Degrees of Separation in Semantic and Syntactic Relationships

Matthew A. Kelly (matthew.kelly@psu.edu), David Reitter (reitter@psu.edu)
The Pennsylvania State University, University Park, PA

Robert L. West (robert.west@carleton.ca)
Carleton University, Ottawa, ON, Canada

Abstract

Computational models of distributional semantics can analyze a corpus to derive representations of word meanings in terms of each word's relationship to all other words in the corpus. While these models are sensitive to topic (e.g., tiger and stripes) and synonymy (e.g., soar and fly), the models have limited sensitivity to part of speech (e.g., book and shirt are both nouns). By augmenting a holographic model of semantic memory with additional levels of representations, we present evidence that sensitivity to syntax is supported by exploiting associations between words at varying degrees of separation. We find that sensitivity to associations at three degrees of separation reinforces the relationships between words that share part-of-speech and improves the ability of the model to construct grammatical sentences. Our model provides evidence that semantics and syntax exist on a continuum and emerge from a unitary cognitive system.

Keywords: semantic memory; mental lexicon; distributional semantics; word embeddings; holographic models; cognitive models; semantic space; part-of-speech; language production

Introduction

How do humans acquire, produce, and comprehend language? To what extent does language require a specialized cognitive capacity? And to what extent do humans learn language the same way that humans learn any other skill, whether it is learning to play chess or to play a piano piece?

Computational cognitive models provide a means of investigating the extent to which basic cognitive functions play a role in language. Computational models of learning and memory have been able to account for a variety of psycholinguistic phenomena without any *a priori* linguistic knowledge.

Linguistics distinguishes *lexical* knowledge (describing words) from *syntactic* processes (describing how words are combined to form sentences). We modify an existing computational model of the acquisition of lexical knowledge to enhance its ability to provide an integrated account of the acquisition of syntactic knowledge.

Our model, the Hierarchical Holographic Model (HHM), is based on BEAGLE (Jones & Mewhort, 2007). BEAGLE is a distributional semantics model that uses holographic memory (Plate, 1995). Distributional models infer the meaning of words from how the words co-occur in a corpus. BEAGLE's algorithm is not specific to language and has been applied to recognition memory (Kelly, Kwok, & West, 2015), learning a decision-making task, math cognition, and playing simple games (Rutledge-Taylor, Kelly, West, & Pyke, 2014).

Building on work by Grefenstette (1994), we define *orders of association* as a measure of the relationship between words. This notion is related to *degrees of separation*, a measure of the distance between two nodes in a connected graph.

First-order (direct) associations are useful for detecting words that are related in topic (e.g., *tiger* and *stripes*) and second-order associations are useful for detecting words that have a degree of synonymy (e.g., *tiger* and *lion*). Distributional semantics models, such as BEAGLE, are sensitive to both first and second-order associations.

Distributional models are weakly sensitive to part-of-speech (e.g., *book* and *shirt* are nouns). In the semantic space of distributional models, words tend to cluster by part-of-speech, such that, using a classifier, these models can be used for automated part-of-speech tagging (e.g., Tsuboi, 2014).

Distributional models are not, strictly speaking, sensitive to these clusters, it is the work of the classifier to detect them. While all words in a cluster will be similar to some other words in the cluster, there may be words in the cluster that are entirely dissimilar to each other. This is because similarity is not transitive. These clusters are evidence of higher-order associations that all words in the cluster have to all other words in the cluster. Thus, we propose a variant of BEAGLE that is sensitive to arbitrarily indirect associations. This allows us to explore how higher-order associations can be utilized to improve on the ability of computational models of distributional semantics to infer syntactic information from a corpus.

Our Hierarchical Holographic Model is not a model of syntax or semantics *per se*, as it does not produce or comprehend utterances. However, HHM generates representations that capture knowledge of how a word is used, what words it can be used with, and how those words should be sequenced to form a grammatical utterance. HHM's representations can be situated in and utilized by a model that operates at the utterance level (e.g., Johns, Jamieson, Crump, Jones, & Mewhort, 2016). The objective of this research is to provide a foundation for a single system account of the acquisition of semantic and syntactic lexical knowledge that is based on a general-purpose computational model of human memory.

In this paper, we explain the theory and mechanics of the Hierarchical Holographic Model and show how the model can be used to learn part of speech relations between words and to order words into grammatical sentences. In sum, we present contributions to a theory of human memory, describe a computational model based on that theory, and evaluate the model on human linguistic behavior.

Theory

In what follows, we define *orders of association* as a measure of the relationship between a pair of words in memory. We describe the BEAGLE model of distributional semantics

Table 1: Example of a third order association between *eagles* and *birds*.

Sentences			
eagles *soar over trees*	airplanes *soar* through skies	dishes are *over* plates	squirrels live in *trees*
birds *fly above forest*	airplanes *fly* through skies	dishes are *above* plates	squirrels live in *forest*

(Jones & Mewhort, 2007), based on the holographic model of memory (Plate, 1995). We then propose the Hierarchical Holographic Model (HHM), a variant of BEAGLE capable of detecting arbitrarily high orders of association.

Orders of Association

Imagine a graph where each word in the lexicon is a node connected to other words. A pair of words are connected once for each time they have occurred in the same context. In human cognition, that context is defined by the limited capacity of working memory. In our model, the context is a window of 5 to 10 words to the left and right of the target word. *Order of association* is the length of a path between two words in the graph. The *strength* of that order of association is the number of paths of that length between the two words.

First order association is when two words appear together. In the sentence "eagles soar over trees", the words *eagles* and *trees* have first order association. Words with strong first order association (i.e., frequently appear together) are often related in topic, such as the words *tiger* and *stripes*.

Second order association is when two words appear with the same words. In the sentences "airplanes soar through skies" and "airplanes fly through skies", *soar* and *fly* have second order association. Words with strong second order association are often synonyms.

Third order association is when two words appear with words that appear with the same words. Given the sentences in Table 1, the words *eagles* and *birds* have neither first nor second order association, but do have third order association.

Fourth order and higher One can keep abstracting to higher orders of association indefinitely. Eventually, all words are related to all other words in the language.

No association A pair of words with no path between them have no association of any order. For an agent that knows only the eight sentences in Table 1 as well as a ninth sentence "cars drive on streets", the words *car* and *eagle* have no association. In real language data, two words will only have no association if they belong to two different languages.

The definition of *orders of association* that we provide here is an application of the concept of *degrees of separation* in a network to words in a language, and is a generalization of Grefenstette (1994)'s first-order, second-order, and third-order affinities between words.

According to Barceló-Coblijn, Corominas-Murtra, and Gomila (2012), the point at which a child transitions from speaking in utterances of one or two words to speaking in full sentences is the point at which the child's knowledge of the relationships between words forms a dense "small world" graph, typical of an adult vocabulary, where all words are several steps from all other words in the graph. We hypothesize that learning these longer range connections between words is necessary to construct novel syntactic utterances.

To define orders of association, we have described the lexicon as a connected graph. This graph is not explicitly represented by the computational models we use. The BEAGLE model defines a space rather than a graph, where words are points in space. Words close together in BEAGLE's space have strong second-order association. Our Hierarchical Holographic Model (HHM) extends BEAGLE by defining a space for each order of association. Level 1 of HHM is BEAGLE, Level 2 represents third-order associations as distance, Level 3 represents fourth-order associations, and so on.

Previous computational models that detect third-order associations (or higher) have been clustering or classification algorithms applied to words organized in a space of second-order associations (e.g., Grefenstette, 1994; Tsuboi, 2014). Conversely, HHM recursively applies the memory and learning principles it uses to detect second order associations to detect higher order associations. As such, even at higher-orders, HHM does not produce discrete categories corresponding to noun, verb, adverb, etc., but instead produces graded representations of lexical syntactic relationships.

We expect that fourth-order associations may be sufficient to capture syntactic relationships. In a semantic network constructed from English word co-occurrence, the average minimum path length between any pair of words is between 3 and 6, depending on how the network is constructed (Steyvers & Tenenbaum, 2005). As such, we expect that by Level 3 of HHM, many words will be related to half the lexicon.

The BEAGLE Model

In the BEAGLE model of semantic memory (Jones & Mewhort, 2007), each word is represented by two vectors: an environment vector that represents the percept of a word and a memory vector that represents the concept of a word.

An environment vector (denoted by **e**) stands for what a word looks like in writing or sounds like when spoken. For simplicity, we chose not to simulate the visual or auditory features of words (but see Cox, Kachergis, Recchia, & Jones, 2011 for a version of BEAGLE that does simulate these features). Instead, we generate the environment vectors using random values, as in (Jones & Mewhort, 2007). In our simulations, environment vectors are generated by randomly sampling values from a Gaussian distribution with a mean of zero and a variance of $1/n$, where n is the dimensionality. These dimensions are meaningless, only the relationships between vectors are meaningful. The number of dimensions, *n*, determines the fidelity with which BEAGLE stores the word co-occurrence data, such that smaller *n* yields poorer encoding.

Memory vectors (denoted by $\mathbf{m}$) represent the associations a word has with other words. Memory vectors are constructed as the model reads the corpus. Memory vectors are holographic in that they use circular convolution (denoted by $*$) to compactly encode associations between words (Plate, 1995). Given a sentence, for each word in the sentence, vectors representing all sequences of words in the sentence (or grams) that include the target word are summed together and added to the target word's memory vector.

For example, given the sentence, "eagles soar over trees", we update the memory vectors for each word in the sentence: *eagles, soar, over,* and *trees*. Each memory vector is updated with a sum of grams. The memory vector for the word *soar*, $\mathbf{m}_{\text{soar}}$, is updated with the bigrams "eagles soar" and "soar over", the trigrams "eagles soar over" and "soar over trees", and the tetragram "eagles soar over trees".

Each gram is constructed as a convolution of the environment vectors of the constituent words, except for the target word, which is represented by the placeholder vector (denoted by ϕ). The placeholder vector is randomly generated and serves as a universal retrieval cue. With the placeholder substituted for the target word, each gram can be understood as a question to which the target word is the answer. So, rather than adding a representation of "eagles soar over" in $\mathbf{m}_{\text{soar}}$, we instead add "eagles ? over", i.e., "What was the word that appeared between *eagles* and *over*?". Each memory vector can be understood as the sum of all questions to which that memory vector's word is an appropriate answer.

For example, given "eagles soar over trees", we add "eagles ?", "? over", "eagles ? over", "? over trees", and "eagles ? over trees" to $\mathbf{m}_{\text{soar}}$ as follows:

$$\begin{aligned}\mathbf{m}_{\text{soar},t+1} = \mathbf{m}_{\text{soar},t} &+ \mathbf{P}_{\text{before}}(\mathbf{e}_{\text{eagles}}) * \phi + \mathbf{P}_{\text{before}}(\phi) \\ &* \mathbf{e}_{\text{over}} + \mathbf{P}_{\text{before}}(\mathbf{P}_{\text{before}}(\mathbf{e}_{\text{eagles}}) * \phi) \\ &* \mathbf{e}_{\text{over}} + \mathbf{P}_{\text{before}}(\mathbf{P}_{\text{before}}(\phi) * \mathbf{e}_{\text{over}}) * \mathbf{e}_{\text{trees}} \\ &+ \mathbf{P}_{\text{before}}(\mathbf{P}_{\text{before}}(\mathbf{P}_{\text{before}}(\mathbf{e}_{\text{eagles}}) * \phi) * \mathbf{e}_{\text{over}}) * \mathbf{e}_{\text{trees}}\end{aligned} \tag{1}$$

where t is the current time step, all vectors $\mathbf{m}$, $\mathbf{e}$, and ϕ have n dimensions, and $\mathbf{P}_{\text{before}}$ is a permutation matrix used to indicate that a word occurred earlier in the sequence. $\mathbf{P}_{\text{before}}$ is constructed by randomly permuting the rows of the n x n identity matrix. Multiplying a vector $\mathbf{v}$ by $\mathbf{P}_{\text{before}}$ results in the permuted vector $\mathbf{P}_{\text{before}}\mathbf{v}$.

While BEAGLE is a model of lexical semantics, variants of BEAGLE have been applied to non-linguistic memory and learning tasks, such as learning sequences of actions for strategic game play (Rutledge-Taylor et al., 2014). We previously proposed a variant of BEAGLE (Kelly et al., 2015) that learns sets of property-value pairs (e.g., *colour:red shape:octagon type:sign label:stop*) of the kind used by the ACT-R cognitive architecture (Anderson & Lebiere, 1998). Thus, the BEAGLE algorithm can be applied to any problem domain that can be translated into discrete symbols. This holds true for the Hierarchical Holographic Model (HHM). While we evaluate HHM in this paper in terms of its ability to account for properties of natural language, HHM is intended as a general model of learning and memory.

Hierarchical Holographic Model

The Hierarchical Holographic Model (HHM) is a series of BEAGLE models, such that the memory vectors of one model serves as the environment vectors for the next model. Level 1 is a standard BEAGLE model with randomly generated environment vectors. Once Level 1 has been run on a corpus, Level 2 is initialized with Level 1's memory vectors as its environment vectors. Level 2 is run on the corpus to generate a new set of memory vectors, which in turn are used as the environment vectors for the next level, and so on, to generate as many levels of representations as desired.

To use the memory vectors of a previous level as the environment vectors for the next, one must normalize and randomly permute the vectors (Kelly, Blostein, & Mewhort, 2013). For level $l+1$, and all words i, the environment vectors for that level are:

$$\mathbf{e}_{l+1,i} = \mathbf{P}_{\text{group}}\left(\frac{\mathbf{m}_{l,i}}{\sqrt{\mathbf{m}_{l,i} \bullet \mathbf{m}_{l,i}}}\right) \tag{2}$$

where $\mathbf{P}_{\text{group}}$ is a random permutation used to transform memory vectors into environment vectors and $\bullet$ is the dot product.

The levels in HHM are virtual mental constructs that could all be represented within a single fully distributed neural structure. There is no limit to the number of such levels that could exist in the mind, as they are not physical constructs.

The levels in HHM can be understood as the products of memory re-consolidation, the process of revisiting experiences and recording new information about those experiences. The different levels of representation are stored separately from each other in the model for the purpose of examining the differential effects of representations that encode lower and higher orders of associations. The different levels are not necessarily separate memory systems.

Experiments

In what follows, we show that the Hierarchical Holographic Model (HHM) is able to detect third-order associations using a small example data set (Experiment 1). Running HHM on a corpus of novels from Project Gutenberg, we show that sensitivity to third or fourth order associations strengthens similarity between words that are the same part of speech (Experiment 2) and improves the ability of the model to order words into grammatical sentences (Experiment 3). These results show that HHM works as intended and that higher-order associations provide useful language data.

Experiment 1: Small Example Data Set

Higher levels of the model are sensitive to higher orders of association, as demonstrated by an example data set consisting of the eight sentences in Table 1 as well as an unrelated control sentence, ”cars drive on streets”. This is a toy example chosen to provide a clear illustration of how HHM works. We believe this toy example is important because understand how HHM behaves in this example is critical to understanding how HHM behaves on real language data.

HHM was run with 1024 dimensional vectors and three levels of representations. In the nine sentences of this example, there are 21 unique words, and therefore 210 unique pairs of words. We can characterize the behavior of HHM by how the word pairs change in similarity across levels. In Figure 1, of the 210 word pairs, we graph the 24 word pairs that have non-negative similarity by Level 3. Of those 24 pairs, we label the 10 pairs with the most similarity.

The memory vectors for words with second order association, such as *soar* and *fly*, are close on Level 1 (cosine = 0.51) and closer by Level 3 (cosine = 0.67). Words *eagle* and *bird*, which have only third order association, are unrelated on Level 1 (cosine = -0.01) but are the fifth most similar word pair by Level 3 (cosine = 0.33).

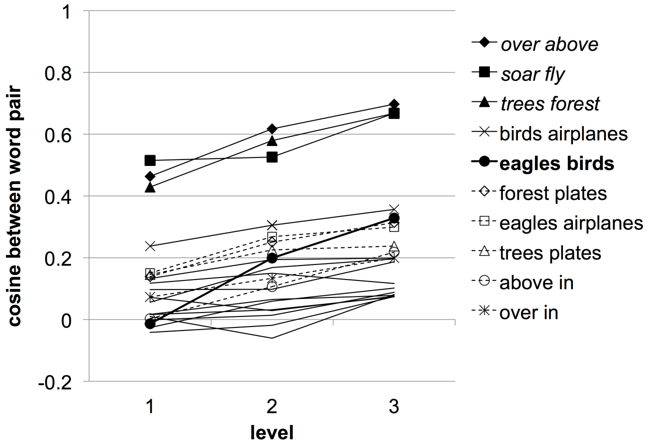

Figure 1: Cosines between word pairs across levels.

These results provide a simple example of the effect of the higher levels. Each memory vector at Level 1 is constructed as a sum of convolutions of environment vectors. As such, the memory vectors at Level 1 encode first order associations with respect to the environment vectors, measuring the frequency with which each word co-occurs with other words and sequences of words. The cosines between memory vectors are a measure of second-order association, the degree to which the two words co-occur with the same words. The algorithm that produces Level 1 transforms data that captures first-order association (co-occurrence) into data that captures second-order associations. The algorithm is a step, and by repeating it to produce higher levels, we can build a staircase.

Level 1 of the model cannot detect third-order associations. A pair of words with third-order association, but not first or second, do not appear together in the same sentence and do not co-occur with the same words. As such, the memory vectors for a pair of words with only third-order association will be constructed from disjoint sets of vectors. At Level 1, $\mathbf{m}_{1,\text{eagles}}$ is a sum of convolutions of $\mathbf{e}_{1,\text{soar}}$, $\mathbf{e}_{1,\text{over}}$, $\mathbf{e}_{1,\text{forest}}$, whereas $\mathbf{m}_{1,\text{birds}}$ is a sum of convolutions of $\mathbf{e}_{1,\text{fly}}$, $\mathbf{e}_{1,\text{above}}$, $\mathbf{e}_{1,\text{trees}}$. As Level 1 environment vectors are approximately orthogonal, the memory vectors constructed from them will also be approximately orthogonal. As a result, $\mathbf{m}_{1,\text{eagles}}$ and $\mathbf{m}_{1,\text{birds}}$ are approximately orthogonal (cosine = -0.01).

But at higher levels, the environment vectors are no longer orthogonal. Level 2 environment vectors are the Level 1 memory vectors. As a result, $\mathbf{e}_{2,\text{soar}}$ is similar to $\mathbf{e}_{2,\text{fly}}$ (cosine = 0.51), $\mathbf{e}_{2,\text{over}}$ is similar to $\mathbf{e}_{2,\text{above}}$ (cosine = 0.46), and $\mathbf{e}_{2,\text{forest}}$ is similar to $\mathbf{e}_{2,\text{trees}}$ (cosine = 0.43). Even though $\mathbf{m}_{2,\text{eagles}}$ and $\mathbf{m}_{2,\text{birds}}$ are still constructed from disjoint sets of environment vectors, because the vectors that they are constructed from are similar, $\mathbf{m}_{2,\text{eagles}}$ and $\mathbf{m}_{2,\text{birds}}$ are somewhat similar (cosine = 0.20). As the memory vectors for the pairs *soar* and *fly*, *above* and *over*, and *forest* and *trees* are more similar at Level 2 than at Level 1 (see Figure 1), the environment vectors for them will be more similar at Level 3 than Level 2, which increases the similarity between *eagles* and *birds* at Level 3 (cosine = 0.33).

Experiment 2: Part of Speech

We trained HHM on a corpus of novels from Project Gutenberg. The corpus is 10 238 600 sentences with 145 393 172 words and 39 076 unique words. HHM read the corpus one sentence at a time. Within each sentence, HHM used a moving window of 21 words, 10 words to the left and right of a target word. In that window, all grams that included the target word, from bigrams up to 21-grams, were encoded as convolutions of environment vectors and summed into the target word's memory vector. We used 1024 dimensional vectors.

Using WordNet (Princeton University, 2010) and the Moby Part-Of-Speech list (Ward, 1996), we assigned a part of speech tag to each word in the 39 076 word vocabulary. Here we use similarity between words that are the same part-of-speech (noun, verb, adjective, etc.) as a proxy measure for knowledge that those words can be used in similar ways.

To examine the effect of third-order associations, we compare Levels 1 and 2. We limit our analysis to words with at least 1000 occurrences in the corpus, as these words will have the most robust vector representations, and to word pairs that increased or decreased in similarity the most between levels.

As shown in Table 2, of the 1000 word pairs that increased the most in similarity from Level 1 to 2, 71% of those words have matching part-of-speech: 48% are partial matches (e.g., *associated* and *searching* are both verbs, but *searching* is also an adjective) and 23% are exact matches (e.g., *focused* and *emerging* can both be an adjective or a verb).

In total, 13% of all pairs of words in the lexicon are exact matches (see Table 2). Among the 1000 word pairs that increased the most from Level 1 to Level 2, there are significantly more (23%) exact matches than would be expected in a random sample from the set of all word pairs ($p < 0.0001$).

Of the 1000 word pairs that decreased in similarity the most from Level 1 to 2, only 1% are exact matches (e.g., both *local*

Table 2: Top 1000 word pairs that changed in similarity the most at each level, categorized by part-of-speech match.

Level	Change	Exact	Partial	Mismatch
total	-	*13%*	*45%*	*42%*
1 to 2	increase	23%	48%	29%
1 to 2	decrease	1%	53%	46%
2 to 3	increase	26%	44%	30%
2 to 3	decrease	0%	1%	99%

and *wizard* can be used as an adjective and a noun), which is significantly fewer than chance ($p < 0.0001$).

From Level 2 to 3, we find that 26% of the word pairs that increased in similarity the most are exact matches, which is significant ($p < 0.0001$). Of the word pairs that decreased in similarity from Level 2 to 3, zero were exact matches and only 1% were partial matches (e.g., *never* and *oh* can both be exclamations, but *never* is more commonly an adverb), which, again, was significantly less than chance ($p < 0.0001$).

In sum, we find that the sensitivity to third order (Level 2) and fourth order associations (Level 3) strengthens similarities between words with matching part of speech and weakens similarities between words with mismatching part of speech.

Experiment 3: Word Ordering Task

Do higher-order associations provide additional useful information about how to sequence words into a sentence? When given an unordered set of words that can be arranged into a sentence, are higher levels of HHM better able to find the grammatical ordering? We replicate a task from Johns et al. (2016). In this task, the model is given a set of n words from an n-word sentence that is not present in the exemplar set. The model must discern which of the n! possible word orderings is the grammatical, original ordering.

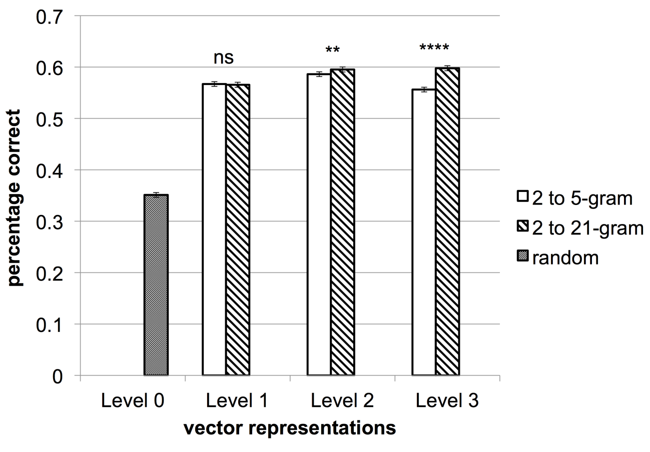

Figure 2: Percentage of test sentences correctly ordered by model as a function of vectors used to represent words.

The exemplar set consists of 125 000 seven-word sentences randomly sampled from the Project Gutenberg corpus. Sentences in the exemplar set have no words with frequency less than 300. All test set sentences and permutations thereof are excluded from the exemplar set.

We embed the word representations generated by each level of HHM in a minimal exemplar model of syntax based on Johns et al. (2016)'s work. Each sentence in the exemplar set is represented as a pair of vectors in the model. One vector is an unordered set of words constructed as a sum of the vectors representing each word in the sentence. The second vector is the ordered sequence of the words in the sentence, constructed as a holographic representation (Plate, 1995).

Test items are a set of 200 seven-word sentences taken from Johns et al. (2016). Test items have simple syntactic construction and consist of words that occur at least 300 times in the corpus. Test items are presented to the model as an unordered set of words. The model first selects the exemplar sentence most similar to the test item, as measured by cosine between the vectors for the unordered sets. Then, of the 7! possible orderings of the words in the test item, the model selects the ordering most similar to the selected exemplar sentence, as measured by the cosine between the vectors representing the ordered sequences of words. The ordering is judged correct if it matches the original ordering of the words in the test item.

HHM is trained on the full Project Gutenberg corpus. We trained HHM twice: once using a 21 word window, computing bigrams to 21-grams within that window, and once using an 11 word window, computing bigrams to 5-grams within that window. The 5-gram window is standard for the BEAGLE model. Words are represented by either random vectors (Level 0), BEAGLE memory vectors (Level 1), Level 2 memory vectors, or Level 3 memory vectors. At Levels 1, 2, and 3, we test both the 5-gram and 21-gram variants.

To ensure that results are not contingent on a particular sample of 125 000 exemplar sentences, results are averaged across 50 random samples. Mean percent correct across the 50 samples is shown in Figure 2 (Error bars indicate standard error). To test for statistical significance across the seven conditions, we used a repeated measures permutation test.

Level 0 gets a mean of 35.1% of the sentences correct using random vectors, i.e., by selecting the exemplar sentence with the most words in common with the test item.

At Level 1, we find no effect of window size ($p > 0.05$). Level 1 outperforms Level 0 ($p < 0.0001$) with a mean of 57% correct. Level 1 uses BEAGLE memory vectors, i.e., selects the exemplar sentence which has the most semantic similarity to the test item.

Level 2 outperforms Level 1 ($p < 0.0001$), demonstrating the value of third-order associations. Here we find an effect of window size ($p < 0.01$). The 21-gram window gets 59.5% correct to the 5-gram window's 58.6% correct.

At Level 3, we find the 21-gram window again outperforms the 5-gram window ($p < 0.0001$). With the 21-gram window, Level 2 and Level 3 are not significantly different (59.5% vs. 60.0%, $p > 0.05$). With the 5-gram window, Level 3 gets only

55.6% correct, significantly worse than Level 2 ($p < 0.0001$).

Our results show that for the task of ordering words into grammatical sentences, a model that uses third or fourth order associations between words outperforms a model that uses first or second order associations. Our results also show that higher levels of HHM benefit from *n*-grams larger than 5-grams (whereas 5-grams may be sufficient for BEAGLE).

Conclusions

We find that the higher levels of the Hierarchical Holographic Model (HHM) exploit higher-order associations to gain syntactic information. Sensitivity to third order (Level 2) or fourth-order associations (Level 3) reinforces relationships between words that share part-of-speech and improves the model's ability to order words into grammatical sentences.

However, we find that higher levels of HHM are more useful when using larger *n*-grams. At higher levels, HHM progressively loses the ability to make fine distinctions between small *n*-grams as the representations for the words that compose the *n*-grams become increasingly similar. For example, "she grinned" and "he smiled" may be represented by identical or nearly identical bigrams at higher levels.

At the same time, higher levels begin to be able to make use of large *n*-grams. At lower levels, large *n*-grams are unique, and thus do not provide useful information about the relationships between words. At higher levels, large *n*-grams are similar to other large *n*-grams. For example, while the 7-gram "you are as gregarious as a locust" may occur only once in a corpus, at higher levels of HHM, this 7-gram comes to resemble other 7-grams, such as "he was as strong as an ox".

Gruenenfelder, Recchia, Rubin, and Jones (2016), modeling word association norms, find that a hybrid model that uses both first and second order associations better matches human data. We note that on the word ordering task, while, on average, Levels 2 and 3 with the 21 word window produced the best results, Level 1 often correctly ordered sentences that Levels 2 or 3 got wrong. We speculate that a model that uses all three levels could outperform a model that uses only one level at a time. We hypothesize that human memory is able to use relations between concepts at varying levels of abstraction as needed to meet task demands.

The Hierarchical Holographic Model is not intended as strictly a language model but as a model of human memory with the ability to detect arbitrarily abstract associations. The present work is a proof of concept of the utility of HHM as a model and preliminary evidence that higher-order associations are relevant to understanding human cognition.

Acknowledgments

We thank Kevin D. Shabahang and D. J. K. Mewhort for the use of their BEAGLE code and Project Gutenberg corpus file. We also thank D. J. K. Mewhort for the use of his server, funded by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC: APA 318). This research has been funded by an Ontario Graduate Scholarship to M. A. Kelly, a National Science Foundation grant (SES-1528409) to D. Reitter, and a grant from Natural Sciences and Engineering Research Council of Canada to R. L. West.

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Lawrence Erlbaum Associates.

Barceló-Coblijn, L., Corominas-Murtra, B., & Gomila, A. (2012). Syntactic trees and small-world networks: syntactic development as a dynamical process. *Adaptive Behavior*, *20*(6), 427-442.

Cox, G. E., Kachergis, G., Recchia, G., & Jones, M. N. (2011). Towards a scalable holographic representation of word form. *Behavior Research Methods*, *43*, 602-615.

Grefenstette, G. (1994). Corpus-derived first, second and third-order word affinities. In *Proceedings of the sixth euralex international congress* (p. 279-290). Amsterdam, The Netherlands: Association for Computational Linguistics.

Gruenenfelder, T. M., Recchia, G., Rubin, T., & Jones, M. N. (2016). Graph-theoretic properties of networks based on word association norms: Implications for models of lexical semantic memory. *Cognitive Science*, *40*(6), 1460–1495.

Johns, B. T., Jamieson, R. K., Crump, M. J. C., Jones, M. N., & Mewhort, D. J. K. (2016). The combinatorial power of experience. In *Proceedings of the 38th annual meeting of the cognitive science society* (p. 1325-1330). Austin, TX: Cognitive Science Society.

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

Kelly, M. A., Blostein, D., & Mewhort, D. J. K. (2013). Encoding structure in holographic reduced representations. *Canadian Journal of Experimental Psychology*, *67*, 79-93.

Kelly, M. A., Kwok, K., & West, R. L. (2015). Holographic declarative memory and the fan effect: A test case for a new memory model for act-r. In *Proceedings of the 13th international conference on cognitive modeling* (p. 148-153). Groningen, the Netherlands: University of Groningen.

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*, 623-641.

Princeton University. (2010). About wordnet. *WordNet*. Retrieved from `http://wordnet.princeton.edu`

Rutledge-Taylor, M. F., Kelly, M. A., West, R. L., & Pyke, A. A. (2014). Dynamically structured holographic memory. *Biologically Inspired Cognitive Architectures*, *9*, 9-32.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*, 41–78.

Tsuboi, Y. (2014). Neural networks leverage corpus-wide information for part-of-speech tagging. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (p. 938-950). Doha, Qatar: ACL.

Ward, G. (1996). *Moby part-of-speech*. University of Sheffield. Retrieved from `http://icon.shef.ac.uk/Moby/mpos.html`