Degrees of Separation in Semantic and Syntactic Relationships

M. A. Kelly (mak582@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

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).

In HHM, semantic and syntactic knowledge are tightly integrated. This is not a new idea: Combinatory categorical grammar (Steedman & Baldridge, 2011), for example, has a transparent interface between semantics and syntax. However, we propose a new approach to examining the relationships between syntax and semantics. 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 (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

Table 1: Example of a fourth 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*

detecting arbitrarily high orders of association.

Orders of Association

Saussure (1916) defines two types of relationships between words: *paradigmatic* and *syntagmatic*. *Syntagmatic* is the relationship a word has with other words that surround it. *Paradigmatic* is when a pair of words can be substituted for each other.

Grefenstette (1994) defines first-order, second-order, and third-order affinities between words, and notes that computational language models are typically sensitive to either first-order (topic) or second-order (synonymy) affinities.

Building on Grefenstette and Saussure, we define the term *order of association* as a measure of the degree of separation of two words in an agent's language experience.

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 (i.e., have a *syntagmatic* relationship), 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 (i.e., have a *paradigmatic* relationship).

Third order association is a first order association plus a second order association. For example, *tiger* and *stripes* have a first-order association and *lion* and *tiger* have a second order association. Thus, *lion* and *stripes* have a third-order association mediated by *tiger*.

Fourth 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* do not have first, second, or third order association, but do have fourth order.

This is merely an artificial example. In natural language, *eagles* and *birds* have strong second-order association (i.e., are highly synonymous). Word pairs that have strong fourth-order association, but do not have first or second order association, are words unrelated in meaning but share part-of-speech (e.g., *focused* and *emerging* can both be used as a verb or adjective, see Table 3).

Fifth 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.

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 spaces for higher orders of association. Level 1 of HHM is BEAGLE, Level 2 represents fourth-order associations as distance, Level 3 represents sixth-order associations, and so on.

We can broaden our scope beyond language and define *order of association* as the degree of separation between any two items in an agent's life experience. If the human mind is sensitive to higher-order associations in language, we would expect that the mind is also sensitive to high-order associations in other knowledge domains.

Related Work

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, the Hierarchical Holographic Model (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.

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.

In language, children acquire first-order associations earlier in development than second-order associations (Brown & Berko, 1960; Ervin-Tripp, 1970; Nelson, 1977; Sloutsky, Yim, Yao, & Dennis, 2017). This process of learning second-order associations later has also been observed in an experimental setting. McNeill (1963) found that when participants are trained on a set of non-words and are tested with a free association task, after 20 trials of training, participants produced only first-order associations between the non-words, but by 60 trials, participants produced both first and second order associations.

Sloutsky et al. (2017) propose a neural network model that captures the early acquisition of first order associations and late acquisition of second order associations. In Sloutsky et al.'s model, learning second order associations is a slower process that operates independently of learning first order associations. Conversely, in our proposed model, HHM, learning higher order associations is a bootstrapping process that requires the lower order associations to have already been acquired.

Note that *order of association* in a language is distinct from *orders of approximation* to a language. *Orders of approximation* is a measure of how closely a probability model approximates a language as measured by the number of words that are taken into account when predicting the next word in a sequence (Shannon, 1951). Depending on the size of the context window we use, in our model, we use up to 5 or 10 preceding words to predict a word as well as up to 5 or 10 of the succeeding words. As such, our model could be described as a 5th or 10th order approximation to English. Independent of this parameter is the model's order of association. In this paper, we explore using up to sixth order associations. Order of approximation and association interact, such that higher orders of approximation (i.e., larger context windows) are more useful in a model sensitive to higher orders of association.

Human Memory

Humans do not have perfect recollection. A recollection is a reconstruction created by the interaction between the current experience (the cue) and the traces in memory of all past experiences related to the cue. This reconstruction process can be simulated as a sum of all memory traces, each trace weighted by a function of its similarity to the cue (Hintzman, 1984). This sum is referred to as an *echo*. When presented with a recall cue, distributed or connectionist models of memory inherently produce echoes (i.e., reconstructions) rather than perfect recollections (Kelly, Mewhort, & West, 2017).

Semantic and episodic memory are not distinct memory systems (Hintzman, 1986; Humphreys, Bain, & Pike, 1989; Anderson & Lebiere, 1998), but rather distinct forms of knowledge. The concepts of semantic memory emerge from the individual experiences of episodic memory. Concepts are formed through the process of reconstructing memories from experiences (Hintzman, 1986). For example, the concept you have of a person you know well is formed from the aggregate of all your experiences with that person. Concepts are vague echoes of all experiences relevant to the cue.

Episodes (i.e., the experiences of specific events) contain first-order associations. For example, if you see an eagle flying above the trees, that episode records in your memory an association between the visual stimuli *eagles* and *trees*. Likewise, if you see the sentence "eagles soar over trees", the words "eagles" and "trees" are associated in your memory of that episode.

Conversely, concepts (or echoes) contain second order associations (Kwantes, 2005). For example, given a memory model that knows only the sentences in Table 1, and the cue "airplanes", the echo will be a mix of the sentences "airplanes fly through skies" and "airplanes soar through skies". As such, the echo for "airplanes" encodes a second-order association between "soar" and "fly".

To account for how recollection reinforces or changes a memory, a standard characteristic of memory models is that the information retrieved from memory is then re-added to memory (e.g., Anderson & Lebiere, 1998; Franklin & Mewhort, 2015). That is to say, echoes retrieved from memory may themselves be stored in memory.

If concepts emerge from aggregation across stored episodes, what information emerges from aggregation across stored concepts? If current memory theory is correct, meta-concepts must arise from retrieval across concepts. These meta-concepts would be sensitive to fourth-order associations between the items of experience. Meta-concepts would be learned later in life. Just as concepts are learned once you have enough experience with a thing to form a concept of it, learning a new stable meta-concept would require sufficient experience with a related cluster of concepts to abstract across them.

The existence of meta-concepts (and thus sensitivity to third and fourth-order associations) is thus a prediction of current theory and models of human memory. Meta-concepts are difficult to investigate experimentally, as they require the learner to have a deep understanding of a knowledge domain, which cannot be instilled in a participant within the the time frame of a laboratory memory and learning experiment.

While meta-concepts and higher-order associations should

influence human behavior across all knowledge domains, in this paper we specifically investigate the influence of meta-concepts and higher-order associations in natural language. Part of speech categories such as *nouns*, *verbs*, and *adjectives*, or the more fine-grained categories used by theories of syntax (e.g., Steedman & Baldridge, 2011), may approximate meta-concepts that arise through learning a natural language. Likewise, the way humans construct sentences may be influenced by higher-order associations.

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 and 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. In BEAGLE, the 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 information from a corpus, such that smaller n yields poorer encoding.

Memory vectors (denoted by **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 **m**, **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 **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. Normalization is to ensure that each word is equally weighted at the next level. Without normalization, high-frequency words would disproportionately dominate the representations at the next level. Permutation is necessary to protect the information encoded at one level from information encoded at the next level (Gayler, 2003). Without using permutation, the different levels of information become confounded and destructively interfere with each other (Kelly, Blostein, & Mewhort, 2013). For level $l+1$, and all words i, the environment vectors for that level are:

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

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 not distinct neural structures. These levels are virtual mental constructs that could all be represented within a single fully distributed neural structure. There is no necessary 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.

Simulations and Experiments

In what follows, we demonstrate that the Hierarchical Holographic Model (HHM) works as intended and is able to detect fourth-order associations in a small artificial data set (Experiment 1). Running HHM on a corpus of novels from Johns, Jones, and Mewhort (2016), we show that sensitivity to higher-order associations strengthens the relationship between words that are the same part of speech (Experiment 2) or Combinatory Categorical Grammar (Steedman & Baldridge, 2011) type (Experiment 3). High order associations also improve the ability of the model to order words into grammatical sentences (Experiment 4). Together these experiments show that HHM can glean useful language data by using higher-order associations.

Experiment 1: Small Example on Artificial Data

Here we show that HHM is able to detect higher order associations as intended. For the purposes of providing a clear illustration of the behavior of the model, we use a small artificial data set that provides a clean example of first, second, and fourth order associations. The data set consists exclusively of the eight sentences in Table 1 as well as an unrelated control sentence, "cars drive on streets". This is merely a toy example, but useful for demonstrating how the model works.

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 thus 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 fourth order association, are unrelated on Level 1 (cosine = -0.01) but are the fifth most similar word pair by Level 3 (cosine = 0.33).

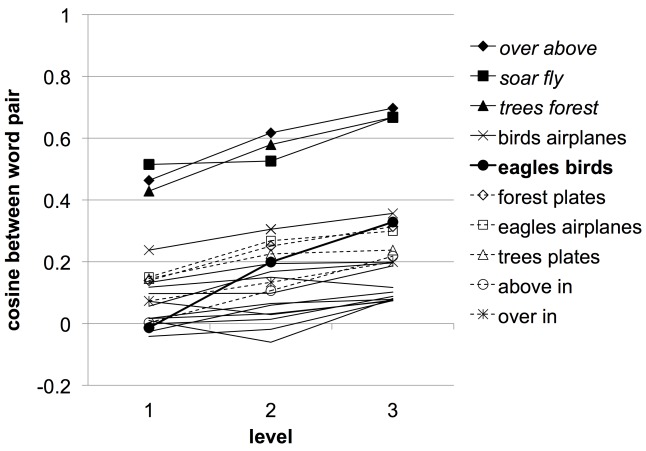

Figure 1: Cosines between word pairs across levels.

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 associations higher than second-order. 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 or higher 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. The environment vectors for Level 2 are the memory vectors for Level 1. 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 further drives up the similarity between *eagles* and *birds* (cosine = 0.33).

Training the Model

We trained HHM on a corpus of novels from Johns, Jones, and Mewhort (2016). 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 centered on a target word. Within that window, all grams that included the target word, from bigrams up to grams the width of the window, were encoded as convolutions of environment vectors and summed into the target word's memory vector. We use 1024 dimensional vectors and three levels of representations.

We experimented with four different window sizes:

1. An 11 word window (5 words to the left and right of the target word) where the model learns all 2- to 5-grams within that window,
2. An 11 word window where model learns all 2- to 11-grams within that window,
3. A 21 word window (10 words to the left and right of the target word) where the model learns all 2- to 21-grams within that window,
4. A sentence-length window, where the model learns all bigrams to sentence length grams within that window.

Experiment 2: Part of Speech

If higher-order associations are useful for knowing how a word can be appropriately used in a grammatical sentence, we should expect see that higher orders of associations enhance the sensitivity of the model to measures of how words are used. In this section, we explore correlations between HHM's representations and part of speech (noun, verb, adverb, adjective, etc.). In the next section, we examine the correlation between HHM's representations and the grammatical types proposed by Combinatory Categorical Grammar (CCG; Steedman & Baldridge, 2011).

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 as a proxy measure for knowledge that those words can be used in similar ways. Properly speaking, part of speech is a theory of language, rather than a behavioral phenomenon, and as such, a cognitive model of language use need not account for part of speech *per se* as long it can account for how humans produce and comprehend sentences. Nevertheless, looking at the relationship between the representations of HHM and traditional part of speech categories can illustrate the effect of the higher levels of the model.

To examine the effect of third and fourth 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). The top five word pairs that increased and decreased the most in similarity between pairs of levels are shown in Table 3.

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 ($p < 0.0001$).

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%

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* 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 summary, we find that the largest effects of sensitivity to third and fourth order associations at Level 2 and fifth and sixth order associations at Level 3 strengthens similarities between high-frequency words with matching part of speech and weakens similarities between high-frequency words with mismatching part of speech.

Experiment 3: Combinatory Categorical Grammar

The familiar part of speech categories (nouns, verbs, adjectives, adverbs) provides a very coarse-grained analysis of how words are used in English. Combinatory Categorical Grammar (CCG; Steedman & Baldridge, 2011) is a theory of grammar that provides a more fine-grained analysis of how words are used.

In CCG, sentences are constructed by combining words using a small number of very simple rules. The complexity of

Table 3: Top 5 word pairs that increased and decreased the most in similarity between levels.

Level 1 to 2	Δ cos	Level 2 to 3	Δ cos
increase			
focused - emerging	+0.91	beings - accord	+0.56
bursting - based	+0.91	breaths - accord	+0.54
searching - associated	+0.91	york - accord	+0.54
away - focusing	+0.90	dollars - accord	+0.52
waiting - associated	+0.90	Orleans - accord	+0.52
decrease			
driver - main	-0.36	oh - resulted	-0.18
driver - outer	-0.36	wow - consisted	-0.16
truth - outer	-0.35	oh - solemnly	-0.15
truth - main	-0.35	ah - impression	-0.15
clerk - local	-0.35	ah - realization	-0.15

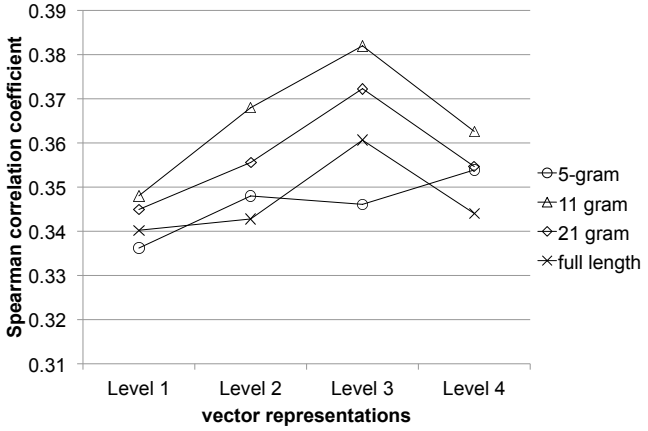

Figure 2: Spearman's rank correlation coefficient between HHM vectors and CCG types.

language arises not from the complexity of the rules, but from the complexity of the words in the language. In CCG, there are hundreds of types of words, and the type of the word determines how it can be combined with other words.

The high dimensional space of HHM provides a rich representation of how a word is used in language. As such, correlation between HHM space and CCG type may be more informative that correlation between HHM space and traditional part-of-speech categories.

To classify the words in HHM by CCG type, we use the Switchboard corpus (Godfrey, Holliman, & McDaniel, 1992). The Switchboard corpus is a collection of 2500 telephone conversations. The syntactic structure of the corpus has been annotated using CCG by Reitter, Hockenmaier, and Keller (2006). There are 10 256 unique words in the corpus. Of those words, we use the 8768 words that are also in the Novels Corpus. Just as a word can be both an adjective and a verb, a word can have multiple CCG types. To represent the CCG type profile of a word, we represent each word in the Switchboard corpus by a vector of 357 dimensions, one dimension for each CCG type in the corpus, where the value of each dimension is a count of the number of times that word is the given CCG type.

These CCG type vectors define similarity relationships between the set of 8768 words. We can compute a 8768 x 8768 similarity matrix by taking the cosine of each pair of vectors. To compare relationships in CCG space to relationships in HHM space, we also compute a 8768 x 8768 similarity matrix for each level of HHM. To measure the correlation between these two spaces, we use Spearman's rank correlation coefficient, which is nonparametric measure sensitive to non-linear relationships in data.

We compute the Spearman's correlation between the CCG cosine matrix and the cosine matrix for each level of each HHM model. As shown in Figure 2, we find that the 11-gram HHM model achieved the highest correlation with CCG types across all levels, peaking at Level 3 with a correlation of 0.382. We find that HHM's correlation to CCG is worse when the model includes up to 21-grams or full sentence grams. We also find that correlation is worse when restricting the model to 5-grams.

In summary, using correlation with CCG types as a metric, we find that higher-order associations (up to Level 3, i.e., sixth order associations) improve the ability of the model to capture part-of-speech relationships, and that large n-grams, in the range from 6-grams up to at least 11-grams, provide useful part-of-speech information.

Experiment 4: Word Ordering Task

The real test of syntactic knowledge is the ability to form grammatical sentences. Do higher-order associations provide additional useful information about how to sequence words into a grammatical 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, Jamieson, et al. (2016). In this task, a model is given an unordered set of n words taken from an n-word sentence. The model must discern which of the $n!$ possible word orderings is the grammatical, original ordering.

To perform this task, we use a simplified version of the exemplar model used by Johns, Jamieson, et al. (2016). The exemplar model is provided with an exemplar set consisting of 125 000 seven-word sentences randomly sampled from the Novels 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 the exemplar model. Each sentence in the exemplar set is represented as a pair of vectors in the exemplar 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 represen-

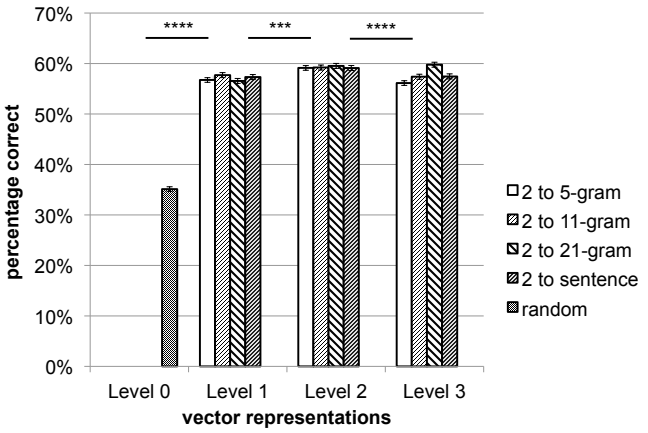

Figure 3: Percentage of test sentences correctly ordered by model as a function of vectors used to represent words.

tation (Plate, 1995).

Test items are a set of 200 seven-word sentences taken from Johns, Jamieson, 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 exemplar 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 that of the selected exemplar sentence, as measured by the cosine between the vectors representing the ordered sequences of words. The ordering produced by the model is judged to be correct if it matches the original ordering of the words in the test item.

We test performance on all versions of HHM from the previous section. 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 3 (Error bars indicate standard error). To test for statistical significance across the seven conditions, we use a repeated measures permutation test.

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

Level 1 outperforms Level 0 across all window sizes ($p < 0.0001$) with a mean of 57.1% 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 across all window sizes ($p < 0.001$) with a mean of 59.2% correct. This improvement demonstrating that third and fourth order associations contribute useful information to the task of ordering words into grammatical sentences.

At Level 3, we find a decline in performance for all models $p < 0.0001$ except the 21-gram HHM, for which performance does not change significantly from Level 2 to 3 $p > 0.05$. Here we see a significant effect of window size. The 21-gram HHM outperforms all other Level 3 models ($p < 0.0001$) and the 5-gram HHM performs worse than all other Level 3 models ($p < 0.0001$).

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 a model that uses first order or higher associations outperforms a model that only uses word overlap (i.e. Level 0).

We find little benefit to using a window beyond 5-grams, possibly because this task is restricted to constructing 7-gram sentences. However, we do find that the 5-gram HHM performs the worst at Level 3 and the 21-gram HHM performs the best. This suggests there are two counter-acting processes at work. At higher levels, HHM is increasingly able to make useful inferences about the relationships between large, low frequency n-grams, while simultaneously losing the ability to make fine discriminations between small, high frequency n-grams. We hypothesize that the decline in performance is due to all HHM models losing the ability to make these fine discriminations. Performance of HHM representations that contain larger n-grams is less affected as those models are simultaneously gaining an ability to better use those n-grams.

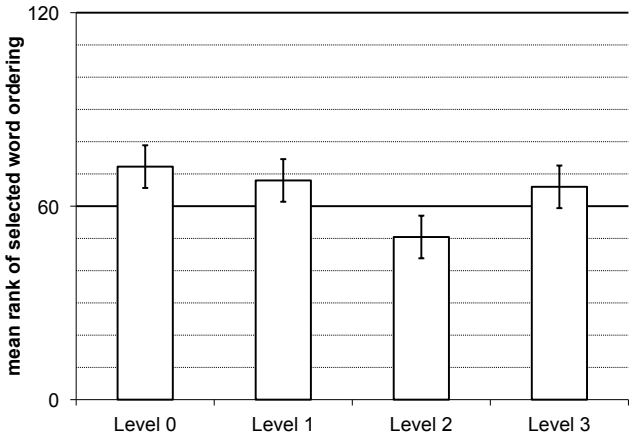

Figure 4: Mean ranking of "Colorless green ideas sleep furiously" assigned by 21-gram HHM.

Experiment 5: Colorless green ideas sleep furiously

But is the information that is gained at higher levels of HHM of a particularly *syntactic* character? Or is this additional information about indirect, abstract relationships between words better understood as semantic?

Chomsky (1956) gives the sentence "Colorless green ideas sleep furiously" as an example of a sentence that is grammatically correct but meaningless. Chomsky uses this sentence as an argument against statistical models of speech. Unless the

sentence "Colorless green ideas sleep furiously" is part of the statistical model's training corpus, a statistical model would neither be able to generate that sentence nor determine that the sentence is grammatical.

HHM is a statistical model. Can the higher levels of HHM discern that "Colorless green ideas sleep furiously" is a grammatical sentence? Given the unordered set of five words *colorless, furiously, green, ideas*, and *sleep*, there are 5! = 120 possible orderings of those words. Does HHM demonstrate a better than chance preference for Chomsky's grammatical but meaningless ordering of the words? If HHM is purely semantic and "Colorless green ideas sleep furiously" is a purely syntactic sentence, performance should be no better than chance at this task.

We use the same exemplar model as in the previous section. To construct the exemplar model's vectors, we the HHM model with a 21-gram window, as the 21-gram HHM has the most robust performance across all levels of representation in the word ordering task. The exemplar model is provided a set of 125 000 five-word sentences and picks the sentence most similar to the unordered set of words colorless, furiously, green, ideas, and *sleep*. The selected sentence's structure is then used to rank the 120 possible orderings of the words.

Mean ranking of the correct ordering by the models is shown in Figure 4. Results are averaged across 50 different random sets of 125 000 sentences. Error bars indicate standard error.

Chance performance at this task has an expected value of 60.5. To test for statistical significance, we use a repeated measures permutation test. We find that Level 0 does not rank the correct ordering as more grammatical than chance ($p > 0.05$), indicating that selecting sentences with words in common with the test set (e.g., *green*, *furiously*, etc.) is insufficient for finding the grammatical ordering. Likewise, Level 1 does not outperform chance ($p > 0.05$), indicating that semantic overlap (i.e., selecting exemplar sentences with similar meanings) is insufficient to find the grammatical ordering. However, Level 2 outperforms both chance ($p < 0.01$) and all other levels of HHM ($p < 0.05$). Level 3, however, is at chance performance ($p > 0.05$), in keeping with the drop in performance at the word-ordering task from Level 2 to Level 3 observed in the previous experiment.

This result suggests that at Level 2 of HHM, sensitivity to third and fourth order associations causes representations for words of the same part of speech to look increasingly alike, such that "Colorless green ideas sleep furiously" begins to look like a familiar, grammatical sentence.

Conclusions

We find that the higher levels of the Hierarchical Holographic Model (HHM) exploit higher-order associations to gain additional syntactic information. We find that sensitivity to third and fourth order (Level 2 of HHM) or fifth and sixth order associations (Level 3) reinforces the relationships between words that share part-of-speech and improves the ability of the model 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. 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, National Science Foundation grants (SES-1528409 and BCS-1734304) 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*. Mahwah, NJ: 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. doi: 10.1177/1059712312455439

Brown, R., & Berko, J. (1960). Word association and the acquisition of grammar. *Child Development*, *31*(1), 1-14.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, *2*, 113-124. doi: 10.1109/TIT.1956.1056813

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. doi: 10.3758/s13428-011-0125-5

Ervin-Tripp, S. M. (1970). Substitution, context, and association. In L. Postman & G. Keppel (Eds.), *Norms of word association* (p. 383 - 467). Academic Press. doi: https://doi.org/10.1016/B978-0-12-563050-4.50012-1

Franklin, D. R. J., & Mewhort, D. J. K. (2015). Memory as a hologram: An analysis of learning and recall. *Canadian Journal of Experimental Psychology*, *69*, 115-135. doi: 10.1037/cep0000035

Gayler, R. W. (2003). Vector symbolic architectures answer jackendoff's challenges for cognitive neuroscience. In P. Slezak (Ed.), *Proceedings of the joint international conference on cognitive science* (p. 133-138). Sydney, Australia: University of New South Wales.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992, Mar). Switchboard: telephone speech corpus for research and development. In *[proceedings] icassp-92: 1992 ieee international conference on acoustics, speech, and signal processing* (Vol. 1, p. 517-520 vol.1). doi: 10.1109/ICASSP.1992.225858

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. doi: 10.1111/cogs.12299

Hintzman, D. L. (1984). Minerva 2: A simulation model of human memory. *Behavior Research Methods, Instruments, and Computers*, *16*, 96-101. doi: 10.3758/BF03202365

Hintzman, D. L. (1986). "schema abstraction" in multiple-trace memory models. *Psychological Review*, *93*, 411-428. doi: 10.1037/0033-295X.95.4.528

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, *96*, 208-233. doi: 10.1037/0033-295X.96.2.208

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

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2016). Experience as a free parameter in the cognitive modeling of language. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *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. doi: 10.1037/0033-295X.114.1.1

Kelly, M. A., Blostein, D., & Mewhort, D. J. K. (2013). Encoding structure in holographic reduced representations. *Canadian Journal of Experimental Psychology*, *67*, 79-93. doi: 10.1037/a0030301

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 N. A. Taatgen, M. K. van Vugt, J. P. Borst, & K. Mehlhorn (Eds.), *Proceedings of the 13th international conference on cognitive modeling* (p. 148-153). Groningen, the Netherlands: University of Groningen. Retrieved from `http://www.iccm2015.org/proceedings/papers/0036/`

Kelly, M. A., Mewhort, D. J. K., & West, R. L. (2017). The memory tesseract: Mathematical equivalence between composite and separate storage memory models. *Journal of Mathematical Psychology*, *77*, 142-155. doi: 10.1016/j.jmp.2016.10.006

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin & Review*, *12*, 703-710. doi: 10.3758/BF03196761

McNeill, D. (1963). The origin of associations within the same grammatical class. *Journal of Verbal Learning and Verbal Behavior*, *2*(3), 250 - 262. doi: https://doi.org/10.1016/S0022-5371(63)80091-2

Nelson, K. (1977). The syntagmatic-paradigmatic shift revisited: A review of research and theory. *Psychological Bulletin*, *84*(1), 93-116.

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*, 623-641. doi: 10.1109/72.377968

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

Reitter, D., Hockenmaier, J., & Keller, F. (2006). Priming effects in combinatory categorial grammar. In *Proceedings of the 2006 conference on empirical methods in natural language processing* (pp. 308–316). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from `http://dl.acm.org/citation.cfm?id=1610075.1610119`

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. doi: 10.1016/j.bica.2014.06.001

Saussure, F. (1916). Cours de linguistique générale. In A. Sechehaye & A. Riedlinger (Eds.), *C. bally.* Lausanne, France: Payot.

Shannon, C. E. (1951). Prediction and entropy of printed english. *Bell System Technical Journal*, *30*(1), 50–64. doi: 10.1002/j.1538-7305.1951.tb01366.x

Sloutsky, V. M., Yim, H., Yao, X., & Dennis, S. (2017). An associative account of the development of word learning. *Cognitive Psychology*, *97*(Supplement C), 1 - 30. doi: https://doi.org/10.1016/j.cogpsych.2017.06.001

Steedman, M., & Baldridge, J. (2011). Combinatory categorial grammar. In R. Borsley & K. Borjars (Eds.), *Non-transformational syntax: Formal and explicit models of grammar* (p. 181-224). Wiley-Blackwell.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78. doi: 10.1207/s15516709cog2901_3

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: Association for Computational Linguistics.

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