Indirect Associations in Learning Semantic and Syntactic Lexical Relationships

M. A. Kelly[a,b,*], Moojan Ghafurian[a,c], Robert L. West[d], David Reitter[a,e]

[a] *The Pennsylvania State University, University Park, PA*
[b] *Bucknell University, Lewisburg, PA*
[c] *University of Waterloo, Waterloo, ON, Canada*
[d] *Carleton University, Ottawa, ON, Canada*
[e] *Google Research, New York City, NY*

Abstract

Computational models of distributional semantics (a.k.a. word embeddings) represent a word's meaning in terms of its relationships with all other words. We examine what grammatical information is encoded in distributional models and investigate the role of indirect associations. Distributional models are sensitive to associations between words at one degree of separation, such as 'tiger' and 'stripes', or two degrees of separation, such as 'soar' and 'fly'. By recursively adding higher levels of representations to a computational, holographic model of semantic memory, we construct a distributional model sensitive to associations between words at arbitrary degrees of separation. We find that word associations at four degrees of separation increase the similarity assigned by the model to English words that share part-of-speech or syntactic type. Word associations at four degrees of separation also improve the ability of the model to construct grammatical English sentences. Our model proposes that human memory uses indirect associations to learn part-of-speech and that the basic associative mechanisms of memory and learning support knowledge of both semantics and grammatical structure.

Keywords: distributional semantics, semantic memory, word embeddings,

*Corresponding author at: Department of Computer Science, Bucknell University, Dana 336, Lewisburg, PA 17837, USA.

Email addresses: m.alex.kelly@bucknell.edu (M. A. Kelly), moojan@uwaterloo.ca (Moojan Ghafurian), robert.west@carleton.ca (Robert L. West), reitter@google.com (David Reitter)

mental lexicon, holographic models, mediated associations

1. Introduction

Syntax (how words are put together) and semantics (what words mean) have traditionally been understood as arising from distinct cognitive processes. The distinction between syntax and semantics was famously illustrated by Chomsky (1956) with the example "Colorless green ideas sleep furiously", a sentence that is grammatical but meaningless.

But can syntax and semantics be understood as arising from a unitary cognitive process? Predictive neural language models (e.g., Ororbia II et al., 2017) appear to be sensitive to both syntax and semantics. Recurrent neural networks are able to make judgements about subject-verb agreement in nonsensical sentences such as "Colorless green ideas sleep furiously" without needing to rely on part-of-speech tagging or other syntactic markers (Gulordava et al., 2018). However, due to the "black box" nature of neural network models, it is difficult to say exactly what information is being exploited by the networks to make decisions about syntax.

Even though the nonsense sentence "Colorless green ideas sleep furiously" has a set of word transitions that do not appear in English language corpora, the sentence has a very common English construction: *adjective*, *adjective*, *noun*, *verb*, *adverb*. How do humans learn that, at an abstract level, the sentence is structurally similar to many other sentences in their life experience?

Jenkins (1964, 1965) and Jenkins and Palermo (1964) hypothesize that knowledge of the syntactic structure of language depends on indirect or *mediated* associations. More specifically, part-of-speech, or the knowledge that nouns can be substituted for other nouns and verbs for other verbs, and so on, depends on learning equivalence classes through mediated association. Although Jenkins (1968, 1974) ultimately abandoned the paradigm of understanding language and memory in terms of associations and equivalence classes altogether, more recent studies with children have found that exploiting equivalence classes is a powerful pedagogical technique for rapidly expanding a learner's language abilities (Sidman, 2009).

To explore the hypothesis that learning the part-of-speech of words is based on a capacity for indirect or mediated association, we propose a "deep" distributional semantics model, the *Hierarchical Holographic Model* (HHM). HHM consists of a stack of holographic vector models that feed one into the

next, which allows HHM to detect arbitrarily indirect associations between words. HHM is based on BEAGLE (Jones et al., 2006; Jones and Mewhort, 2007), one of the few distributional semantics models sensitive to the order of words in sentences, a critical part of English syntax.

Holographic models of human memory have a long history (Murdock, 1982; Pribram, 1969) and have been applied to a wide range of paradigms (e.g., Eliasmith, 2013; Franklin and Mewhort, 2015; Jamieson and Mewhort, 2011). Holographic vectors allow for easy implementation of a recursive model capable of learning arbitrarily indirect associations. Our approach can be understood as an extension of Jenkins and Palermo (1964)'s work, though instead of using artificial grammar experiments, we use a computational approach applied to an English-language corpus.

In what follows, we provide theoretical background on the Hierarchical Holographic Model and then evaluate the model. We give a proof-of-concept demonstration of HHM on a small artificial dataset and then train HHM on an English-language corpus. We analyze the relationship between the representations produced by the higher levels of HHM and part-of-speech (e.g., nouns, adjectives, adverbs, etc.) and the syntactic types proposed by Combinatory Categorical Grammar (CCG; Steedman and Baldridge, 2011). We show that HHM's representations can be used to order words into grammatical sentences and we test HHM on the sentence "Colorless green ideas sleep furiously". HHM is an account of the mental lexicon based on a general-purpose computational model of human memory. HHM demonstrates how a single system can incorporate knowledge of both how a word is used (i.e., part-of-speech) and what a word means (i.e., distributional semantics).

2. Theory

In this section, we describe the BEAGLE model of distributional semantics (Jones and Mewhort, 2007), based on the holographic model of memory (Plate, 1995). We propose the Hierarchical Holographic Model (HHM). HHM is a recursively constructed variant of BEAGLE capable of detecting arbitrarily high orders of association. We then define *orders of association* as a measure of the relationship between a pair of words in memory.

2.1. The BEAGLE Model

The BEAGLE model (Jones and Mewhort, 2007) belongs to the family of distributional semantics models, also known as *word embeddings*. Distri-

butional models include Latent Semantic Analysis (Landauer and Dumais, 1997), the Hyperspace Analogue to Language (Burgess and Lund, 1997), the Topics Model (Griffiths et al., 2007), *word2vec* (Mikolov et al., 2013), GloVe (Pennington et al., 2014), as well as word embeddings extracted from neural language models such as BERT (Devlin et al., 2019). Distributional models use the word co-occurrence statistics of a large corpus to construct high-dimensional vectors that represent the meanings of words. Each vector can be understood as a point in a high-dimensional space and distance in the space serves as a measure of similarity in meaning. Words that are closer together have more similar meanings. Such a space, where distance measures similarity in meaning, is referred to as a *semantic space*.

In BEAGLE, each word is represented by two vectors: an *environment vector* that represents the percept of a word (i.e., the word's perceptual features) and a *memory vector* that represents the concept of a word (i.e., the word's meaning and associations).

An environment vector (denoted by $\mathbf{e}$) stands for what a word looks like in writing or sounds like when spoken. For simplicity, we do not simulate the visual or auditory features of words (but see Cox et al., 2011, for a version of BEAGLE that does simulate features). Instead, we generate the environment vectors using random values, as in Jones and Mewhort (2007). Thus, in our simulations, words with similar morphology (e.g., *walk* and *walked*) have dissimilar environment vectors, such that the model needs to learn from the corpus that the two words are related.

Environment vectors are generated by randomly sampling values from a Gaussian distribution with a mean of zero and a variance of $1/d$, where d is the dimensionality. Individually, the dimensions of the vectors have no inherent meaning: they do not stand for specific words or features. A word is represented as a pattern of values across all dimensions. The number of dimensions, d, determines the fidelity with which BEAGLE stores word co-occurrence information, such that smaller d yields poorer encoding.

Memory vectors (denoted by $\mathbf{m}$) represent the associations a word has with other words. As the model reads the corpus, memory vectors are continuously updated. For example, the words *walk* and *walked* are represented by dissimilar, randomly-generated environment vectors. But, because the words are used in similar ways, *walk* and *walked* develop highly similar memory vectors. That said, the two memory vectors will not be identical, as *walked* is more likely to appear in contexts with other past-tense verbs and *walk* with other present-tense verbs (e.g., "I *walked* to the store and *bought* bread" vs.

"I *walk* to the store and *buy* bread.").

BEAGLE stores two kinds of information in a memory vector: *context* and *order*. The *context* information for a target word in a sentence is the sum of the environment vectors of the other words in the sentence. Conversely, the *order* information for a word in a sentence is a sum of sequences of words that include the target word. A sequence of words is represented by a vector that is a *convolution* of the environment vectors of the words in the sequence.

2.1.1. Order information

The memory vectors are termed *holographic* because they use circular convolution to compactly encode associations between words (Plate, 1995). According to holographic theories of memory (Eliasmith, 2013; Murdock, 1982; Pribram, 1969), patterns of neural activity in the brain interfere to create new associations in a manner mathematically analogous to how light waves interfere to create a hologram (Gabor, 1969). Given two patterns of neural activity represented as vectors, the interference pattern produced by the association of the two is computed as the *convolution* of the vectors.

To compute the *order* information for a target word, a sum of n-grams is added to the target word's memory vector. The n-grams are at minimum bigrams consisting of the target word and the word immediately preceding or following. The n-grams also have a maximum size that can be set. Jones and Mewhort (2007) use a maximum of 7-grams. We experiment with maximum n-gram sizes ranging from 5-grams to the full length of the sentence.

For example, given the sentence, "eagles soar over trees", BEAGLE updates the memory vectors for each word in the sentence: *eagles, soar, over,* and *trees*. For *soar*, the following n-grams are added into the memory vector $\mathbf{m}_{\text{soar}}$: the bigrams "eagles soar" and "soar over", the trigrams "eagles soar over" and "soar over trees", and the tetragram "eagles soar over trees".

Each n-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 $\mathbf{\Phi}$). The placeholder vector is randomly generated and serves as a universal retrieval cue. With the placeholder substituted for the target word, each n-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" into $\mathbf{m}_{\text{soar}}$, 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.

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

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

where $*$ is circular convolution, t is the current time step, all vectors $\mathbf{m}$, $\mathbf{e}$, and $\mathbf{\Phi}$ have d dimensions, and $\mathbf{P}_{\text{before}}$ is a permutation matrix used to indicate that a word occurred earlier in the sequence (see Appendix for discussion). $\mathbf{P}_{\text{before}}$ is made by randomly reordering the rows of the d x d identity matrix. Multiplying a vector $\mathbf{v}$ by $\mathbf{P}_{\text{before}}$ results in the permuted vector $\mathbf{P}_{\text{before}}\mathbf{v}$.

2.1.2. Context information

Context information is a sum of environment vectors. For example, the context information for $\mathbf{m}_{\text{soar}}$ and the sentence "Eagles soar over trees" is:

$$\mathbf{m}_{\text{soar},t+1} = \mathbf{m}_{\text{soar},t} + \mathbf{e}_{\text{eagles}} + \mathbf{e}_{\text{over}} + \mathbf{e}_{\text{trees}} \tag{1}$$

For the purposes of the simulations reported in this paper, we only use the *order* information and exclude the *context* information, as we found little benefit to including *context* information in the word ordering task that we use to evaluate the models.

2.1.3. Applications of BEAGLE

BEAGLE can model semantic priming (Jones et al., 2006), the pattern of semantic memory deficits in Alzheimer's disease (Johns et al., 2013), as well as basic memory phenomena, such as release from proactive interference (Mewhort et al., 2018).

While BEAGLE is a model of the mental lexicon, Dynamically Structured Holographic Memory (Rutledge-Taylor et al., 2014) is a variant of BEAGLE applied to non-linguistic memory and learning tasks, such as learning sequences of actions for strategic game play. Kelly et al. (2015) and Kelly and Reitter (2017) propose another BEAGLE variant, Holographic Declarative Memory, that learns sets of property-value pairs (e.g., *colour:red*

shape:octagon type:sign) of the kind used by the ACT-R cognitive architecture (Anderson, 2009), showing that BEAGLE's algorithm can be applied to any problem domain that can be expressed in discrete symbols.

The Hierarchical Holographic Model (HHM) can, like BEAGLE, be applied to a wide range of problem domains. 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.

2.2. Hierarchical Holographic Model

The Hierarchical Holographic Model (HHM) is a series of BEAGLE-like models, such that the memory vectors of one model serve as the environment vectors for the next model. Level 1 is a standard BEAGLE model with randomly generated environment vectors, except that we only use order information to construct the memory vectors. Level 2 and higher are order-only BEAGLE models where the environment vectors are the memory vectors of the previous level. Once Level 1 has been run on a corpus, Level 2 is initialized with Level 1's memory vectors as its environment vectors. Then 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. Vectors are normalized to unit Euclidean length 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 et al., 2013). The destructive interference arises because convolution distributes over addition. If we convolve a memory vector with another vector, that vector will distribute across all of the component n-gram vectors that are summed into the memory vector. If the other vector is also a memory vector, all of its n-grams will distribute across all of the memory vector's n-grams to create a multitude of spurious n-gram representations.

Thus, to transform memory vectors to environment vectors, the elements of all memory vectors are re-ordered according to a randomly generated per-

mutation, $\mathbf{P}_{\text{group}}$. 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 $\mathbf{e}$ and $\mathbf{m}$ are environment and memory vectors and $\bullet$ is the dot product.

The levels in HHM can be understood as the products of memory reconsolidation, 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, but instead could constitute different kinds of knowledge within a single memory system.

2.3. Orders of Association

Saussure (1916) defines two types of relationships between words: *paradigmatic* and *syntagmatic*. *Syntagmatic* describes a relationship a word has with other words that surround it. *Paradigmatic* describes a relationship in which 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. Grefenstette defines third-order affinities as semantic groupings among similar words, which can be discovered using cluster analysis techniques.

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. *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.

A pair of words are connected once for each time they have occurred in the same context. In human cognition, the context is defined by the associations in mind at the time of encoding. Ideally, we would use a model of memory to determine when words are or are not in the same context (see §5.2 for discussion). However, for simplicity, we use a context that is a window of five or more words to the left and right of the target word.

First-order association describes two words that 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 describes two words that appear with the same words. Given "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 (i.e., a paradigmatic relationship plus a syntagmatic relationship). 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*.

Statistical smoothing algorithms use third-order associations to estimate the acceptability of novel bigrams (Pereira, 2000; Roberts and Chater, 2008). For example, *unsightly bumbershoot* is a perfectly acceptable adjective-noun pair, but is unlikely to appear in a corpus that doesn't include this paper. But an *unsightly bumbershoot* is very similar to an *unsightly umbrella*. The third-order association between *unsightly* and *bumbershoot* mediated by *umbrella* can be used to judge that *unsightly bumbershoot* is an acceptable bigram.

Fourth-order association describes two words that appear with words that appear with the same words. A fourth-order association is two second-order (or paradigmatic) associations added together.

The sentences in Table 1 provide an artificial example of a fourth-order association. Words with fourth-order association are indicated in **bold** and words with second-order association are indicated in *italics*. The word pairs *soar* and *fly*, *over* and *above*, and *trees* and *forest* each have second-order associations. Given only the sentences in Table 1, the words *eagles* and *birds* do not have first-, second-, or third-order association, but do have fourth-order. The web of associations between the words in Table 1's sentences is illustrated in Figure 1.

Table 1 is an artificial example. In natural language, *eagles* and *birds* have strong second-order association (i.e., are highly synonymous). Fourth-order association indicates that two words can be substituted for each other, but at a more abstract level than second-order association (synonymy). We hypothesize that word pairs that have strong fourth-order association, but do not have first- or second-order association, are words unrelated in meaning

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

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

but are grammatically acceptable to substitute for each other. We expect that words with fourth-order association are likely to share part-of-speech or syntactic type (e.g., *focused* and *emerging* can both be used as a verb or adjective, see Table 2). We explore this hypothesis in Sections 3.3 and 3.4.

Fifth-order and higher associations can be obtained by abstracting indefinitely. Eventually, all words are related to all other words in the language.

Even-numbered associations are paradigmatic or *super-paradigmatic* relationships that indicate a semantically valid or, we hypothesize, syntactically valid substitution.

Odd-numbered associations are syntagmatic or *super-syntagmatic* relationships describing the association between a word and other words that could appear either with the word directly (first-order) or with another word like it (third-order, fifth-, etc.).

No association describes a pair of words that have no path between them of any length. For an agent that knows only the nine sentences in Table 1, 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 (e.g., the words *goyang-i* from Korean and *borroka* from Basque have no association with each other).

In our description of *orders of association* we have glossed over the question of the distinct nature of syntagmatic versus paradigmatic associations. For two words to have a syntagmatic association, it is sufficient for the words to co-occur in any way. Conversely, for paradigmatic associations, the two words should be interchangeable for each other, which is contingent on position in the sentence or phrase.

HHM, as implemented in this paper, is specifically a model of super-paradigmatic associations between words. Examining super-syntagmatic as-

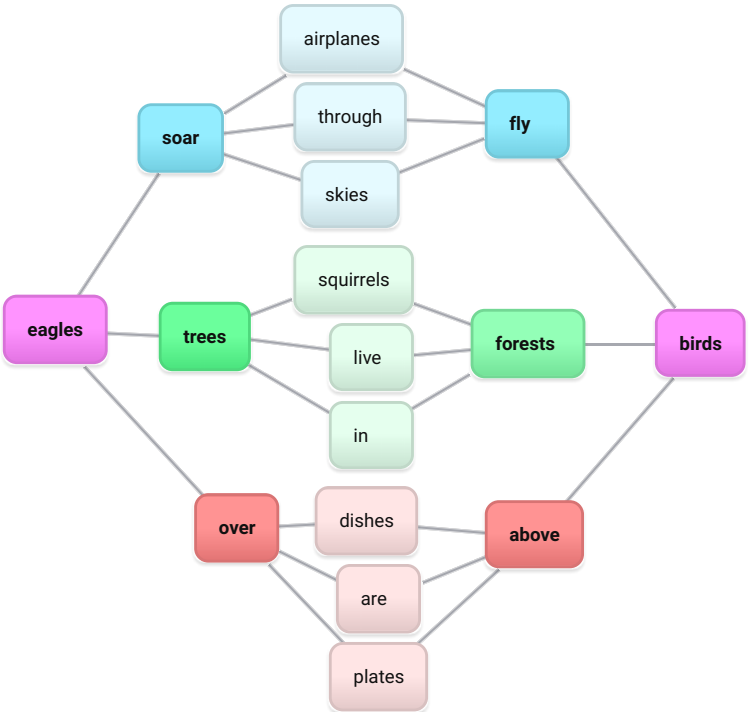


Figure 1: Web of associations between words in Table 1.

sociations is beyond the scope of our work, as our interest is in part-of-speech and syntactic type relationships, which are valid substitution relationships, rather than co-occurrence (or syntagmatic) relationships. For the purposes of this paper, we only use *order* vectors in HHM. However, we have found that odd-numbered orders of association are captured by recursively constructing levels of representation using *context* vectors.

To define orders of association, we have described the lexicon as a connected graph. This graph is not explicitly represented by HHM. HHM defines a semantic space rather than a graph. Words close together at Level 1 of HHM have strong second-order association, Level 2 represents fourth-order associations, Level 3 represents sixth-order associations, and so on.

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 HHM context window, we use up to 5, 10, or k preceding words to predict a word as well as up to 5, 10, or k of the succeeding words, where k is the length of the sentence. As such, HHM could be described as a 5th, 10th, or kth order approximation to English. Independent of this parameter is the order of association. In this paper, we explore using up to eighth-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.

3. Simulations and Experiments

We test two hypotheses:

1. Level 2 (fourth-order associations) or higher levels of the Hierarchical Holographic Model (HHM) significantly outperform Level 1 (second-order associations) on tests of correlates of syntactic knowledge.
2. Whereas second-order associations are semantic in character, fourth-order associations or higher provide knowledge that is primarily part-of-speech or a word's syntactic type.

We contrast the two hypotheses with two alternatives:

1. Fourth-order associations or higher do not improve performance on tests of correlates of syntactic knowledge.
2. Fourth-order associations or higher merely provide additional lexical semantic knowledge, such that given more data, a model sensitive only to second-order associations would discover the same word relationships.

To test these hypotheses, we begin by validating HHM as a model of orders of association. We show that HHM works as intended and is able to detect fourth-order associations in a small artificial data set (Section 3.1).

To demonstrate that higher-order associations are lexical syntactic in character, we investigate the relationship between higher-order associations and part-of-speech (Experiment 1).

However, part-of-speech provides only a coarse-grained analysis of the types of words in English. Conversely, Combinatory Categorical Grammar (CCG; Steedman and Baldridge, 2011) postulates hundreds of different word types. In CCG, a word type captures what types of phrases the word may

combine with to the left or to the right (and the associated semantic operations). Thus grammatical information is stored along with the word in the lexicon, providing fine-grained information about how each word is used. The theory proposes a very limited set of syntactic and semantic operations in parsing and sentence production that is parameterized for the specific language. CCG is a broad-coverage formalism that allows us to study the granularity of grammatical information that might be represented in the vectors generated by higher-order associations (Experiment 2).

While comparisons between HHM, part-of-speech, and CCG types are illuminating, part-of-speech and CCG are theories of language, not language itself. To evaluate the role of higher-order associations in producing grammatical sentences, we situate HHM's word representations in a simple exemplar model that operates on sentences. We use a word ordering task where the exemplar model must order a given set of words into a grammatical sentence. By varying the level of HHM used by the exemplar model, we investigate the effect of higher-orders of association and n-gram size on the ability of the model to find the grammatical ordering of the words (Experiment 3).

Chomsky (1956) famously gave "Colorless green ideas sleep furiously" as an example of a sentence that is grammatical but meaningless. If the sentence is truly meaningless we would expect that second-order (semantic) associations would be insufficient for finding the grammatical ordering of the words *colorless*, *furiously*, *green*, *ideas*, and *sleep*. However, if fourth-order associations are syntactic in character, we should expect to find that the exemplar model can find the grammatical ordering of the words using representations from Level 2 of HHM (Experiment 4).

Through these simulations and experiments, we seek to demonstrate the validity of HHM as a model, HHM's relationship to established theories of syntax, and the role of higher-order associations in constructing grammatical sentences. Code for running HHM[1][2] and the exemplar model is available online, along with data and figures[3].

3.1. 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

[1] `https://github.com/ecphory/BEAGLE-HHM`

[2] `https://github.com/moojan/Python-BEAGLE-HHM`

[3] `https://github.com/ecphory/Indirect-Associations`

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 sentences in Table 1. This is merely a toy example, but useful for demonstrating how the model works. This example has been designed such that the word pairs *soar* and *fly*, *over* and *above*, and *trees* and *forest*, have second-order associations, whereas the word pair *eagles* and *birds*, have a fourth-order association.

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.

Figure 2 shows cosine similarity between the word pairs as a function of level of representation in HHM. 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 and rank the 10 pairs with the most similarity, from *over above* (cosine = 0.70 at Level 3) to *over in* (cosine = 0.20 at Level 3). Word pairs with fourth-order association are in **bold** and word pairs with strong second-order association are in *italics*.

The memory vectors for words with second-order association are close on Level 1 (e.g., *soar* and *fly*, cosine = 0.51) and closer by Level 3 (cosine = 0.67). The 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).

The 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 or higher, 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-

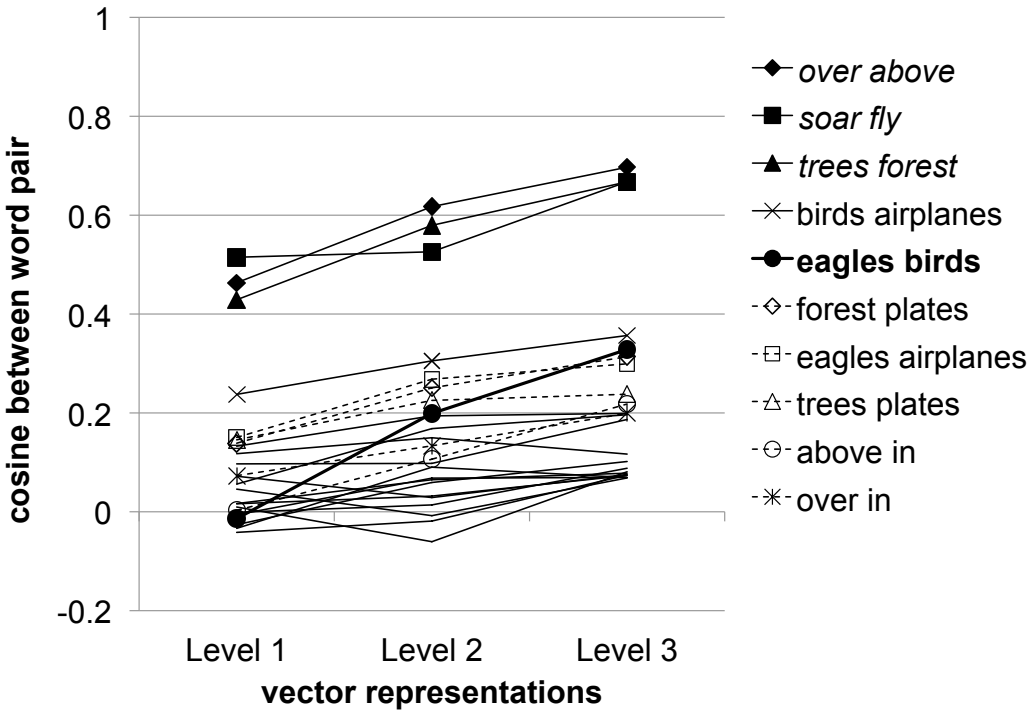


Figure 2: Cosines between word pairs across levels.

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 because 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).

Because the Level 2 environment vectors are more similar to each other than the Level 1 environment vectors, the memory vectors for the pairs *soar*

and *fly*, *above* and *over*, and *forest* and *trees* are also more similar at Level 2 than at Level 1 (see Figure 2). As a result, the Level 3 environment vectors for the three word pairs will be more similar at Level 3 than Level 2, which drives up the similarity between *eagles* and *birds* (cosine = 0.33). Eventually, at even higher levels, each pair *soar* and *fly*, *above* and *over*, and *forest* and *trees* will converge approximately to a point (cosine ≈ 1), causing *eagles* and *birds* to converge as well. The similarity between *eagles* and *birds* will never exceed the similarity between the three words pairs that the fourth-order association is contingent upon because it is the strengthening of those second-order associations that drives the strength of the fourth-order association.

3.2. Training the Model

We train HHM on the Novels Corpus from Johns et al. (2016b) with 10 238 600 sentences, 145 393 172 words, and 39 076 unique words. HHM reads the corpus one sentence at a time. Within each sentence, HHM uses a moving window centered on a target word. Within the window, all n-grams that include the target word, from bigrams up to n-grams of window width, are encoded as convolutions of environment vectors and summed into the target word's memory vector. We use 1024 dimensional vectors and four levels of representations, where Level 1 is sensitive to second-order associations, Level 2 to fourth-order, Level 3 to sixth-order, and Level 4 to eight-order.

At each level of HHM, we experiment with four maximum n-gram sizes:

1. **5-gram HHM**: an 11 word window (5 words to the left and right of the target) where the model learns all 2- to 5-grams within that window,
2. **11-gram HHM**: an 11 word window where the model learns all 2- to 11-grams within that window,
3. **21-gram HHM**: a 21 word window where the model learns all 2- to 21-grams within that window, and
4. **Sentence HHM**: a sentence-length window, where the model learns all bigrams to sentence length n-grams within that window.

For all models, the window cannot cross sentence boundaries (e.g., in a five-word sentence, the 21-gram HHM uses a five-word window). Note that for the 5-gram HHM, the maximum n-gram size (5) is distinct from the window size (11), whereas for the three other models the window size is also the maximum n-gram size. We consider large window sizes to account for human sensitivity to long-range dependencies in language, though given that humans can, in

principle, be sensitive to arbitrarily long-range dependencies, we consider the fixed context window to be an approximation (see §5.2 for discussion).

We use the four HHMs for the following experiments.

3.3. Experiment 1: 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 to 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 syntactic types proposed by Combinatory Categorical Grammar (CCG; Steedman and Baldridge, 2011).

Using WordNet (Princeton University, 2010) and the Moby Part-Of-Speech List (Ward, 1996), we assign a set of part-of-speech tags to each word in the 39 076 word vocabulary. 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 as long as it can account for how humans produce and comprehend sentences. Nevertheless, looking at the relationship between the representations of HHM and part-of-speech categories can illustrate the effect of the higher levels of the model.

Here we analyze the 11-gram HHM, as it is the model with the highest correlation to CCG types in Experiment 2. However, we inspected other windows sizes for this analysis and did not observe substantive differences. To examine the effect of higher-order associations, we compare Levels 1 and 2 (i.e., second- vs fourth-order associations), Levels 2 and 3 (i.e., fourth vs sixth), and Levels 3 and 4 (i.e., sixth vs eighth).

To provide clear examples of higher-order associations and their relationships to part-of-speech, we limit our intial analysis to words with at least 1000 occurrences in the corpus, as these words have the most robust vector representations. While the part-of-speech of some words (e.g., *manager*, a noun) may be easy to learn from only a few examples, words with more flexible part-of-speech (e.g., *course*, which can be used as a noun, verb, or adjective) may require more examples to learn all the ways in which the word

can used, particularly if one of the uses is obscure (e.g., *entire*, *n.*, an uncastrated horse). We also limit our initial analysis to the 500 unique word pairs that increase or decrease in similarity the most between levels. By unique, we mean that we select word pairs where neither word in the pair is present in any of the other 500 word pairs to ensure statistical independence.

By limiting our analysis in these ways, we focus on unambiguous examples of relationships between words that are affected by fourth-order associations. However, by limiting our analysis, we limit the scope of our conclusions in this analysis to high-frequency words and strong associations. As such, we also conduct more general analyses in this and later sections.

To illustrate the nature of higher-order associations, the word pairs that changed the most in similarity between pairs of levels are shown in Table 2. The word pairs that increase the most from Level 1 to 2 can be understood as the most pure examples of words with fourth-order associations but no second-order associations. For example, *focusing* and *derived* have a cosine of -0.10 at Level 1, indicating no second-order association, but have a cosine of 0.86 at Level 2, indicating a strong fourth-order association. Likewise, the word pairs that increase the most from Level 2 to 3 can be understood as the most pure examples of sixth-order associations, and from Level 3 to 4, eighth-order associations.

We can see in Table 2, that the four word pairs that increase the most in similarity from Level 1 to 2 are unrelated in meaning, which suggests that second-order associations are sufficient for semantics. However, the top four word pairs that increased the most in similarity from Level 1 to 2 each have exactly matching part-of-speech. While the words *focusing* and *derived* are unrelated in meaning, they are both typically verbs that can also be used as adjectives (e.g., a *focusing lens* or a *derived equation*). Likewise, *focused* and *emerging* can both be used as either an adjective or a verb.

By contrast, from Level 3 to 4, the word pair that increases the most in similarity is *across* and *druid*, which has neither meaning nor part-of-speech in common. The word pairs that increase and decrease the most from Level 3 to 4 suggest that Level 4 may not provide useful linguistic information.

From Level 1 to Level 2, the three word pairs that decrease the most in similarity have partially matching part-of-speech: *clerk* and *local* can both be used as nouns (e.g., *local* in the sense of a *local union branch*), as can *manager* and *main* (e.g., *main* as in *a water main*), and *operator* and *entire* (i.e., *entire* as in *an uncastrated horse*). However, the use of *local*, *main*, and *entire* as nouns is highly infrequent, whereas each is commonly used as

Table 2: The four word pairs that increased or decreased the most in similarity between each level, with each word's parts of speech (POS) and each word pairs' change in cosine similarity between levels (Δ cos). Matching part-of-speech in **bold**.

levels	word 1	word 2	POS 1	POS 2	Δ cos
1 to 2	focusing	derived	***v., adj.***	***v., adj.***	+0.95
	searching	associated	***v., adj.***	***v., adj.***	+0.93
	focused	emerging	***v., adj.***	***v., adj.***	+0.92
	perched	emerged	***v.***	***v.***	+0.92
	clerk	local	***n.***, *v.*	***n.***, *adj.*	-0.37
	manager	main	***n.***	***n.***, *adj.*	-0.37
	operator	entire	***n.***	***n.***, *adj.*	-0.37
	truth	outer	*n.*	*adj.*	-0.37
2 to 3	beings	accord	*n. plural*	*n., v.*	+0.55
	course	cent	***n.***, *v., adv.*	***n.***	+0.50
	lone	amounts	*adj.*	*n. plural, v.*	+0.50
	prime	bye	***n.***, *v., adj.*	*exclam.*, ***n.***	+0.48
	eh	velvet	*exclam.*	*n., adj.*	-0.14
	huh	silk	*exclam.*	*n.*	-0.12
	creaked	hemisphere	*v.*	*n.*	-0.11
	erupted	regions	*v.*	*n. plural*	-0.11
3 to 4	across	druid	*prep., adv.*	*n.*	+0.37
	course	ought	***n., v., adv.***	***n., v., adv.***	+0.37
	been	Russians	*v.*	*n. plural*	+0.36
	must	fraction	***n., v.***	***n., v.***	+0.36
	huh	which	*exclam.*	*det., pron.*	-0.05
	eh	however	*exclam.*	*adv.*	-0.04
	distinction	nineteenth	***n.***	***n.***, *adj.*	-0.03
	but	furthermore	***adv.***	***adv.***	-0.03

adjectives. As such, these three word pairs are better understood as examples of mismatching part-of-speech (nouns vs. adjectives). Because a partial part-of-speech match is not indicative of the relative frequency of the multiple uses of the word, it is difficult to interpret whether a partial match is more like a match or a mismatch. Thus, we focus our analysis on exact matches.

For the 500 word pairs that increased or decreased the most in similar-

ity between each level, Figure 3 shows how many are exact part-of-speech matches, partial matches, or have mismatching part of speech. In total, 13% of all words pairs in the lexicon are exact part-of-speech matches. Among the 500 unique word pairs that increased the most from Level 1 to Level 2, there are significantly more (18%) exact matches than would be expected in a random sample of word pairs ($p < 0.01$). Of the 500 unique word pairs that decreased in similarity the most from Level 1 to 2, 9% are exact matches (e.g., both *great* and *stranger* can be used as an adjective and a noun), which is significantly fewer than expected in a random sample ($p < 0.01$).

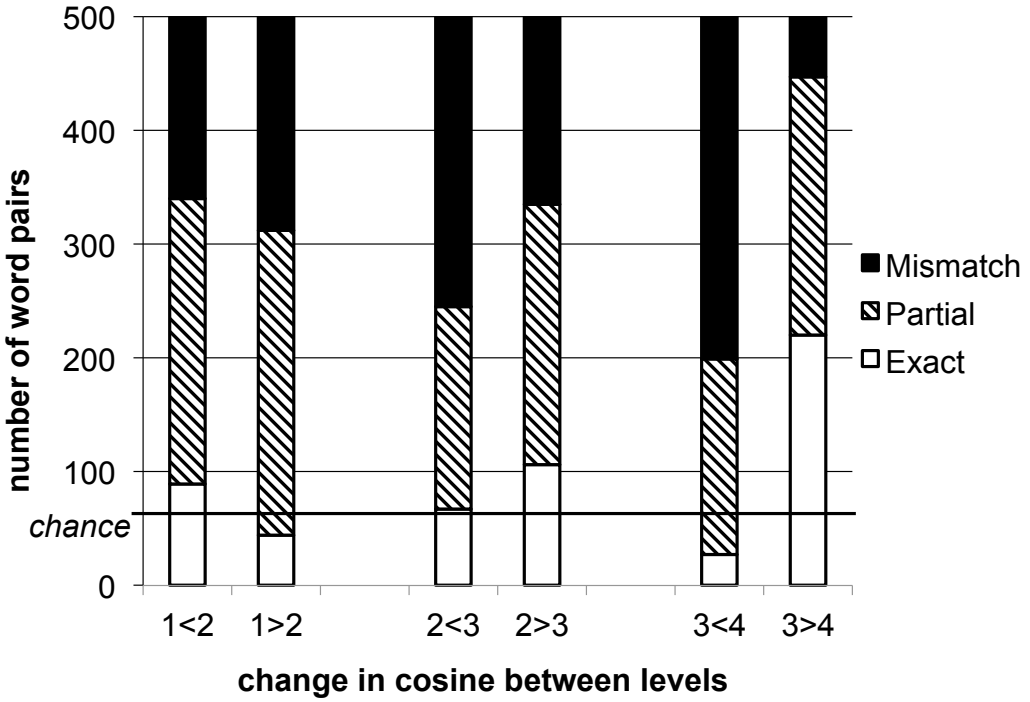


Figure 3: 500 unique word pairs that increased/decreased the most in similarity at each level, categorized by part-of-speech match.

However, from Level 2 to 3 and from Level 3 to 4, significantly more exact matches than expected in a random sample are among the top 500 word pairs that decrease the most ($p < 0.0001$). From Level 3 to 4, significantly fewer exact matches are among the 500 word pairs that increase the most ($p < 0.0001$).

The reversal suggests that fourth-order associations are sufficient to discover most exact part-of-speech matches. Indeed, from Level 2 to 3, among the 500 word-pairs that decrease the most, the exact matches have a mean decrease in similarity of 0.00, with a mean cosine of 0.90 between the words at both Levels 2 and 3. Likewise, from Level 3 to 4, the mean decrease in similarity is 0.00 for exact matches, with a mean cosine of 0.98 at both Levels 3 and 4. The exact matches are already highly similar by Level 2 and remain highly similar at Levels 3 and 4, and as such their similarity is increased little by sensitivity to sixth- and eighth-order associations.

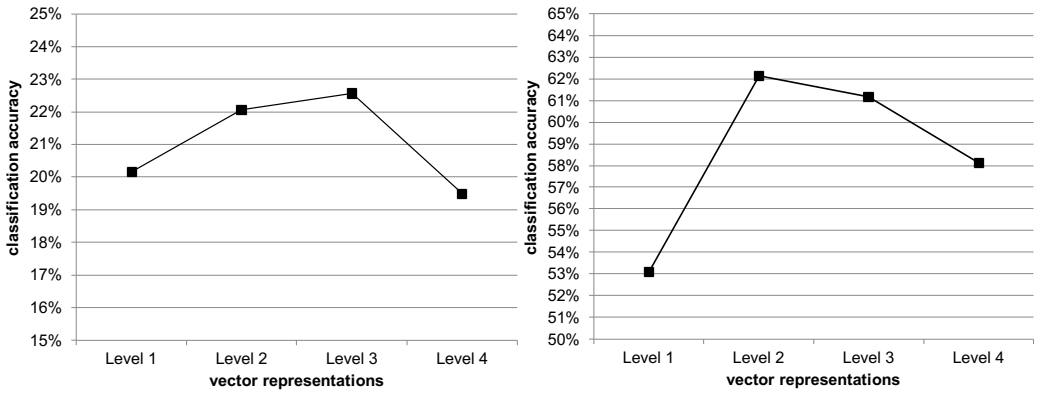
Our analysis thus far has focused on a highly select sample of the corpus: words that occur at least 1000 times and word pairs whose similarity changes dramatically between levels. For the purposes of a more general analysis, we test the ability of HHM to classify the parts-of-speech of all words that occur in the corpus at least five times. Among the 37 543 words that occur at least five times, there are 104 unique sets of part-of-speech tags. We construct a prototype for each part-of-speech tag set as a sum of the vectors for each word that has the exact same tag set. We then classify each word in the lexicon according to the closest prototype, as measured by cosine similarity.

As shown in Figure 4a, at Level 1, 20% of words are closest to the prototype that matches the word's parts-of-speech. Classification accuracy modestly improves at Levels 2 (22%) and 3 (23%) before declining at Level 4 (19%). Accuracy is not high, however, as there are 104 part-of-speech prototypes, chance classification accuracy is at 1% correct.

We re-run the classification only using each word's most frequent part-of-speech tag in the corpus. We identify the dominant tag using the Stanford Log-Linear Part-of-Speech Tagger (Toutanova et al., 2003) from the Stanford CoreNLP package (Manning et al., 2014)[4]. Again, we exclude words that occur less than five times in the corpus, as well as words with unique part-of-speech tags (e.g., the word "to" is the only word assigned the tag "TO") for a total of 37 539 words and 29 part-of-speech tags. We compute a prototype for each of the 29 tags and assign words to the closest tag. Chance classification accuracy is 3%.

As shown in Figure 4b, at Level 1, 53% of word are classified correctly. Accuracy increases to 62% at Level 2, plateaus at Level 3 (61%) and decreases at Level 4 (58%). Misclassifying nouns and adjectives as each other is the

[4] `https://stanfordnlp.github.io/CoreNLP/`



(a) all part-of-speech tags (b) dominant part-of-speech tag

Figure 4: Classification accuracy of words by closest part-of-speech prototype.

largest single source of classification errors at Level 1. At Level 1, 14% of all classifications are errors from confusing nouns and adjectives, compared to only 5% of all classifications at Level 2. The gain in classification accuracy from Level 1 to 2 is mostly due to correctly distinguishing adjectives and nouns. Conversely, confusing singular and plural nouns is a source of error across all levels (7% of all classifications at Level 1 vs. 10% at Level 4), likely due to HHM's insensitivity to case marking (see §5.4 for discussion).

In summary, strong fourth-order associations (Level 2) strengthen similarities between words with matching part of speech and weaken similarities between words with mismatching part of speech. However, sixth- and eighth-order associations (Levels 3 and 4) do little to further increase similarity between words with the same part-of-speech, and eight-order (Level 4) associations may even obfuscate part-of-speech information.

3.4. Experiment 2: Combinatory Categorical Grammar

Part-of-speech categories (nouns, verbs, adjectives, adverbs) provide a coarse-grained analysis of how words are used in English. Combinatory Categorical Grammar (CCG; Steedman and 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 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 than correlation between HHM space and part-of-speech categories.

To classify the words in HHM by CCG type, we use the *Switchboard* Corpus (Godfrey et al., 1992). The Switchboard Corpus is a collection of 2500 telephone conversations. The syntactic structure of the corpus has been annotated using CCG (Reitter et al., 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 in each dimension is a count of the number of times that word appears as the given CCG type in the corpus.

The CCG type vectors define similarity relationships between the set of 8768 words. We 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 CCG space and HHM spaces, we use Spearman's rank correlation coefficient, which is a non-parametric measure of monotonic (linear or non-linear) relationships in data.

We compute the Spearman's correlation between the CCG cosine matrix and the cosine matrix for each level of HHM. Figure 5 shows the correlation for each Level of HHM and each maximum n-gram size. The 11-gram HHM achieved the highest correlation with CCG types across all levels, peaking at Level 3 with a correlation of 0.382.

A correlation of 0.382 is not especially high, but it is worth noting that a low correlation does not indicate that HHM is wrong or that CCG is wrong. HHM's representations contain semantic information that CCG types do not contain. Likewise, CCG types may contain some particulars of syntax that it may be difficult for HHM to learn from a corpus using a sliding context window (see §5.2 for a discussion of using a memory model instead of a fixed window). Other differences may arise simply from how words are used in the Switchboard Corpus versus the Novels Corpus.

HHM's correlation to CCG is worse when the model includes up to 21-grams or full-sentence-grams, or when restricting the model to 5-grams.

Larger n-grams are not always better: as larger n-grams are more unique, they may be less useful for making inferences about new sentences.

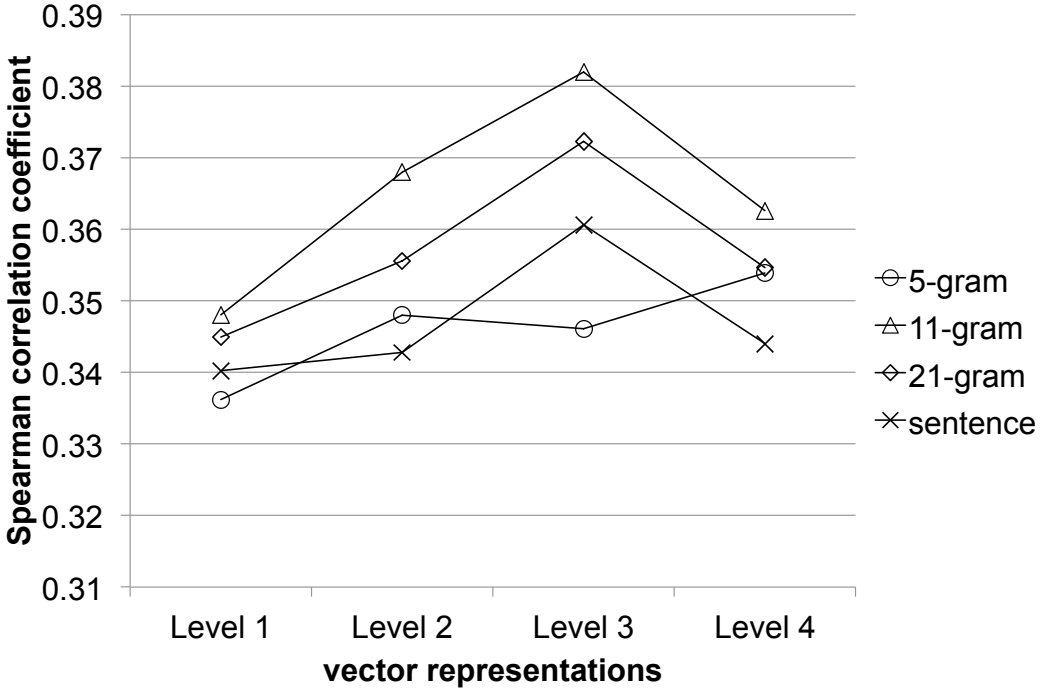


Figure 5: Spearman's rank correlation coefficient between HHM vectors and CCG types.

We see the same pattern across the four levels for the 11-gram, 21-gram, and full-sentence HHM, with increasing correlation until Level 3, and then a decrease at Level 4. Though correlation to CCG types is lowest at Level 1 for all models, the increase in correlation is modest, indicating that Level 1 can account for much of the information captured by CCG types.

The 5-gram model does not replicate this pattern of correlation, which we attribute to differences in how the 5-gram model is constructed. Whereas the other models compute all n-grams within the window, the 5-gram model computes only 2 to 5-grams within an 11-word window. The dissociation between window size (11) and maximum n-gram (5), appears to produce a different behavioural profile than when window size and maximum n-gram size are scaled together, though we do still see a general upward trend in correlation at higher levels. The 5-gram model is essentially just the 11-gram model with the 6- to 11-grams removed, as both models look forward

and backward 5 words from the target word. However, the 5-gram model's correlation to CCG types is lower than the 11-gram model at all levels, which suggests that the 6- to 11–grams are, in fact, providing useful information about syntax, independently from the size of the context window.

In summary, higher-order associations (up to Level 3, i.e., sixth-order associations) improve the ability of the model to capture syntactic type relationships and that large n-grams, in the range from 6-grams up to at least 11-grams, provide useful information about the syntactic type of words.

3.5. Experiment 3: 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 et al. (2016a). 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 original ordering.

HHM is not, by itself, able to perform the word ordering task, because HHM does not operate on sentences. However, HHM's representations contain word-level information that can be leveraged to perform the task when situated within a sentence-level model. We use a simplified version of the exemplar model used by Johns et al. (2016a). 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 HHM's memory vectors representing each word in the sentence. The second vector is the sum of all ordered sequences of words in the sentence, from individual words up to 7-grams. Each sequence is constructed as a convolution of HHM's memory vectors for each word in the sequence. Before use, all HHM vectors are normalized to a Euclidean length of one and permuted, as shown in Equation 2.

Test items are a set of 200 seven-word sentences as used by Johns et al. (2016a). 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 exemplar model first selects the exemplar sentence most similar to the test item, as measured by the 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.

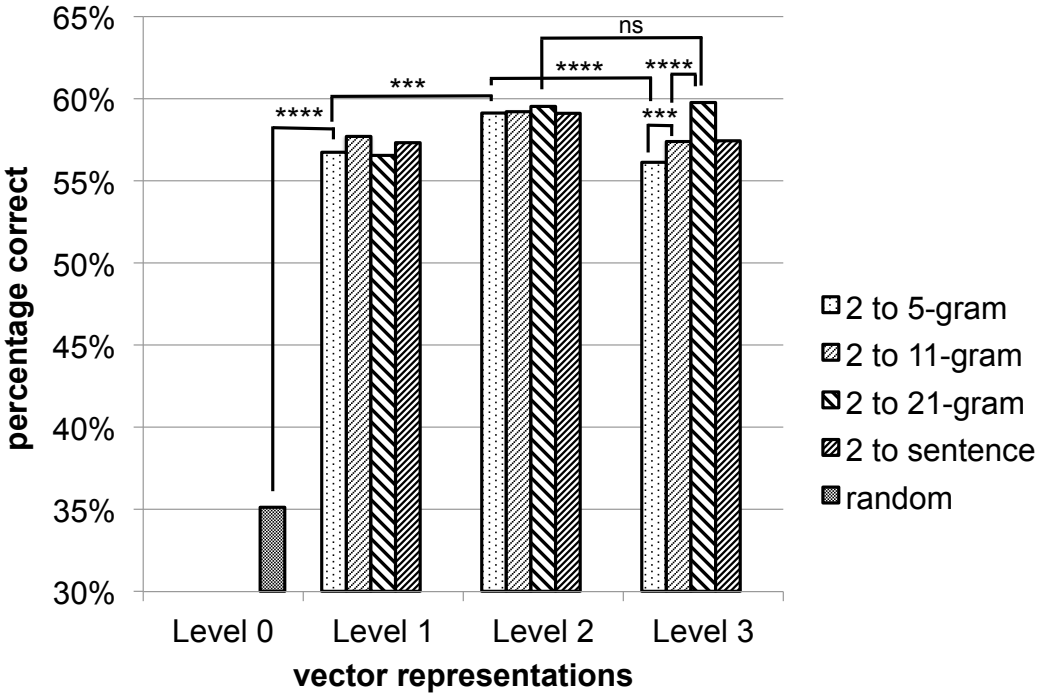


Figure 6: Test sentences correctly ordered by model as a function of vectors used to represent words.

We test all four versions of HHM from Level 1 to Level 3. 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 6. To test for statistical significance across the seven conditions, we use a repeated-measures permutation test, a non-parametric measure (Mewhort et al., 2010, 2009).

We also include a "Level 0" as a baseline for performance. Level 0 represents individual words as randomly generated vectors and the sentence vectors are constructed from those vectors. In effect, at Level 0, the model selects the exemplar sentence with the most words in common with the test item and applies the word ordering of the selected exemplar to the test item. Level 0 provides a baseline where the model is sensitive to neither semantic similarity nor higher-order associations but is sensitive to word overlap between the test item and exemplars. Level 0 gets a mean of 35.1% correct.

Level 1 outperforms Level 0 across all window sizes ($p < 0.0001$) with a mean of 57.1% correct. Level 1 selects the exemplar sentence that has the most semantic similarity to a given test item.

Level 2 outperforms Level 1 across all window sizes ($p < 0.001$) with a mean of 59.2% correct, demonstrating that fourth-order associations contribute to the task of ordering words into grammatical sentences.

At Level 3, performance declines 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 ($p < 0.0001$).

Inspecting by hand the errors made in a single run of Level 0 and each level of the 21-gram HHM, we find that the pattern of errors varies little across levels. The Levenshtein edit distance from a produced error to a correct ordering has a mean of 3 at each level of the model. All levels occasionally suggest a grammatical ordering of the words different from the original ordering (e.g., "he opened the door and got up", an incorrect ordering at Level 3, is grammatical even if "he got up and opened the door" would be a more typical sequence of actions). At Level 0, we found that an additional 6.5% of the 200 sentences produced were grammatical but not the original ordering. At Level 1, we found an additional 11.5% to be grammatical, at Level 2, an additional 11.0%, and at Level 3, an additional 7.5%. The remaining incorrect orderings are ungrammatical, typically due to a misplaced word (e.g., "came a serious look over his face", at Level 1, misplaces the verb *came*, or "I do not much trust you that", at Level 2 misplaces *much*).

Our results show that for the task of ordering words into grammatical sentences, a model that uses fourth-order associations between words outperforms a model that uses second-order associations. Our results also show that a model that uses second-order associations or higher outperforms a model that only uses word overlap (i.e. Level 0).

The results show little benefit to using a window beyond 5-grams, possibly because the task is restricted to constructing 7-gram sentences. However, the 5-gram HHM performs the worst at Level 3 and the 21-gram HHM performs the best, which 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 task performance from Levels 2 to 3 is due to all HHMs losing the ability to make fine discriminations for small n-grams. Performance of HHM representations that contain larger n-grams is less affected as those models are simultaneously gaining an ability to better use those large n-grams.

To test this hypothesis, we break down HHM into its constituent n-gram components. While the HHMs previously discussed learned 2-grams up to n-grams for some n, here we train each HHM on one and only one size of n-gram. For Level 0, we use random vectors. For Level 1, we use Level 0's random vectors to construct a 2-gram only HHM, a 3-gram only HHM, etc., up to a 7-gram only HHM. For Level 2, we construct the HHMs out of Level 1 of the 2- to 21-gram HHM. For Level 3, we construct the HHMs out of Level 2 of the 2- to 21-gram model. We use the 2- to 21-gram HHM as it is the model with the most robust performance across all levels on this task.

Figure 7 shows the percentage of test sentences ordered correctly across different sizes of n-gram and levels of HHM. Results are averaged across 10 sets of 125 000 exemplar sentences. Higher levels of the model are better able to use larger n-grams. Level 1 of HHM is best able to use 2-grams and 3-grams. Conversely, at Level 3 of HHM, the model is able to make use of large n-grams, but performance declines for smaller n-grams. Task performance at Level 2 of HHM peaks for 3- to 6-grams, and declines for 2- and 7-grams. Level 0 is included for a baseline performance of 35.1% correct.

Figure 7 illustrates that 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. However, 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

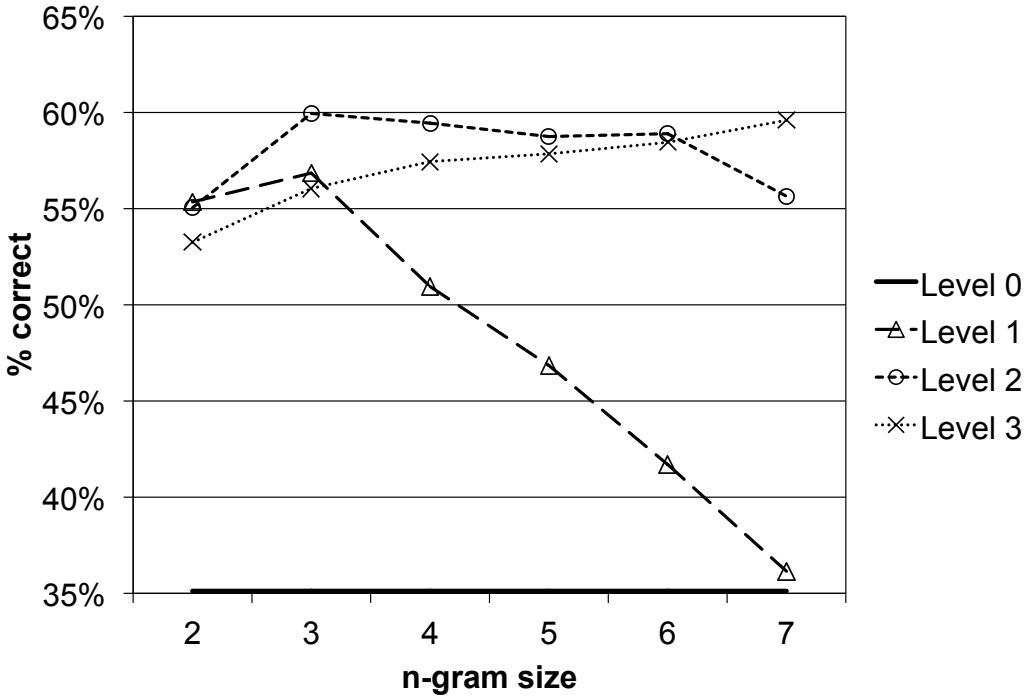


Figure 7: Test sentences correctly ordered as a function of n-gram size and HHM level.

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".

Correctness is a noisy metric of model skill as it is binary. We can get a more precise measure of model skill by using the cosine scores assigned to each of the 7! alternative orderings. To measure the degree of confidence with which the model endorses a given ordering as grammatical, we use the deviation of the grammatical ordering's cosine from the cosines of the other orderings. The deviation is a graded measure, sensitive to how close the model is to wrong when it's correct and how close to correct the model is when it's wrong, giving us a better picture of the model's decisions. We normalize the deviation by the standard deviation to control for differences in the spread of cosine values at different levels.

As shown in Figure 8, the deviations yield the same pattern of results as Figure 7. The ability of the Level 1 models to discriminate between the correct answer and alternatives is highest for 2-grams and 3-grams and declines

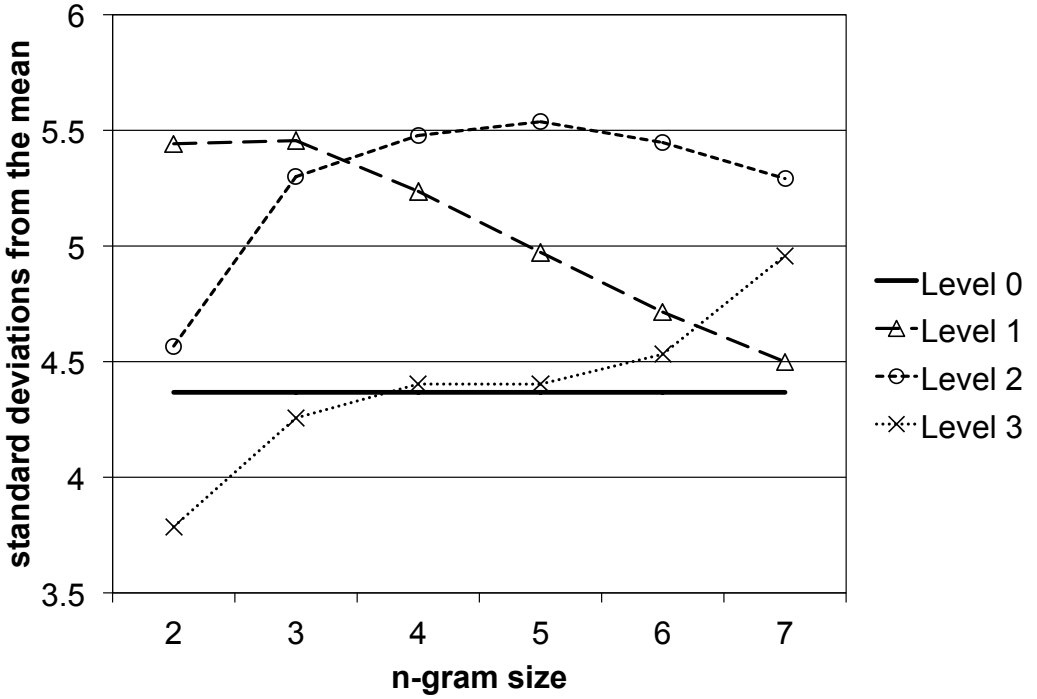


Figure 8: Deviation of correct word ordering from alternatives as a function of n-gram size and HHM level.

for larger n-grams. At Level 3, we observe the opposite: the deviation of the correct answer is highest for 7-grams and declines for smaller n-grams. At Level 2, deviation peaks at 5- grams, declining for smaller or larger n-grams.

The results in Figures 7 and 8 demonstrate that the higher levels of HHM allow for better use of large n-gram information, at the cost of a declining ability to make discriminations between small n-grams. Specifically, Level 1 (i.e., second-order associations) representations make the best use of 2-grams and 3-grams, Level 2 (i.e., fourth-order associations) makes the best use of 4-grams to 6-grams, and Level 3 makes the best use of 7-grams.

Note that none of the combined n-gram models in Figure 6 outperform the 3-gram only Level 2 model or the 7-gram only Level 3 model, which suggests that simpler HHMs may be sufficient for the word ordering task.

Accuracy on the task can be increased by using more data. By increasing the size of the exemplar set from 125 000 sentences to 500 000, accuracy for the 21-gram Level 2 HHM improves from 60% correct to 68% correct.

Performance can also be improved by using a more complex model. Rather than selecting the best exemplar, Johns et al. (2016a) use a weighted sum of all exemplars to make word-ordering decisions. Each exemplar is weighted by the cosine similarity between the exemplar and the unordered set of words in the test item, raised to a power (a fitting parameter).

Johns et al. (2016a) report a best accuracy of 60% with 500 000 exemplars, random vectors (Level 0), and an exponent of 9. We find a best accuracy of 76% correct with 500 000 exemplars, the 21-gram Level 2 HHM vectors, and an exponent of 450. However, our aim is not to optimize accuracy on the word-ordering task, but to illustrate the role of higher-order associations in constructing grammatical sentences.

3.6. Experiment 4: Colorless green ideas sleep furiously

Chomsky (1956) gives "Colorless green ideas sleep furiously" as an example of a sentence that is grammatically correct but meaningless. By contrast, Chomsky notes that "Furiously sleep ideas green colorless" is ungrammatical. Chomsky uses this example 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 the sentence nor determine that it is grammatical.

Pereira (2000) demonstrates that a statistical model can, in fact, discriminate between "Colorless green ideas sleep furiously" and the ungrammatical "Furiously sleep ideas green colorless". Pereira uses an *aggregate bigram model* that estimates the probability of each bigram in "Colorless green ideas sleep furiously" by using second-order associations to known bigrams and an expectation-maximization algorithm (Dempster et al., 1977). Pereira's aggregate bigram model finds that "Colorless green ideas sleep furiously" is about 20 000 times more likely than "Furiously sleep ideas green colorless".

HHM is also a statistical model that can be understood as estimating the probability of unseen n-grams through the use of higher-order associations. 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 over "Furiously sleep ideas green colorless" or the 118 other orderings? 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 use the 21-gram HHM. The exemplar model is provided a set of 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 score the 120 possible orderings.

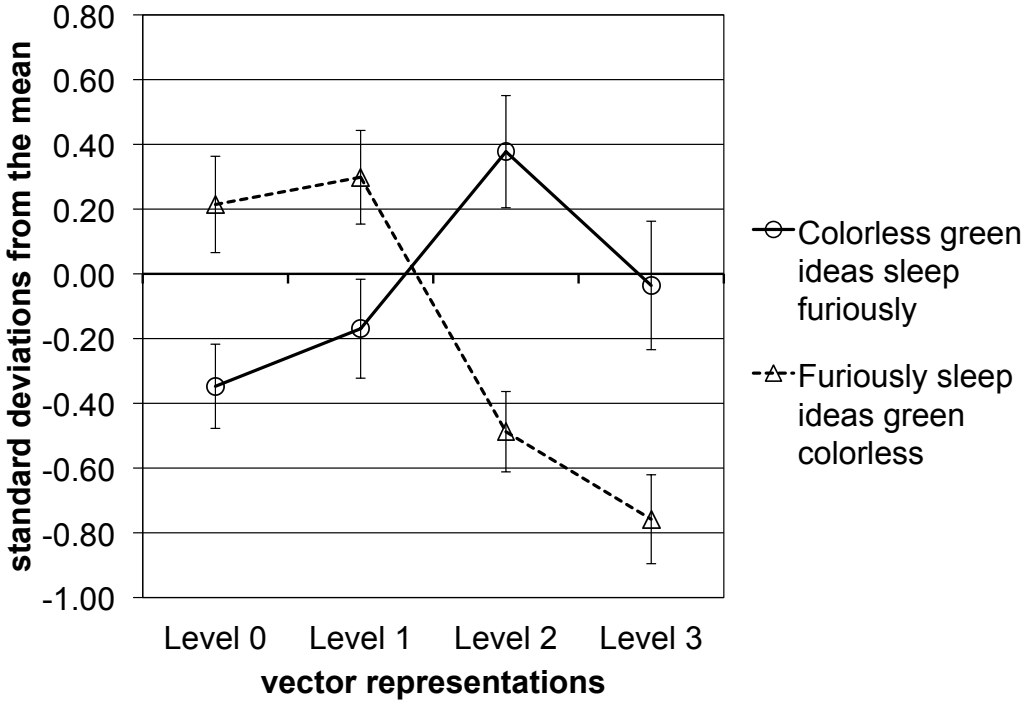


Figure 9: Deviation from the mean cosine score as a function of HHM level.

Mean deviation of both "Colorless green ideas sleep furiously" and "Furiously sleep ideas green colorless" at each level of HHM is shown in Figure 9. Results are averaged across 50 different random sets of 125 000 sentences. Error bars indicate standard error. Word orderings above zero are judged to be more grammatical than the mean of the 120 possible sentence orderings and orderings below zero are judged to be less grammatical than the mean.

Is "Colorless green ideas sleep furiously" more grammatical, according to the exemplar model, than "Furiously sleep ideas green colorless"? To test for statistical significance, we use a repeated-measures permutation test. At Level 0, "Furiously sleep ideas green colorless" is more grammatical ($p <$

0.05) whereas at Levels 2 and 3, "Colorless green ideas sleep furiously" is more grammatical ($p < 0.05$). Given that the two sentences are exact reverse orderings of each other, it is not surprising that the model's confidence in each sentence is roughly the inverse of the other's.

Thus, selecting exemplar sentences with words in common with the test set (e.g., *green*, *furiously*, etc.), as the model does at Level 0, is not enough to make the correct grammatical distinction. Selecting sentences with similar meanings (e.g., *red*, *angrily*, etc.), as Level 1 does, is likewise insufficient. Higher-order associations at Levels 2 and 3, seem to be necessary to determine that "Colorless green ideas sleep furiously" is the more grammatical alternative of the pair.

Identifying 'Colorless green ideas sleep furiously" as the *most* grammatical ordering of the 120 possible orderings is a more difficult problem. Only at Level 2 does the model judge "Colorless green ideas sleep furiously" to be more likely than average ($p < 0.05$), selecting it as the most likely ordering 7 times out of 50. The rate at which Level 2 selects "Colorless green ideas sleep furiously" as the preferred alternative might be improved by either increasing the number of exemplars over the current 125 000 or by using a more sophisticated model (e.g., Johns et al., 2016a; Gulordava et al., 2018).

The results suggest that "Colorless green ideas sleep furiously" cannot be judged as grammatical by analogy to sentences with either the same words or words with similar meanings. However, sensitivity to fourth-order associations causes representations for words with similar syntactic type to look increasingly alike, such that "Colorless green ideas sleep furiously", or *adjective adjective noun verb adverb*, begins to look like an English sentence.

4. Other Models of Higher-Order Associations

While we have based HHM on BEAGLE, it is possible to use other models to detect higher-order associations in language.

The Associative Smoothing Network (Roberts and Chater, 2008), for example, is a spreading activation model that uses third-order associations to make sentence acceptability judgments. The network has no inherent limitation to how far activation can spread, and so can be easily applied recursively to detect fourth-, fifth-, sixth-order associations and higher.

However, for some models, there's no trivial way to recursively apply the model to incorporate higher-order associations. For example, the *word2vec* neural network expects, for each word it takes as input, a vector that uses

one-hot encoding[5]. Conversely, the semantic vectors *word2vec* generates have d dimensions, where d is much smaller than the size of the lexicon, and each dimension is real valued and individually meaningless.

Because semantic vectors and one-hot vectors have such different properties, the semantic vectors cannot be re-used as input to *word2vec* to recursively detect higher-order associations. While it's almost certainly possible to design a neural network model of distributional semantics that can be recursively applied in much the same manner as HHM, *word2vec* cannot be used to do so as standardly implemented.

BEAGLE and HHM also have a unique property that may make replicating our results with other models difficult. Other models of distributional semantics only learn relationships between pairs of words, whereas BEAGLE and HHM learn a relationship between a word and sequences of words.

Models limited to knowing the relationships between pairs of words can certainly benefit from third- or fourth-order associations. The Associative Smoothing Network, for example, is strictly a bigram model, but third-order associations allow the model to make judgments about the acceptability of novel word pairs (Roberts and Chater, 2008). However, in the word ordering task, we find that improvements in performance at higher orders of association largely result from improving the ability of HHM to make use of the information in larger n-grams, $n \geq 3$ (see Figures 7 and 8).

While we are not committed to the specific implementation details of how the Hierarchical Holographic Model learns higher-order associations, HHM has two desirable properties for modelling higher-orders of association:

1. HHM can be recursively applied an arbitrary number of times to learn arbitrarily high orders of association, and
2. HHM is able to learn arbitrarily large n-grams with linear time complexity and constant space complexity.

5. Future Work

The Hierarchical Holographic Model (HHM) has a number of limitations, namely, (1) HHM learns batch-style rather than online, (2) HHM's fixed window is unrealistic, (3) HHM does not combine levels of representation,

[5]In *one-hot encoding*, a vector has one dimension for each word in the lexicon. To represent a word, that word's dimension is set to 1 and all other dimensions are set to 0.

and (4) HHM is applied only to English. HHM also has possible applications beyond what we explore in this paper, such as (5) modelling developmental language acquisition. We discuss each avenue for future research in turn.

5.1. Online Learning

HHM is not an online model of learning. HHM learns each level of representation sequentially. In keeping with research on the acquisition of first and second order associations in children and adults (McNeill, 1963; Sloutsky et al., 2017), we would expect learning to happen at each order of association continuously and in parallel. Though we hypothesize that higher-order associations scaffold off lower-order associations, we hold that the scaffolding is such that higher-order associations are adjusted as new lower-order associations are learned.

HHM may be able to learn all levels in parallel. Doing so will introduce more noise into the higher levels of the model, as early on, the level(s) below will not have stable memory vectors yet, but over a large enough corpus, stable representations should emerge first at Level 1 and then propagate upward to higher levels.

5.2. Window Size

Humans are sensitive to long-range dependencies in language. For example, in anaphora resolution, readers are able to identify the referent of a pronoun such as *she* even over a large number of intervening words or sentences. Readers selectively and strategically maintain pertinent information in memory from much earlier in a sentence, paragraph, or passage (Kintsch and Van Dijk, 1978).

We include large window sizes in our simulations as proxies for the capacity of memory to selectively retain long-range information. The sliding context window of HHM is best understood as an inexact proxy for the linguistic associations and dependencies available to a reader (or listener) when the target word in a sentence is encoded. However, human memory does not behave like a verbatim list of the last 21 words read (or heard).

To build a more detailed model of human sentence processing, we would need to replace the sliding window with a model of the linguistic information maintained in working memory and stored in long-term memory, as in Kintsch and Van Dijk's (1978) model of sentence processing. We would also need a model of selective attention to account for what information is

retained in long-term memory and maintained in working memory, as informed by the model's experience of what is likely to be useful for resolving the syntax and semantics of future utterances. Computational, holographic approaches to modelling working memory (Franklin and Mewhort, 2015) and episodic memory (Jamieson and Mewhort, 2011) could potentially be integrated with HHM to provide a more detailed processing model.

5.3. Combining Levels

Gruenenfelder et al. (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, Level 2 with any window, or Level 3 with the 21 word window, produces the best results, Level 1 often correctly ordered sentences that Levels 2 or 3 got wrong. Perhaps a model that uses all three levels could outperform a model that uses only one level at a time. A neural network model that combines input from varying n-gram sizes and from varying orders of association might be able to outperform a neural network that strictly takes traditional word embeddings as input. We hypothesize that human memory is able to use relations between concepts at varying levels of abstraction as needed to meet task demands.

5.4. Other Languages

In languages with extensive case marking (e.g., Latin), case markers are used to indicate the part-of-speech of a word instead of relying on word order, as English does. To learn the case markers, HHM would need to either process the corpus parsed into sub-word units (e.g., Cotterell and Schütze, 2015), splitting off the case marker from the root word, or to use non-random environment vectors that represent the orthography of the word, as in Cox et al. (2011).

The utility of HHM's sensitivity to word sequence and higher-order associations for modelling case-marked languages is an open question. Case-marked languages typically use word sequence to convey non-syntactic information (e.g., emphasis or new information), such that while preserving word order may not be important for syntax, *per se*, order remains important for conveying meaning. Thus while the type of information captured by HHM's sensitivity to word sequence and abstract associations may differ in case-marked languages, we expect that sequence and associations will still play an informative role. HHM's central hypothesis is that human memory has

the capacity for sensitivity to abstract associations, even if those associations are potentially used differently across languages.

5.5. Language Acquisition

Children acquire first-order associations earlier in development than second-order associations (Brown and Berko, 1960; Ervin-Tripp, 1970; Nelson, 1977; Sloutsky et al., 2017). Likewise, 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 produce only first-order associations between the non-words, but by 60 trials, participants produce both first- and second-order associations.

Sloutsky et al. (2017) propose a neural network model that captures the gradual acquisition of second-order associations contingent on learning first-order associations sensitive to sequential word order, as well as the acquistion of order-independent first-order (syntagmatic) associations. Similarly, the Syntagmatic-Paradigmatic Model (Dennis, 2004, 2005) is a computational model of human memory and language learning that postulates two long-term memory systems: one for sequences and one for (order-independent) relations, which respectively account for knowledge of first-order (syntagmatic) and second-order (paradigmatic) associations

According to Barcel-Coblijn et al. (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 transitions from a sparsely connected graph to 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 longer range connections between words is necessary to construct novel syntactic utterances. We speculate that a model that captures higher-order associations, such as an online variant of HHM that uses both *context* and *order* vectors, and is therefore sensitive to both super-paradigmatic and super-syntagmatic associations, may be able to account for the dynamics of a child's language learning process.

6. Conclusions

We define orders of association and explore the hypothesis that higher-order associations in language capture syntactic relationships between words. We propose a "deep" model of distributional semantics, the Hierarchical

Holographic Model (HHM), sensitive to higher-order associations. We evaluate the correlation between HHM's representations, part-of-speech, and the lexical syntactic types of Combinatory Categorical Grammar (CCG; Steedman and Baldridge, 2011). We find that strong fourth-order associations are likely to increase similarity between words with the same part-of-speech and decrease similarity between words with mismatching part-of-speech. Fourth- and sixth-order associations increase correlation with CCG type relative to second-order (i.e., paradigmatic) associations.

Fourth-order associations also improve the ability of HHM's representations to order words into grammatical sentences, including nonsense sentences such as Chomsky (1956)'s "Colorless green ideas sleep furiously". The usefulness of higher-order associations interacts with the window size of the distributional semantics model, such that larger n-grams require higher orders of association in order to contribute useful information, whereas smaller n-grams are best represented using lower orders of association.

In summary, we find consistent evidence that fourth-order associations (Level 2) provide useful linguistic information of a syntactic character. Conversely, the evidence is mixed for sixth-order (Level 3), and we find no evidence that eighth-order associations (Level 4) are useful for linguistic tasks.

We hypothesize that humans are also sensitive to higher-order associations in non-linguistic domains. Humans have the ability to abstract away from the specifics of an experience (i.e. episodic memories) to infer concepts (i.e., semantic memories) from the patterns that occur across multiple experiences (e.g., Hintzman, 1986). The theoretical claim of HHM is that the pattern inference process is recursive, such that human memory can also infer meta-concepts from patterns across concepts, and that these meta-concepts play an important role in human behaviour, such as language.

7. Acknowledgments

We thank Eilene Tomkins-Flanagan, Kevin D. Shabahang, and the late D. J. K. Mewhort for the use of their BEAGLE code. We are also indebted to D. J. K. Mewhort for comments on the paper and the use of his server, funded by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC: APA 318). We thank Jeremy R. Cole for use of his CCG vectors and for his comments on the paper. We thank Saranya Venkatraman for her insights regarding the *word2vec* model. This research has been funded by an Ontario Graduate Scholarship and NSERC Post-Doctoral Fellowship

to M. A. Kelly, a National Science Foundation grant (BCS-1734304) to David Reitter and M. A. Kelly, and a grant from NSERC to Robert L. West.

References

Anderson, J.R., 2009. How can the human mind occur in the physical universe? Oxford University Press, New York, NY.

Barcel-Coblijn, L., Corominas-Murtra, B., Gomila, A., 2012. Syntactic trees and small-world networks: Syntactic development as a dynamical process. Adaptive Behavior 20, 427–442. doi:10.1177/1059712312455439.

Brown, R., Berko, J., 1960. Word association and the acquisition of grammar. Child Development 31, 1–14.

Burgess, C., Lund, K., 1997. Modelling parsing constraints with high-dimensional context space. Language and Cognitive Processes 12, 177–210.

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

Cotterell, R., Schütze, H., 2015. Morphological word-embeddings, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Denver, Colorado. pp. 1287–1292. doi:10.3115/v1/N15-1140.

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.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological) , 1–38URL: http://www.jstor.org/stable/2984875.

Dennis, S., 2004. An unsupervised method for the extraction of propositional information from text. Proceedings of the National Academy of Sciences 101, 5206–5213. doi:10.1073/pnas.0307758101.

Dennis, S., 2005. A memory-based theory of verbal cognition. Cognitive Science 29, 145–193. doi:10.1207/s15516709cog0000_9.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Minneapolis, Minnesota. pp. 4171–4186. doi:10.18653/v1/N19-1423.

Eliasmith, C., 2013. How to build a brain: A neural architecture for biological cognition. Oxford University Press, New York, NY.

Ervin-Tripp, S.M., 1970. Substitution, context, and association, in: Postman, L., Keppel, G. (Eds.), Norms of Word Association. Academic Press, pp. 383 – 467. doi: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.

Gabor, D., 1969. Associative holographic memories. IBM Journal of Research and Development 13, 156–159. doi:10.1147/rd.132.0156.

Gayler, R.W., 2003. Vector symbolic architectures answer Jackendoff's challenges for cognitive neuroscience, in: Slezak, P. (Ed.), Proceedings of the Joint International Conference on Cognitive Science. University of New South Wales, Sydney, Australia, pp. 133–138. URL: http://cogprints.org/3983/.

Godfrey, J.J., Holliman, E.C., McDaniel, J., 1992. Switchboard: Telephone speech corpus for research and development, in: [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 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, Association for Computational Linguistics, Amsterdam, The Netherlands. pp. 279–290.

Griffiths, T.L., Steyvers, M., Tenenbaum, J.B., 2007. Topics in semantic representation. Psychological Review 114, 211–244. doi:10.1037/0033-295X.114.2.211.

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, 1460–1495. doi:10.1111/cogs.12299.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., Baroni, M., 2018. Colorless green recurrent networks dream hierarchically, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics. pp. 1195–1205. URL: http://aclweb.org/anthology/N18-1108, doi:10.18653/v1/N18-1108.

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

Jamieson, R.K., Mewhort, D.J.K., 2011. Grammaticality is inferred from global similarity: A reply to kinder (2010). The Quarterly Journal of Experimental Psychology 64, 209–216. doi:10.1080/17470218.2010.537932.

Jenkins, J.J., 1964. A Mediational Account of Grammatical Phenomena. Journal of Communication 14, 86–97. doi:10.1111/j.1460-2466.1964.tb02352.x.

Jenkins, J.J., 1965. Mediation theory and grammatical behavior, in: Rosenberg, S. (Ed.), Directions in psycholinguistics. MacMillan, New York, pp. 66–96.

Jenkins, J.J., 1968. The challenge to psychological theorists, in: Dixon, T.R., Horton, D.L. (Eds.), Verbal behavior and general behavior theory. Prentice-Hall, Inc., Englewood Cliffs, N.J., pp. 538–549.

Jenkins, J.J., 1974. Remember that old theory of memory? well, forget it. American Psychologist 29, 785–795. doi:10.1037/h0037399.

Jenkins, J.J., Palermo, D.S., 1964. Mediation processes and the acquisition of linguistic structure. Monographs of the Society for Research in Child Development 29, 141–169. doi:10.2307/1165762.

Johns, B.T., Jamieson, R.K., Crump, M.J.C., Jones, M.N., Mewhort, D.J.K., 2016a. The combinatorial power of experience, in: Papafragou, A., Grodner, D., Mirman, D., Trueswell, J.C. (Eds.), Proceedings of the 38th Annual Meeting of the Cognitive Science Society, Cognitive Science Society, Austin, TX. pp. 1325–1330.

Johns, B.T., Jones, M.N., Mewhort, D.J.K., 2016b. Experience as a free parameter in the cognitive modeling of language, in: Papafragou, A., Grodner, D., Mirman, D., Trueswell, J.C. (Eds.), Proceedings of the 38th Annual Meeting of the Cognitive Science Society, Cognitive Science Society, Austin, TX. pp. 1325–1330.

Johns, B.T., Taler, V., Pisoni, D.B., Farlow, M.R., Hake, A.M., Kareken, D.A., Unverzagt, F.W., Jones, M.N., 2013. Using cognitive models to investigate the temporal dynamics of semantic memory impairments in the development of Alzheimer's disease, in: West, R., Stewart, T. (Eds.), Proceedings of the 12th International Conference on Cognitive Modeling. Carleton University, Ottawa, Canada, pp. 23–28. URL: http://iccm-conference.org/2013-proceedings/papers/0004/index.html.

Jones, M.N., Kintsch, W., Mewhort, D.J.K., 2006. High-dimensional semantic space accounts of priming. Journal of Memory and Language 55, 534–552. doi:10.1016/j.jml.2006.07.003.

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: Taatgen, N.A., van Vugt, M.K., Borst, J.P., Mehlhorn,

K. (Eds.), Proceedings of the 13th International Conference on Cognitive Modeling. University of Groningen, Groningen, the Netherlands, pp. 148–153. URL: https://iccm-conference.neocities.org/2015/proceedings/papers/0036/paper0036.pdf.

Kelly, M.A., Reitter, D., 2017. Holographic declarative memory: Using distributional semantics within act-r, in: Laird, J., Lebiere, C., Rosenbloom, P.S. (Eds.), The 2017 AAAI Fall Symposium Series: Technical Reports, The AAAI Press, Palo Alto, California. pp. 382–387. URL: https://aaai.org/ocs/index.php/FSS/FSS17/paper/view/16001.

Kintsch, W., Van Dijk, T.A., 1978. Toward a model of text comprehension and production. Psychological Review 85, 363–394. doi:10.1037/0033-295X.85.5.363.

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

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D., 2014. The Stanford CoreNLP natural language processing toolkit, in: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Baltimore, Maryland. pp. 55–60. doi:10.3115/v1/P14-5010.

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

Mewhort, D.J.K., Johns, B.T., Kelly, M., 2010. Applying the permutation test to factorial designs. Behavior Research Methods 42, 366–372. doi:10.3758/BRM.42.2.366.

Mewhort, D.J.K., Kelly, M., Johns, B.T., 2009. Randomization tests and the unequal-n/unequal-variance problem. Behavior Research Methods 41, 664–667. doi:10.3758/BRM.41.3.664.

Mewhort, D.J.K., Shabahang, K.D., Franklin, D.R.J., 2018. Release from pi: An analysis and a model. Psychonomic bulletin & review , 932–950doi:10.3758/s13423-017-1327-3.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality, in: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 26. Curran Associates, Inc., pp. 3111–3119. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Murdock, B.B., 1982. A theory for the storage and retrieval of item and associative information. Psychological Review 89, 609–626. doi:10.1037/0033-295X.89.6.609.

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

Ororbia II, A.G., Mikolov, T., Reitter, D., 2017. Learning simpler language models with the differential state framework. Neural computation 29, 3327–3352.

Pennington, J., Socher, R., Manning, C.D., 2014. GloVe: Global vectors for word representation, in: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

Pereira, F., 2000. Formal grammar and information theory: Together again? Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 358, 1239–1253. doi:10.1098/rsta.2000.0583.

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

Pribram, K.H., 1969. The neurophysiology of remembering. Scientific American 220, 73–86.

Princeton University, 2010. About wordnet. WordNet URL: 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, Association for Computational

Linguistics, Stroudsburg, PA, USA. pp. 308–316. URL: http://dl.acm.org/citation.cfm?id=1610075.1610119.

Roberts, M.A., Chater, N., 2008. Using statistical smoothing to estimate the psycholinguistic acceptability of novel phrases. Behavior Research Methods 40, 84–93.

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. Rapports syntagmatiques et rapports associatifs, in: Bally, C., Sechehaye, A. (Eds.), Cours de linguistique générale. Payot, Paris, France, pp. 170–175.

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

Sidman, M., 2009. Equivalence relations and behavior: An introductory tutorial. The Analysis of Verbal Behavior 25, 5–17. doi:10.1007/bf03393066.

Sloutsky, V.M., Yim, H., Yao, X., Dennis, S., 2017. An associative account of the development of word learning. Cognitive Psychology 97, 1 – 30. doi:10.1016/j.cogpsych.2017.06.001.

Steedman, M., Baldridge, J., 2011. Combinatory categorial grammar, in: Borsley, R., Borjars, K. (Eds.), Non-Transformational Syntax: Formal and Explicit Models of Grammar. Wiley-Blackwell, pp. 181–224.

Toutanova, K., Klein, D., Manning, C.D., Singer, Y., 2003. Feature-rich part-of-speech tagging with a cyclic dependency network, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Association for Computational Linguistics, Edmonton, Canada. pp. 173–180. doi:10.3115/1073445.1073478.

Ward, G., 1996. Moby Part-of-Speech. University of Sheffield. URL: http://icon.shef.ac.uk/Moby/mpos.html.

Appendix: Encoding order with one versus two permutations

Our approach to encoding the sequential order of words differs from Jones and Mewhort (2007). Convolution is commutative, that is, invariant to the sequential order of the operands, i.e., $\mathbf{v}_1 * \mathbf{v}_2 = \mathbf{v}_2 * \mathbf{v}_1$. However, the sequence of the words can be preserved by permuting each operand. Jones and Mewhort, p.35, using a method proposed by Plate (1995, p.12), apply two different permutations to the left and right operands of convolution, such that $(\mathbf{P}_{\text{left}}\mathbf{v}_1) * (\mathbf{P}_{\text{right}}\mathbf{v}_2) \neq (\mathbf{P}_{\text{left}}\mathbf{v}_2) * (\mathbf{P}_{\text{right}}\mathbf{v}_1)$.

We apply a permutation only to the left operand as it is simpler and sufficient for preserving sequence: $(\mathbf{P}_{\text{before}}\mathbf{v}_1) * \mathbf{v}_2 \neq (\mathbf{P}_{\text{before}}\mathbf{v}_2) * \mathbf{v}_1$. Our one-permutation method is isomorphic to using two permutations. Vectors constructed using one permutation will have, in expectation, the same spatial relationships to each other as vectors constructed using two permutations,

$$\begin{aligned}&\text{cosine}\left((\mathbf{P}_{\text{before}}\mathbf{v}_1) * \mathbf{v}_2, (\mathbf{P}_{\text{before}}\mathbf{v}_3) * \mathbf{v}_4\right) \approx \\ &\text{cosine}((\mathbf{P}_{\text{right}}\mathbf{v}_1) * (\mathbf{P}_{\text{left}}\mathbf{v}_2), (\mathbf{P}_{\text{right}}\mathbf{v}_3) * (\mathbf{P}_{\text{left}}\mathbf{v}_4))\end{aligned}$$

where spatial relationships are measured by the cosine similarity. Differences in the cosine similarity between the two methods will be due to small, zero mean variations introduced by using the $\mathbf{P}_{\text{right}}$ permutation. The isomorphism arises because convolution and permutation preserve cosine similarity relationships, such that $\text{cosine}(\mathbf{v}_1, \mathbf{v}_2) = \text{cosine}(\mathbf{P}\mathbf{v}_1, \mathbf{P}\mathbf{v}_2)$ and,

$$\text{cosine}((\mathbf{P}\mathbf{v}_1) * \mathbf{v}_2, (\mathbf{P}\mathbf{v}_3) * \mathbf{v}_4) \approx \text{cosine}(\mathbf{P}\mathbf{v}_1, \mathbf{P}\mathbf{v}_3) \times \text{cosine}(\mathbf{v}_2, \mathbf{v}_4) \quad (3)$$

for any vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$ and permutation $\mathbf{P}$.