Compositional Generalization in Distributional Models of Semantics: Transformer-based Language Models are Architecturally Advantaged

Shufan Mao (smao9@illinois.edu)
Department of Psychology, 603 E Daniel St
Champaign, IL 61820 USA

Philip A. Huebner (huebner3@illinois.edu)
Department of Psychology, 603 E Daniel St
Champaign, IL 61820 USA

Jon A. Willits (jwillits@illinois.edu)
Department of Psychology, 603 E Daniel St
Champaign, IL 61820 USA

Abstract

An important aspect of language comprehension is learning and generalizing complex lexical relations. For instance, having learned that the phrase *preserve cucumbers* predicts *vinegar* and that *preserve berries* predicts *dehydrator*, one should be able to infer that the novel phrase *preserve peppers* is more compatible with *vinegar*, because *pepper* is more similar to *cucumber*. We studied the ability to perform such (compositional) generalization in distributional models trained on an artificial corpus with strict semantic regularities. We found that word-encoding models failed to learn the multi-way lexical dependencies. Recurrent neural networks learned those dependencies but struggled to generalize to novel combinations. Only mini GPT-2, a minified version of the Transformer GPT-2, succeeded in both learning and generalization. Because successful generalization in our tasks requires capturing the relationship between a phrase and a word, we argue that mini GPT-2 acquired hierarchical representations that approximate phrase structure. Our results show that, compared to older models, Transformers are architecturally advantaged to perform compositional generalization.

Keywords: distributional semantics; language model; semantic plausibility; language comprehension; compositionality

Introduction

Knowing how words relate to each other in a sentence is crucial to language comprehension. For instance, the knowledge that *The baby is sleeping* is semantically more plausible than *The table is sleeping* is in part a consequence of knowing that *sleep* is a better fit for *baby* compared to *table*. Moreover, judgments of semantic plausibility often require understanding of the relation among more than two lexical items. For example, in the sentence *John preserves the cucumber with vinegar*, the choice of the instrument (*vinegar*) is constrained jointly by the verb (*preserve*) and the patient noun (*cucumber*). If the verb were replaced by *cut*, or the patient noun were replaced by *antiques*, the instrument *vinegar* would be less compatible. This suggests that the lexical relation *preserve-cucumber-vinegar* cannot be reduced to relations between word pairs, and it is better framed as relation between the instrument and the verb-patient combination (VP). Psycholinguistic experiments have shown that representing such multi-way lexical relations — relations that involve three or more words — is an important component for understanding transitive sentences and complex language (Rayner, Warren, Juhasz, & Liversedge, 2004; Bicknell, Elman, Hare, McRae, & Kutas, 2010).

The productivity of natural language means that humans do not observe all possible instances of multi-way relations (Fodor & Pylyshyn, 1988). However, humans are capable of making useful inferences about combinations of familiar lexical items that they have never encountered. For instance, upon learning the multi-way relation *preserve-cucumber-vinegar*, one may generalize this knowledge to unobserved sentences like *John preserves pepper with vinegar* to judge how well the instrument fits with the VP *preserve pepper*. How can this be performed from a computational perspective? There are two potential approaches: (1) Identify known phrases that are similar to *preserve pepper* at the phrase level, and make an inference based on the relatedness between the proxy phrase and the instrument *vinegar* (Figure 1a). (2) Decompose *preserve pepper* into words, identify similar words that were previously observed in identical syntactic positions, and combine the relatedness of the proxy verb-instrument pair and the proxy patient-instrument pair to make an inference (Figure 1b). In this work, we refer to the former approach as holistic generalization, and the latter as **compositional generalization**.

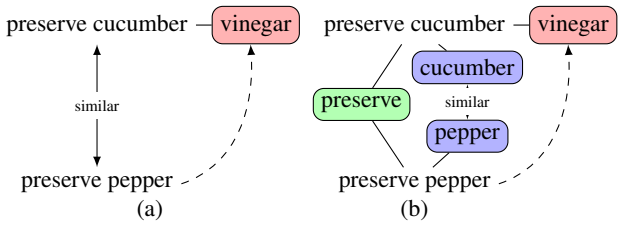


Figure 1: The holistic (a) and compositional (b) perspective on semantic inference. The task is inferring the plausible but omitted instrument *vinegar* given the verb phrase *preserve pepper* which was never observed with *vinegar*. Solid and dashed lines indicate familiar, and inferred relations, respectively. Similarity relations are labeled 'similar'.

While holistic generalization can be enormously useful, it may fall short relative to compositional generalization in some situations. First, complex phrases tend to occur less often than the words that make up those phrases. This means that constructing representations of words requires less data compared

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

to phrases. Because compositional generalization relies primarily on representations of individual words and not phrases, the representations used by compositional generalization tend to be more available and useful in a wider range of situations. Another advantage of the compositional approach is the ability to compute relatedness on the fly. It has been argued that natural language is productive in the sense that infinite meaningful sentences can be generated from a finite vocabulary (Fodor & Pylyshyn, 1988; Fodor & Lepore, 2002). The compositional approach provides a way to judge the semantic plausibility of novel lexical combinations based on a representation of lexical relations formed from finite language input. Psycholinguists have shown that humans rely on this computational capability to understand complex sentences (Bicknell et al., 2010; McRae, Hare, Elman, & Ferretti, 2005).

Despite these advantages, the intricacies of the computational mechanisms that enable compositional generalization in both humans and artificial systems are not thoroughly characterized. What representational structure and processing mechanisms are needed to learn multi-way lexical relations, and to generalize such knowledge to novel lexical combinations? In this work, we evaluate the abilities of various distributional models to perform compositional generalization, in order to hone in on the representational and processing mechanisms needed for compositional generalization. We first discuss a recently introduced graphical distributional model, the Constituent Tree Network (CTN), which performed perfectly in compositional generalization (Mao, Huebner, & Willits, 2022). The CTN is an important tool as its internal operations are fully understood, and are governed by a small set of highly interpretable equations. This is a stark contrast to most other distributional models, especially those that rely on computation of similarity in vector space. However, the advantage of most existing distributional models is that they operate on raw text, as opposed to syntactic trees (which the CTN relies on). We examined three distributional models which are based on raw text: Hyperspace Analogue to Language (HAL), Recurrent Neural Networks (RNN), and a miniature version of GPT-2 (mini GPT-2). Having found that only mini GPT-2 achieved consistently accurate results in a task that evaluates compositional generalization, we argue that a hierarchical attentional structure analogous to the constituent trees encoded in the CTN may have emerged in mini GPT-2 over the course of training. We discuss the implication of these findings to the study of human semantic development and language modeling.

Models

The Constituent Tree Network

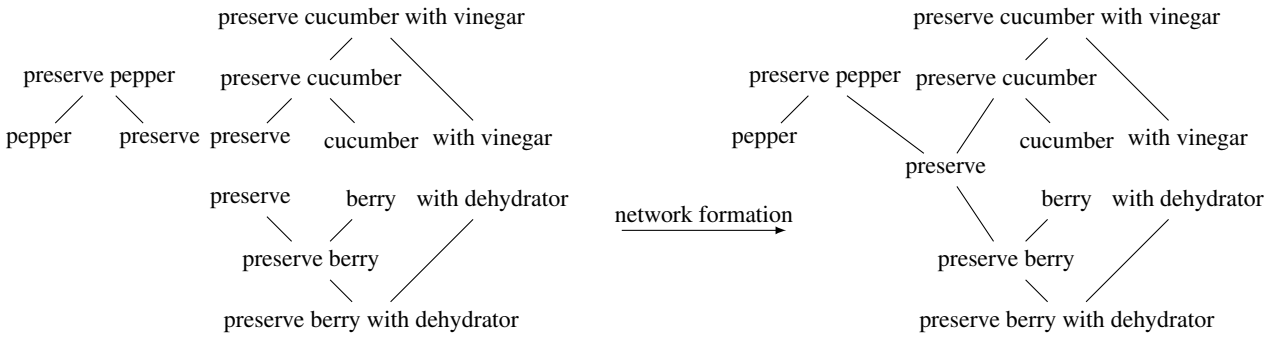
The Constituent Tree Network (CTN) encodes distributional language data in a graphical format. Given a corpus of constituency-parsed sentences (Figure 2a), a semantic network is constructed by joining the constituent parse-trees at shared nodes (Figure 2b). As a result, constituent structure is explicitly encoded in the topology of the network. By joining parse-trees into a single network, the CTN is able to leverage phrasal nodes to infer the relation between structures that did not occur in the corpus, e.g.. *preserve-berry-vinegar* and *preserve-cucumber-dehydrator* (Figure 2b). Critically, constituents that belong to the same phrase are connected via higher-order phrasal nodes, which helps to differentiate relations between words and phrases in terms of graph-theoretic metrics. For instance, after constructing the network in a way that preserves phrasal nodes, the constituent *cucumber* (part of the phrasal node *preserve cucumber*), is closer to *vinegar* than to *dehydrator* by 2 steps. To compute graded graphical semantic relatedness, we adopted the spreading-activation algorithm used by Mao, Huebner, and Willits (2023), and known to account for a diverse range of human semantic judgments (De Deyne, Navarro, Perfors, & Storms, 2016).

Previous work demonstrated that constituent structure is essential for networks built from text to perform compositional generalization (Mao et al., 2022). This is not surprising given a rich literature that has supported the notion that constituent structure is critical to making semantic inferences involving complex phrases. As mentioned, this is accomplished by the presence of nodes that represent complex phrases in the CTN; phrasal nodes enrich network connectivity so that semantically similar phrases become more tightly connected. That said, a noteworthy disadvantage of prior work on the CTN is that it does not specify how information about constituency is derived from raw text.

Assuming such a sophisticated structure as a modeling prerequisite not only makes it difficult to use the model in applied settings, but also distances it from studying human semantic development. As a remedy, this work investigates the CTN in comparison to distributional models trained on raw text. Particularly, what representational structures emerge when training on un-parsed linguistic input, and how do such structures relate to the constituency explicitly provided to the CTN?

Distributional Models Processing Raw Text

Word-Encoding Models The simplest class of distributional models derive representations of linguistic units by counting co-occurrences between linguistic elements in a raw corpus. HAL (Lund & Burgess, 1996), alongside LSA (Landauer & Dumais, 1997) and BEAGLE (Jones & Mewhort, 2007) are examples of this class of models. Individual words are represented as vectors in a high-dimensional space, enabling the quantification of the similarity between two words (Landauer & Dumais, 1997) by operations in the vector space. Models in this class have been used in cognitive science, including predicting human behavior in linguistic tasks (Baroni, Dinu, & Kruszewski, 2014; Evert & Lapesa, 2021), and providing plausible mechanisms for word learning in humans (Mandera, Keuleers, & Brysbaert, 2017; Lupyan & Lewis, 2019). We refer to models of this type as word-encoding models, and investigate the HAL model as a representative. While the word-encoding models are useful for modeling pair-wise lexical relations, they may struggle to learn multi-way relations, e.g., *preserve-cucumber-vinegar*. How might HAL correctly select *vinegar* instead of *dehydrator* as the best-fitting instru-



(a) Constituent parse trees **(b) CTN**

Figure 2: Formation of the network structure in the Constituent Tree Network (CTN) given the mini corpus *preserve pepper, preserve cucumber with vinegar, preserve berry with dehydrator*. **(a)** The input to the CTN consists of constituency-parsed trees for sequences in the mini corpus. **(b)** The network structure of the CTN is formed by joining the constituent trees at shared nodes.

ment? One way is to find an effective function to compose the vector representations of the verb and the patient into a vector representation of the VP. Then the VP-instrument relatedness can be evaluated. Historically, this approach has been less successful, due to the difficulty of manually selecting a composition function (Mitchell & Lapata, 2010).

Recurrent Neural Network One way to overcome the need to manually select or construct a composition function is to train models to learn an implicit function on their own. The class of models that excel at this are recurrent neural networks (RNNs) (Elman, 1990). By being trained to predict which words are likely to come next, such models learn implicit composition functions for how to combine words in their input. These functions are encoded in the network's weights, and are gradually tuned to the statistics of the training corpus. RNNs have been extremely useful in accounting for a broad range of behavioral phenomena in the psycholinguistic literature (Chang, Dell, & Bock, 2006; Elman, 1990, 1991; Huebner & Willits, 2018). Despite the advantage of RNNs over word-encoding models in learning which words tend to go together in a sequence, it can be a challenge for RNNs to generalize such knowledge compositionally to novel phrases. Since RNNs tend to form a holistic representation of a sequence, they may only compare *preserve pepper* and *preserve cucumber* as a whole (Figure 1a), instead of constructing the phrasal similarity in terms of word-word similarities (Figure 1b). Because of this, an RNN that has never previously observed *preserve pepper with vinegar* may struggle to infer the most plausible instrument when given only *preserve pepper*.

Transformer While trained on similar prediction tasks used for RNNs, Transformers do not forcibly compress representations of individual items into one holistic representation of the input sequence, but organize and process tokens in parallel, using self-attention (Vaswani et al., 2017). Rather than learning a single transition function, (e.g., the recurrent weights for *preserve cucumber* to predicting *with vinegar* in the RNN), a Transformer learns how to combine arbitrary tokens in the input using unique keys and queries for each token, e.g., *with* attends *cucumber* and/or *preserve* across different attention layers. The representations of individual tokens (i.e., words) are combined as a function of the match between the query of one token and the key of another. One major advantage of self-attention is that integration of information is unique to a given pair of words and is sensitive to the context in which those words have occurred. In this way, Transformers may keep the track of the individual identities of all the words in the model's input, allowing for flexible combination of lexical representations.

We made the following predictions. Word-encoding models like HAL will fail to learn multi-way lexical relations due to the lack of an effective composition function. In contrast, RNNs should be able to learn multi-way relations by composing lexical items into a holistic representation, at the cost of representing individual items which leads to generalizing learned relations to novel lexical combinations in a compositional manner. Finally, the Transformer based mini GPT-2 should succeed in learning and compositional generalization, because its architecture enables learning of representations for each element in a sequence as well as how to flexibly combine them to represent larger chunks of language.

Materials and Methods

All models were trained on an artificial corpus that incorporated controlled semantic constraints. This allow us to concludes about the model mechanisms, which would not have been possible if using a naturalistic corpus. Each sentence in the artificial corpus is of the form agent-verb-patient-instrument, e.g., *John preserve cucumber (with) vinegar*. For each trained model, all pairwise semantic relatedness scores between VP pairs and instruments were computed, and evaluation is based on how well a model predict the structurally-licensed instrument of certain VPs.

Corpus

The artificial corpus is based on a set of 48 verbs and sets of nouns that define possible arguments for each verb. Each verb is associated with three nouns in the agent position, three nouns in the patient position, and zero, one, or two nouns in instrument position (depending on verb-type, defined below). The agent nouns (*John*, *Mary*, and *Fatima*) are not verb-specific, while the patient and instrument nouns are verb-specific. In total, there are three possible nouns in the agent position, 48 nouns in the patient position, and 24 nouns in the instrument position, and the vocabulary consists of 123 word types, not counting the preposition *with* (that optionally preceded instruments) and the period symbol marking sentence boundaries. Words are bound to specific positions; for instance, words that occur in the agent position never occur in the patient position and vice versa. Sentences used for training were derived by iteratively sampling from all possible (576) agent-verb-patient-instrument combinations over 400 blocks. In each block, one of 48 verbs was selected without replacement, and arguments were filled by choosing among legal candidates randomly. For the purpose of statistical comparison, 30 instances of each model were trained, each on a unique corpus generated with a different random seed.

Patients Each patient (e.g. *cucumber*, *berry*) belongs to one of 16 semantic categories, such as FRUIT and VEGETABLE. Each category consists of 3 patients, and defines the verbs with which a patient could co-occur. For each category, two category members were designated as 'control' patients, and one member was designated as the 'experimental' patient (Table 1). The experimental patient never occurs with an instrument, while control patients always occur with an instrument in the training corpus. Thus, the corpus included *Mary preserve cucumber with vinegar*, and *Mary preserve pepper* instead of *Mary preserve pepper with vinegar*. In this way, VP-instrument relations are expressed in the pairing between control VPs and instruments. The critical test was whether models generalized these relations to the experimental VPs that had not been seen with these instruments.

Table 1: Two of 16 patient categories and their members. Experimental patients are in bold-face.

Patient Category	Members		
VEGETABLE	potato	cucumber	**pepper**
FRUIT	apple	berry	**orange**

Verbs There are four verb types in the corpus. Type-0 verbs only occur with patients that belong to the same category, to provide distributional evidence for similarity among patients belonging to the same category. Similarly, type-1 verbs can occur with two related patient categories (e.g. FRUIT and VEGETABLE), and their purpose is to induce a distributional semantic hierarchy for patients. The type-0 and type-1 were created for the sole purpose of forming the semantic structure in our corpus, and therefore never occur with instruments. In contrast, Type-2 and type-3 verbs occur with instruments and are therefore used during evaluation. The difference is that type-2 verbs can only occur with one instrument, whereas type-3 verbs can occur with two. The instrument that can occur with type-3 verbs is contingent on the choice of patient; for instance, while *'preserve cucumber'* can only occur with *'vinegar'*, *'preserve berry'* can only occur with *dehydrator*. In this way, we created (i) systematic dependencies between the verb, patient and instrument by patient category, and (ii) a semantic taxonomy of the patients basing on their distributional associations to the verbs (e.g., for *pepper*, *cucumber* and *potato* are most similar patients as they are in the same category; this is followed by FRUIT such as *berry* that shares partial verbs with VEGETABLE, and finally all other patients. We refer to the two categories which share verbs, e.g. FRUIT and VEGETABLE as forming a 'super-category'. There are 8 super-categories in total, and Table 2 shows the sentences for each verb-type in a particular super-category.

Model Training

Three types of models were investigated, the HAL model (Lund & Burgess, 1996), the Long Short-Term Memory model, or LSTM (Hochreiter & Schmidhuber, 1997), and a miniature version of GPT-2 (Radford et al., 2019). We trained 30 instances of each model, varying the random seed used to generate the corpus each time.

In the HAL model, vector representations of the words (patients, verbs and instruments) are first formed.Then the vector representations of VPs are composed from word representations, with two composition functions: addition and point-wise multiplication. The semantic relatedness between the VP and the instrument was computed as the cosine similarity between their vector representations. To form word vectors in HAL, we tuned on minor parameters that may affect the performance in the downstream tasks (Bullinaria & Levy, 2007, 2012). We identified the best performing HAL model by tuning the window weight (flat, **linear**), window type (forward, backward, **sum**, concatenated), and the number of singular dimensions (16,22,24,... **30**,32,...,36,64). Bold-face indicates the chosen parameters. The window size for tracking co-occurrences was constrained such that all words in a sentence are available for analysis — other window sizes were not considered.

For the LSTM and the mini GPT-2, we identified one highest-performing hyper-parameter configuration for each model after extensive tuning on the generalization task. Hyper-parameter search was restricted to 2-layer architectures, and 64 and 32 hidden units performed best for the LSTM and the mini GPT-2 model respectively. In keeping with the format of the task used for training, we operationalized the relatedness of a VP-instrument pair by the network's activation at the output layer. To obtain the network's prediction of the best fitting instrument, we provided the model with sentences like *John preserve pepper with*, and chose the instrument with the highest activation as the model's prediction. This is in accordance with previous studies where constraints on predic-

Table 2: Example sentences from the artificial corpus, for 2 patient categories (1 super-category) only. Each category is associated with 4 verb types. Type-2 and 3 verbs always occur with instruments except when patient is experimental (indicated by bold-face).

Patient Category	type-0	type-1	type-2	type-3
VEGETABLE	J dice cucumber J dice potato J dice **pepper**	J ferment cucumber J ferment potato J ferment **pepper**	J grow cucumber with fertilizer J grow potato with fertilizer J grow **pepper**	J preserve cucumber with vinegar J preserve potato with vinegar J preserve **pepper**
FRUIT	J dice berry J dice apple J dice **orange**	J pick berry J pick apple J pick **orange**	J spray berry with insecticide J spray apple with insecticide J spray **orange**	J preserve berry with dehydrator J preserve apple with dehydrator J preserve **orange**

tive processing are considered to reflect knowledge of typical events (McRae et al., 2005). In addition, RNNs tend to ignore the context on the left. In our artificial corpus, the patients are categorized by the verb before them, so that the ignorance of the left-side semantic constraints may affect the model's performance. To rule out this potential effect, we further tuned and trained the same LSTM model on corpora with reversed sentences. To be more specific, for a sentence like *John preserve cucumber with vinegar* in the original corpus, we included both the sentence and its reversed sequence *vinegar with cucumber preserve John*. We report the performance of LSTMs trained on the original and the augmented corpus.

Evaluation Tasks

After training, we evaluate each model's knowledge of the VP-instrument relations. In each super-category, there are 12 'instrument-relevant' VPs, like *grow cucumber* and *preserve berry* that associate to instrument, and 4 instruments (Table 1), resulting in a total of 96 VPs and 32 instruments in the corpus. For each VP, we computed the semantic relatedness between the VP and all 32 instruments, and ranked the instruments by the model's relatedness score. A hit is recorded if a model assigns the largest relatedness score to the pair that contains the structurally licensed instrument.

There is only one structurally licensed instrument for each VP. A model's ability to choose the one is used as a measure for its ability to learn and/or generalize lexical relations from distributional statistics. Which instrument is structurally licensed depends on the VP. We focus on the 48 VPs that include type-3 verbs, e.g., *preserve*, among which 32 VPs have been allowed to co-occur with instruments in the corpus, and 16 VPs did not. We reserved the first group of VPs to evaluate the ability of models to learn multi-way relations, and the latter group of VPs to evaluate their ability to generalize the learned knowledge to novel lexical combinations.

In the learning tasks, the structurally licensed instrument is the one that co-occurred with a given VP in the corpus, i.e. *preserve cucumber - vinegar*. The model needs to combine the clues from *preserve* and *cucumber* to pinpoint on the licensed instrument *vinegar*. If the model only pay attention to the verb *preserve*, it would be distracted by the competitor instrument *dehydrator*, which also associated with the verb, but only when the patient is in FRUIT, e.g., *preserve **berry** with dehydrator*. Alternatively, if the model only focused on the patient, it might fall prey to the instrument attached to the patient but under the competing type-2 verb, e.g., *fertilizer* for ***grow** cucumber*. In the generalization tasks, the structurally licensed VP is the one that was associated with the verb and the category of the given patient, i.e., *preserve pepper - vinegar*. To succeed in the task, the model needs to at first ace the learning task, i.e., knowing *preserve cucumber -vinegar*. On top of that, the model must capture the distributional similarity between *cucumber* and the experimental patient *pepper*, and then infer that *preserve pepper* should be similar to *preserve cucumber*, to select on the licensed instrument (*vinegar*).

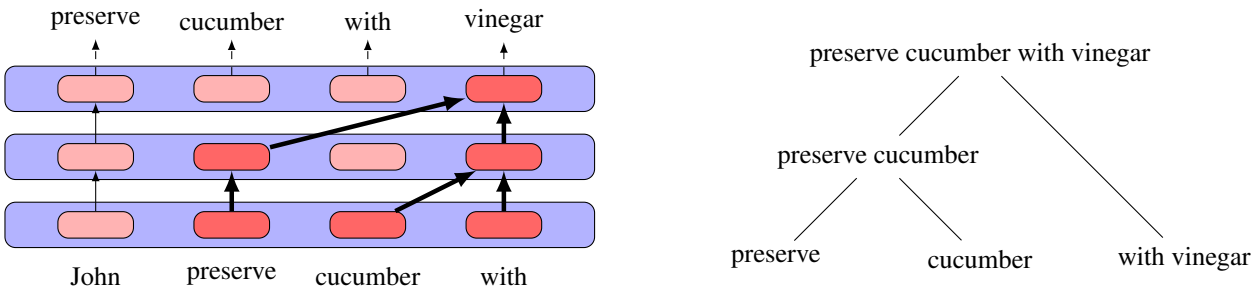
Results and Analysis

All results are summarized in Table 3. The HAL models performed poorly (invariant to the composition function) in both tasks. Critically, they failed to learn the multi-way relations implicit in the corpus, e.g., the relation between the VP *preserve cucumber* and *vinegar*. Further analysis showed that HAL models consistently confused the two instruments associated with the same super-category. In other words, HAL models predicted that *vinegar* and *dehydrator* are equally likely instruments for *preserve cucumber* and *preserve berry*. The same conflation was also observed in the generalization task.

Table 3: Accuracy of inferring the structurally-licensed instruments in the learning and generalization tasks. Accuracies are averages across 30 seeds.

	Learning	Generalization
HAL-addition	0.25 (0.07)	0.28 (0.18)
HAL-multiplication	0.27 (0.07)	0.23 (0.12)
LSTM	0.84 (0.28)	0.26 (0.15)
LSTM, add_reversed	0.86 (0.29)	0.57 (0.29)
Mini GPT-2	**1.00** (0.00)	**0.87** (0.15)

As predicted, the LSTMs learned the multi-way relations presented in the input, but failed to generalize. While performance was enhanced with the addition of reversed sentences to the training corpus, accuracy was nowhere near optimal (100% accuracy). Further analysis showed that the most frequent error made by LSTMs (19% of all errors) is predicting an instrument that belongs to a related category (e.g., FRUIT instead of VEGETABLE). In contrast, mini GPT-2 achieved perfect



(a) Transformer (b) Constituent Tree

Figure 3: Structural representations in the Transformer and the CTN. **(a)** We propose that the Transformer learned a pattern of self-attention across layers that resembles a hierarchically structured parse tree of the input sentence. **(b)** The corresponding constituent tree used to build the CTN.

accuracies in our learning tasks and near-perfect accuracies during generalization. These results support our hypothesis that the parallel architecture of Transformers is better suited for compositional generalization than recurrence in RNNs. We expand on this point below and argue that a hierarchical structure in terms of learned attentional weights might have emerged in the Transformer through the process of learning the distributional regularities in our artificial corpus.

Discussion

This work examined the ability of distributional models to perform compositional generalization — transferring knowledge of familiar multi-way lexical relations to novel combinations. Our results show that word-encoding models like HAL struggle to learn multi-way relations, resonating with previous work showing that it is difficult to construct a single function to effectively compose word representations (Mitchell & Lapata, 2010). While the LSTM learned an implicit composition function for combining next-words and their linguistic context, it failed when generalizing to novel combinations. Finally, the Transformer-based mini GPT-2 succeeded in both learning and generalization tasks.

How do we explain the success of mini GPT-2 in contrast to the LSTM? Inspired by the previous finding that the CTN succeeds in compositional generalization with its explicit encoding of constituent structure (Mao et al., 2022), we argue that a similar hierarchical structure emerged in the self-attention weights of the Transformer architecture. As illustrated in Figure 3, each layer in the Transformer may correspond to a depth in a hierarchically structured parse tree. Hypothetically, the token *with* could attend to *cucumber* in layer 1 to form a composite representation in layer 2. The composite representation could, in turn, attend to *preserve* to form a representation of the VP in layer 3. Recall that self-attention enables Transformers to combine token representations at arbitrary time steps. At successively higher layers, these representations become more contextualized, capturing information about increasingly larger chunks of the input. This ability to flexibly combine lexical representations is one proposal for how a system can make semantic inferences after exposure to finite language inputs. Lacking self-attention, RNNs must compress all information about an unfolding sequence into a single hidden layer representation that is forcibly updated at each time step.

Taken together, the performance of the CTN and mini GPT-2 support our argument that distributional models can enrich their semantic processing with information about constituent structure, and that this information can emerge automatically as a consequence of ingesting raw text. This phenomenon not only provides models the ability to generalize multi-way relations to novel lexical combinations, but could be used to explain how humans may make accurate semantic plausibility judgements about sentences they have never heard. Further, the presence of this ability is often interpreted in terms of the productivity of human language (Fodor & Pylyshyn, 1988). In this regard, our work provides a motivation for investigating productivity as an emergent result of training connectionist systems. Moreover, our interpretation can explain why state-of-the-art LLMs can generate semantically plausible text that strongly resembles language produced by humans.

Follow-up work is required to address limitations of the current study. First, diagnostic research is needed to scrutinize our claim that the attentional patterns that mini GPT-2 learned are indeed hierarchical, and analogous to constituent trees. Future work on opening the black box of Transformers should make use of toy corpora so that semantic regularities that are used for testing can be more strictly controlled. The second direction involves behavioral experiments that explore the interplay between syntax and semantics in language development. Our modeling results suggest that exposure to linguistic distributional information may spur the development of early syntactic abilities, and that syntactic abilities, in turn, enrich semantic development. More work is needed to establish at which age children start to track and use multi-way lexical relations. Result of such studies can be aligned with hallmarks of syntactic development. Lastly, the combination of modeling and behavioral work will be an important strategy in advancing our understanding of syntax-semantics interactions in computational models and human semantic cognition.

References

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 238–247).

Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of memory and language*, *63*(4), 489–505.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*, 510-526.

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

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological review*, *113 2*, 234-72.

De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2016). Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General*, *145*(9), 1228.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Elman, J. L. (1991). Distributed representations, recurrent nets, and grammatical structure. *Mach. Learn.*, *7*, 195–225.

Evert, S., & Lapesa, G. (2021). FAST: A carefully sampled and cognitively motivated dataset for distributional semantic evaluation. In *Proceedings of the 25th conference on computational natural language learning* (pp. 588–595). Online: Association for Computational Linguistics.

Fodor, J. A., & Lepore, E. (2002). *The compositionality papers*. Oxford University Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3-71.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735-1780.

Huebner, P. A., & Willits, J. A. (2018). Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*, *9*, 133.

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

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*(2), 211.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*, 203–208.

Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: the role of language in semantic knowledge. *Language, Cognition and Neuroscience*, *34*(10), 1319–1337.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78.

Mao, S., Huebner, P., & Willits, J. (2022). Compositional generalization in a graph-based model of distributional semantics. In (pp. 1993–1999). (Publisher Copyright: © 2022 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY); 44th Annual Meeting of the Cognitive Science Society: Cognitive Diversity, CogSci 2022 ; Conference date: 27-07-2022 Through 30-07-2022)

Mao, S., Huebner, P. A., & Willits, J. A. (2023). Spatial versus graphical representation of distributional semantic knowledge. *Psychological Review*.

McRae, K., Hare, M., Elman, J. L., & Ferretti, T. R. (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, *33*, 1174–1184.

Mitchell, J., & Lapata, M. (2010). Composition in distributional models. *Cognitive science*, *34*(8), 1388–1429.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004, November). The effect of plausibility on eye movements. *J. Exp. Psychol. Learn. Mem. Cogn.*, *30*(6), 1290–1301.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).