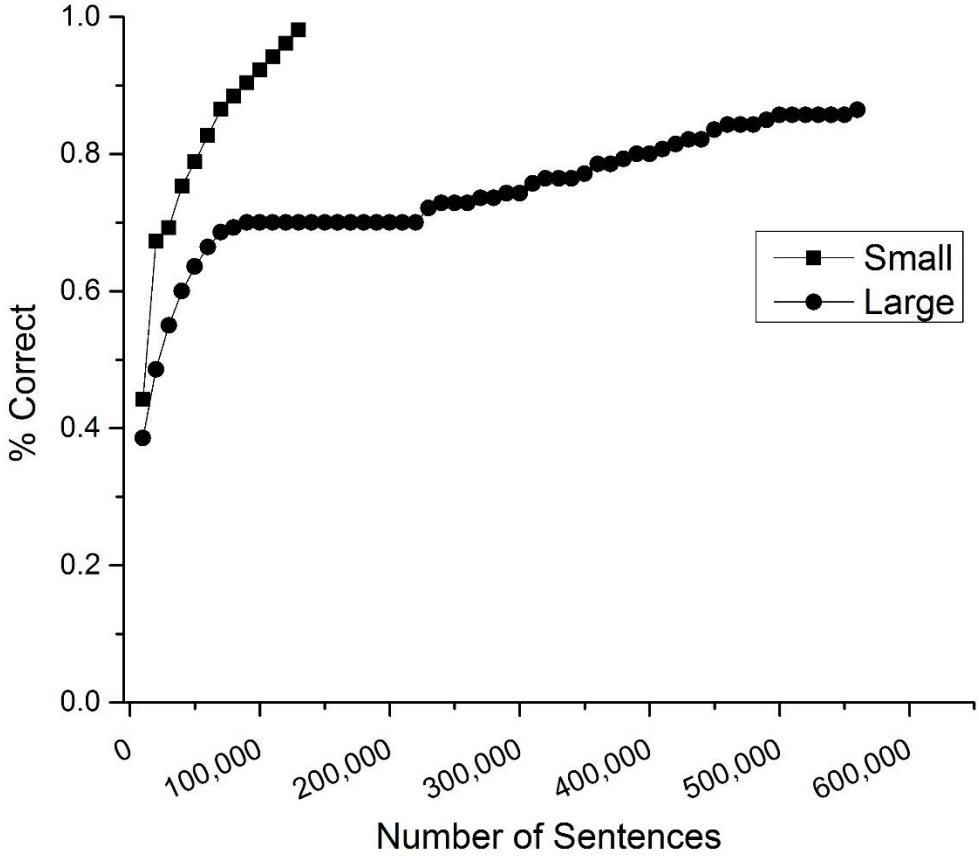


**Figure 8.** Performance on the AFC test as a function of the number of sentences sampled.