Compositional Generalization in a Graph-based Model of Distributional Semantics

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

A critical part of language comprehension is inferring omitted but plausible information from linguistic descriptions of events. For instance, the verb phrase *'preserve vegetable'* implies the instrument *vinegar* whereas *'preserve fruit'* implies *dehydrator*. We studied the ability of distributional semantic models to perform this kind of semantic inference after being trained on an artificial corpus with strictly controlled constraints on which verb phrases occur with which instruments. Importantly, the ability to infer omitted but plausible instruments in our task requires compositional generalization. We found that contemporary neural network models fall short generalizing learned selectional constraints, and that a graph-based distributional semantic model trained on constituency-parsed data and equipped with a spreading-activation procedure for calculating semantic relatedness, achieves perfect performance. Our findings shed light on the mechanisms that give rise to compositional generalization, and using graphs to model semantic memory.

Keywords: distributional semantics; semantic inference

Introduction

In language, the meaning of the whole is often determined by some function of its parts. For example, the set of plausible continuations of a *'John preserves the pepper with _'* is constrained by the selectional preferences (Katz & Fodor, 1963) of the verb (e.g. *preserve*), and the theme (e.g. *pepper* in the sentence). While behavioral evidence for sensitivity to multiple constraints during language processing abound (Rayner, Warren, Juhasz, & Liversedge, 2004), there is a long-standing debate concerning how such constraints are represented and integrated during learning and generalization (McRae, Hare, Elman, & Ferretti, 2005). In the connectionist approach, semantic constraints on processing are typically considered to emerge from open-ended interactions among linguistic units without limits on the level at which such units are represented (e.g. word, phrase, sentence). In contrast, the compositional approach emphasizes principled decomposition of such constraints (e.g. independent constraints due to the verb, and due to the theme). The primary difference (see Figure 1) is that the connectionist approach permits relations among complex expressions (e.g. *'preserve pepper'* $\leftrightarrow$ *'preserve cucumber'*), whereas, in the compositional approach, such relations are decomposed into smaller relations (e.g. *cucumber* $\leftrightarrow$ *pepper*), and encoded separately.

This difference has potential consequences for how models generalize to novel stimuli. When a model has learned that *pepper* is semantically similar to *cucumber* or another theme

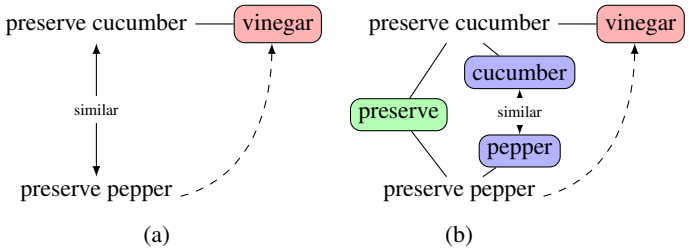

Figure 1: The connectionist (a) and compositional (b) perspective on semantic inference. The task is inferring the plausible but omitted instrument *vinegar* given the verb phrase *'preserve pepper'* which was never observed with *vinegar*. Solid and dashed lines indicate familiar, and inferred relations, respectively. Similarity relations are labeled 'similar'. The compositional approach emphasizes relations between constituents (e.g. themes, shown in blue) while the connectionist approach also considers relations between larger chunks of language.

previously associated with *vinegar*, the model may perform so-called compositional generalization, to infer that *vinegar* is a plausible continuation of *'preserve pepper'* despite never having observed the two during training. If, instead, a model has only learned the similarity between the verb phrases *'preserve pepper'* and *'preserve cucumber'*, the model must rely on phrasal similarity instead of compositional generalization to make the same inference. While both the connectionist and compositional approaches may be, in principle, able to account for compositional generalization, connectionist models often do not converge on the needed lexical similarity structure that would allow them to perform this kind of generalization. To better understand why some models succeed and others fail, we compared models in both traditions in a task that explicitly requires compositional generalization. By comparing a novel graph-based distributional semantic model to existing connectionist models, our work sheds light on what data structures and processes are better suited for generalizing knowledge of individual words to novel word combinations.

While connectionist language models have proven successful at predicting upcoming words in prediction tasks, they often fall short when generalizing under conditions too dissimilar from those during training (Lake & Baroni, 2018; Kim &

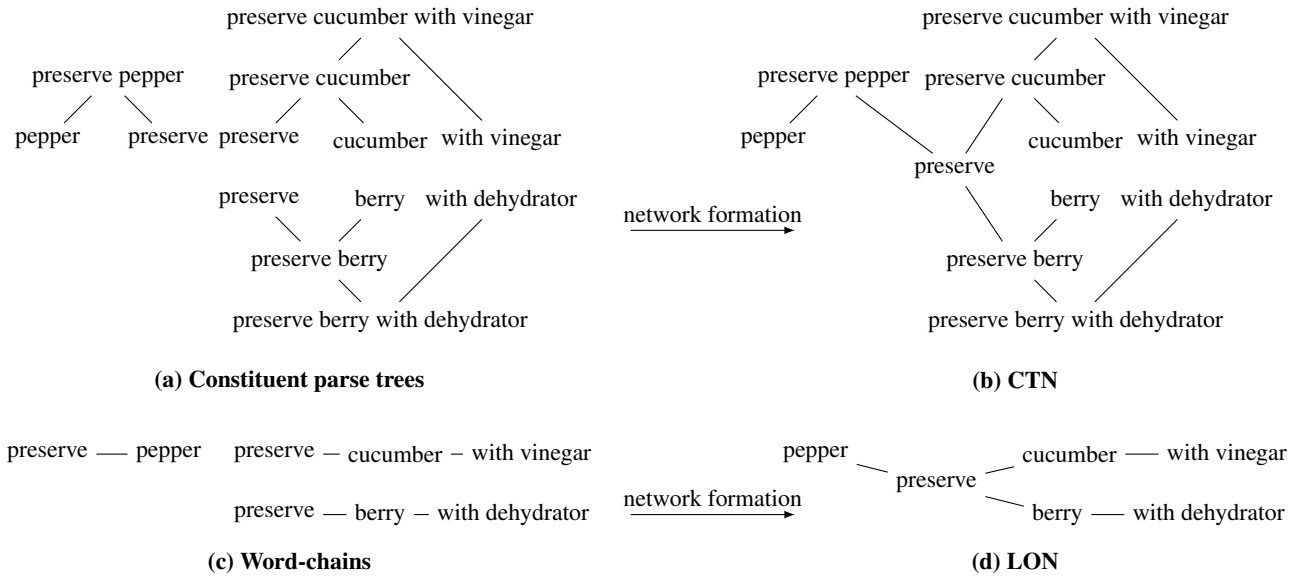

Figure 2: Formation of the network structure in the Constituent Tree Network (CTN) and the Linear Order Network (LON) given the mini corpus *'preserve pepper', 'preserve cucumber with vinegar', 'preserve berry with dehydrator'*. **(a)** The input to the CTN consists of constituency-parsed trees for sequences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

Linzen, 2020; but see Russin, Jo, O'Reilly, & Bengio, 2020). One explanation is that such models often memorize larger chunks of language without learning how they relate to similar chunks (Arnon & Snider, 2010; Elman, 2014). For instance, it is possible that when a neural network has encoded the sequential dependency between *'preserve cucumber'* and *vinegar*, it does so in a way that does not readily extend the same knowledge to similar phrases such as *'preserve pepper'*. The success of inferring that *vinegar* is a plausible continuation for *'preserve pepper'* likely depends on whether or not a model has learned similar representations for these two phrases. The likelihood of this, in turn, rests on the statistical properties of the training data, which is not always suitable for the induction of compositional representations.

In contrast, compositional approaches typically eschew relationships among unanalyzed wholes in favor of relations among smaller components. Consider again the sentence *'John preserves the pepper with _'*. To infer the omitted instrument, compositional approaches prescribe a principled computational procedure whereby selectional preferences due to (i) the verb and (ii) theme are applied in a step-wise fashion. For instance, to correctly infer the target instrument *vinegar*, despite never having observed that instrument in that context, a compositional system would first use the verb *preserve* — independently of the theme — to access verb phrases which were previously associated with an instrument (e.g. *'preserve cucumber'* occurred with *vinegar*; *'preserve berry'* occurred with *dehydrator*). The second step involves the theme. To correctly choose the instrument, the system would exploit the semantic similarity between the given theme and the two candidate themes. Because the similarity between *pepper* and *cucumber* is greater than that between *pepper* and *berry*, the system could infer that *vinegar* is a more plausible instrument for *'preserve pepper'*. In such a system, the selectional preferences of the parts (the verb, and theme) separately contribute to the selectional preference of the whole. This idea follows from the principle of compositionality (Fodor & Lepore, 2002; Carnap, 1947), and has previously been investigated for modeling meaning combination (Mitchell & Lapata, 2010).

Models that approach semantic inference from the perspective of compositionality have varied in how their representations are formed. Traditionally, such models have included hand-crafted rules (Fodor & Lepore, 2002). More recent models derive their knowledge from corpus data via domain-general learning algorithms (Mitchell & Lapata, 2010), similar to the connectionism approach. To control for variation along this dimension, we compared the compositional and connectionist approach within the same framework — distributional semantic modeling — in which models encode the meaning of linguistic units in terms of their relation to other units in linguistic data. Specifically, we examined two popular connectionist distributional semantic models, the simple recurrent neural network and the Transformer. We contrasted them with a novel graphical distributional model, which we refer to as the Constituent Tree Network. The graphical model is both compositional and distributional as it explicitly encodes constituent structure of distributional linguistic data in a network (Figure 2b). While connectionist systems are in principle capable of

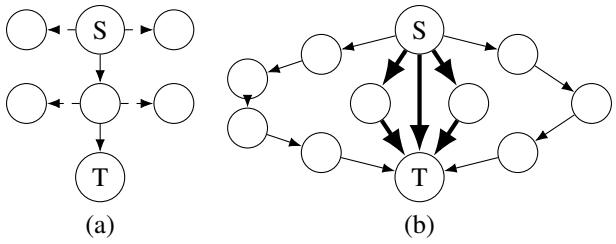

Figure 3: Spreading-activation based measure of semantic relatedness between nodes in a hypothetical network. S and T indicate source and target nodes, respectively. (a) Because activation diffuses along all edges (dashed lines), longer paths lead to less activation arriving at T. (b) Multiple paths from S to T may exist; only first and second shortest paths (thick lines) are considered when calculating relatedness.

solving challenging semantic inference tasks, we hypothesized that the Constituent Tree Network may surpass connectionist models on some tasks requiring out-of-distribution generalization.

The Constituent Tree Network

We propose a novel distributional semantic model, which we call the Constituent Tree Network (CTN). The Constituent Tree Network is a semantic network that encodes distributional language data in a graphical format we hypothesized is useful for semantic tasks that require compositional generalization. Given a corpus of constituency-parsed sentences (Figure 2a), where nodes in the parse-trees represent constituents (e.g. words, phrases and sentences), and edges represent constituency relations, the network is constructed by joining the constituent parse-trees at shared nodes (Figure 2b). As a result, constituent structure is explicitly encoded in the network topology; constituents that belong to the same phrase are connected via higher-order phrasal nodes. These phrasal nodes are helpful for encoding the selectional preference of a whole phrase. For instance, the verb phrase *preserve cucumber* is closer to *vinegar* than to *dehydrator* by 2 steps in the network. However, without the phrasal node *preserve cucumber*, graphical distance cannot be used to discriminate *vinegar* among other instruments which are less related, but nonetheless associated with the verb or theme (Figure 2d). By joining parse-trees into a single network, the Constituent Tree Network is able to leverage phrasal nodes to infer the relation between structures that did not occur in the corpus. For instance, *preserve pepper*, which did not co-occur with any instrument in the corpus, becomes indirectly connected to instruments that did occur.

While graphical distance is typically adopted as a proxy for the relatedness between linguistic units in networks, we implemented a spreading-activation algorithm (De Deyne, Navarro, Perfors, & Storms, 2016) to compute semantic relatedness. Relatedness between a source node (S) and a target node (T) is defined as the activation that reaches T and originates at S. By so doing, relatedness in the network is graded and sensitive to (i) the length of the path connecting node S and T, (ii) the number of paths between S and T (see Figure 3). Activation-spreading based measures on semantic networks have been shown to be a better fit to empirical data, and are better grounded in cognitive theory (De Deyne et al., 2016).

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

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

Methods and Materials

All models were trained on an artificial corpus, in which occurrences of instruments were precisely controlled. This enabled us to draw strong conclusions about model differences, which would not have been possible if using a naturalistic corpus. Each sentence in the artificial corpus is of the form agent-verb-theme-instrument, where agent and theme refer to the semantic roles of the subject, and direct-object, respectively. For each trained model, all pairwise semantic relatedness scores between verb-theme (VP) pairs and instruments were computed, as a proxy for the selectional preferences of VPs on instruments. Evaluation is based on how well a model's rank-ordering of relatedness scores matches the expected rank-ordering implicit in the training data. Relatedness measures for graphical and connectionist models are obtained using activation-spreading, and next-word prediction error, respectively.

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 theme position, and zero, one, or two nouns in instrument position (depending on verb-type, defined below). The set of agent nouns is not verb-specific, and includes *John*, *Mary*, and *Fatima*. In contrast, nouns that can occur in the theme and instrument positions are verb-specific. In total, there are three possible nouns in the agent position, 48 nouns in the theme position, and 24 nouns in the instrument position. In total, the vocabulary consists of 123 word types, not counting the preposition *with* (which optionally preceded instruments) and the period symbol, which marks sentence boundaries. We varied whether *with* is inserted prior to each instrument, and only report results that provide better performance. Words are bound to specific positions; for instance, words that occur in the agent position never occur in the theme position and vice versa. Sentences used for training were derived by iteratively sampling from all possible (576) agent-verb-theme-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, multiple instances of each model were trained, each on a unique corpus generated

with a different random seed. This introduced variation in learning outcomes for the graphical models, which would otherwise produce the same results given identical input.

Themes Each theme (e.g. *cucumber*, *berry*) belongs to one of 16 semantic categories, such as FRUIT and VEGETABLE. Each category consists of 3 themes, and defines the verbs with which a theme could co-occur. For each theme category, two category members were designated as 'control' themes, and one member was designated as the 'experimental' theme (Table 1). The difference is that an experimental theme never occurs with an instrument, while control themes always occur with an instrument in the training corpus. Thus, while the sentence *'Mary preserve cucumber with vinegar'* was seen during training, the sentence *'Mary preserve pepper with vinegar'* was not; instead the corpus was limited to sentences like *'Mary preserve pepper'* where the instrument had been omitted. In this way, the selectional preferences of the VP on instruments are expressed in the pairing between control VPs and instruments. The critical test was whether models generalized their knowledge of these preferences to the experimental VPs that had not been seen with these instruments.

Verbs There are four verb types in the corpus. Type-0 verbs only occur with themes that belong to the same theme category; their purpose is to provide distributional evidence for similarity among themes that belong to the same category. Similarly, type-1 verbs can occur with two related theme categories (e.g. FRUIT and VEGETABLE), and their purpose is to induce a distributional semantic hierarchy for theme nouns. Neither type-0 not type-1 verbs occur with instruments, as these verbs were created for the sole purpose of forming the semantic structure in our corpus. Type-2 and type-3 verbs occur with instruments and are therefore used during evaluation of selectional preferences on instruments. 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 theme; for instance, while *'preserve cucumber'* can only occur with *'vinegar'*, *'preserve berry'* can only occur with *dehydrator*. Example sentences for each verb-type are shown in Table 2.

Model Training and Evaluation

Four classes of models were investigated. We included two connectionist language models, the simple RNN (Elman, 1991) and the Transformer (Vaswani et al., 2017), and two novel graph-based models, the Constituent Tree Network (CTN), and a reduced variant of the CTN, the Linear Order Network (LON). We trained 10 instances of each model, varying the random seed used to generate the corpus each time. After training completed, for each model, we computed all pairwise semantic relatedness scores between type-2 VPs and instruments, and type-3 VPs and instruments. All subsequent analyses are based on these scores.

RNN and Transformer We examined two connectionist language models, the simple RNN (Elman, 1991) and the more recent Transformer (Vaswani et al., 2017). For the latter, we adopted a miniature version of the GPT-2 Transformer architecture (Radford et al., 2019). We identified one highest-performing hyper-parameter configuration for each model after extensive tuning on the selectional preference task. We found that 64 hidden units, and 32 hidden units performed best for the RNN and Transformer, respectively. Hyper-parameter search was restricted to 1-layer architectures. In particular, the Transformer required a significant amount of hyper-parameter tuning to reduce variance over different seeds. After best hyper-parameters were identified, we re-trained all models on corpora generated using different random seeds compared to the one used for tuning. While we observed strong performance of the Transformer during tuning (near 100% accuracy in all conditions), we observed a considerable drop in performance in the generalization condition of Experiment 2 (see Results and Analysis) after re-training on 10 novel random seeds. In keeping with the format of the training task, we operationalized selectional preferences of the VP on instruments in terms of prediction error: Given as input *'John preserve pepper with _'*, we computed the prediction error at the last time step, substituting '_' with an instrument. This is in accordance with previous proposals where constraints on predictive processing are considered to reflect knowledge of typical events (McRae et al., 2005).

CTN and LON Knowledge in the graphical models is encoded in a network consisting of nodes connected by edges. Nodes correspond to lexical or phrasal units in the training corpus, and edges correspond to co-occurrence (LON) or constituency (CTN) relations between units. Training in both models involves converting sentences into graphical form and joining the resulting sub-graphs at shared nodes (Figure 2a-d). The LON is a degenerate version of the CTN in which word chains (words adjacent in the training data) instead of constituency-parse trees are joined during training (Figure 2c-d). In contrast to CTN that encodes phrasal (i.e. higher-order) dependencies explicitly, the absence of phrasal nodes in LON makes it difficult to represent the dependency between a VP and associated instruments (see Figure 2b and 2d for a comparison). Thus, the LON was included to diagnose the effect of ablating constituent structure.

In contrast to vector-based models, where relatedness corresponds to distance in vector-space, relatedness in the CTN and LON is based on a spreading-activation algorithm that closely resembles the procedure described by Mao and Willits (2020). In this work, we extend their measure of lexical relatedness — defined for two words — to relatedness between a phrase and a word, and refer to this extension as 'phrasal relatedness'. To compute phrasal relatedness between a VP and an instrument, the lexical relatedness scores of verb-instrument pairs and theme-instrument pairs were computed separately, and then combined via multiplication. In this way, both constituents of the VP contribute independently to the composite relatedness — in accordance with compositional generalization.

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

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

Experiments

We conducted two experiments to investigate the ability of models to learn the structural relationship between VPs and instruments. In Experiment 1, we evaluated the ability of models to predict the structurally-licensed instrument when the verb alone provides sufficient information. In Experiment 2, the verb alone is not sufficient, and to succeed, a model must use the combined VP — the verb and the theme — to correctly predict the structurally-licensed instrument. In each experiment, we separate performance by control and experimental themes; high performance in the former condition indicates successful encoding of the training data, while high performance in the latter condition indicates successful generalization to out-of-distribution data (unobserved VP-instrument pairs).

Experiment 1: Verb-based Learning and Generalization In Experiment 1, we investigated the ability of models to learn and infer the selectional preferences of VPs with type-2 verbs. Given that type-2 verbs can only occur with one instrument, their selectional preferences on instruments is reducible to a simple lexical dependency — the theme provides no additional information to succeed in this task. In the control condition, we assessed learning by evaluating the ability of models to assign greater relatedness scores to observed VP-instrument pairs compared to unobserved pairs. In the experimental condition, we assessed generalization by evaluating only themes that never occur with instruments in the training data. In both conditions, accuracy indicates how often the structurally-licensed instrument is ranked higher than all other instruments.

Experiment 2: VP-based Learning and Generalization In Experiment 2, we investigated the ability of models to learn and infer the correct selectional preferences of VPs with type-3 verbs. This task is more difficult because each type-3 verb can occur with two instruments in the corpus. This means the correct instrument is a function of both the verb and the theme. For instance, both *vinegar* and *dehydrator* are plausible (i.e structurally licensed) instruments for the verb *preserve*, but themes in the VEGETABLE category exclusively license the former and themes in the FRUIT category exclusively license the latter. On the flipside, the theme alone is not sufficient either, given that themes occur with more than one instrument in the training data. In the control condition, performance reflects learning; accuracy is calculated based on the proportion of times a model assigns the highest semantic relatedness to VP-instrument pairs that were observed in the training data (given all possible VP-instrument pairs made of type-3 verbs and control themes). To perform well in the control condition, a model need only learn the dependency between VPs and instruments directly accessible in the training data.

In the experimental condition, performance reflects generalization; not only must a model encode the dependency between VPs and instruments, but do so in a manner that preserves the similarity between same-category themes (e.g. *cucumber* and *pepper*). Specifically, for each unobserved VP-instrument pair, a model must infer a target rank-ordering of instruments that aligns with the structure of the artificial corpus. For instance, given *'preserve pepper'*, the correct rank-ordering is *vinegar* $>$ *dehydrator* $>$ all other instruments, because *vinegar* co-occurs with VEGETABLE themes, and *dehydrator* co-occurs with semantically related FRUIT themes. High accuracy in this condition requires compositional generalization because a model must substitute the experimental theme *pepper* — never paired with instruments in the corpus — with a control theme from the same theme category without compromising the categorical verb-theme relation. The structure of the corpus was created so that *lexical* semantic models fail the generalization portion of Experiment 2. We confirmed this by training Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013) on our data, and computing phrasal relatedness using element-wise multiplication as in Mitchell and Lapata (2010)[1].

Results and Analysis

All results are summarized in Table 3. We observed that the CTN reached ceiling performance in all four conditions. As expected, constituent structure is essential for compositional generalization, as evidenced by the lower accuracy of the LON in both conditions of Experiment 2. The Transformer scored second-best, but performed considerably worse in the experimental portion of Experiment 2. The RNN, excelled in the control conditions, but fell short in the experimental conditions which requires generalization. These findings shed light on the capabilities of different classes of DSMs, and in particular, which factors are useful for compositional generalization.

[1]The corpus does not contain any first-order cue (i.e. dependency between two words) useful for generalization in Experiment 2. Successful generalization requires breaking the symmetry between the two candidate instruments (e.g. *vinegar* and *dehydrator*), but neither the verb or the theme alone is sufficient to do so. For instance, the type-3 verb *preserve* is equally compatible with both instruments, and the experimental theme *pepper* is never observed with either.

Interestingly, while we observed strong performance of the Transformer during tuning (near 100% accuracy in all conditions), we observed a considerable drop in performance in the generalization condition of Experiment 2 after re-training on 10 novel random seeds. This suggests that more sophisticated hyper-tuning strategies are needed to make the Transformer more robust against such performance discrepancy. Note that the CTN, in contrast, requires no tuning.

Table 3: Accuracy of inferring the structurally-licensed rank-ordering of instruments in the control (ctl) and experimental (exp) condition. Accuracies are averages across 10 seeds.

	Experiment 1		Experiment 2	
	Ctl	Exp	Ctl	Exp
CTN	**1.00**	**1.00**	**1.00**	**1.00**
Transformer	1.00	1.00	1.00	0.43
RNN	1.00	0.40	1.00	0.06
LON	0.50	0.51	0.50	0.00

Limited Generalization in Neural Language Models

Limited compositional generalization has long been a major weak point for neural networks (Symons & Calvo, 2014; Lake & Baroni, 2018), and this work proved no exception. Both the RNN and Transformer learned the selectional preferences governing observed VP-instrument pairs, but failed to generalize to novel pairs. To succeed in the generalization task in Experiment 2 (i.e. where the instrument *vinegar* should be ranked higher than *dehydrator* and *fertilizer*), a model must utilize the distributional similarity between *cucumber* and *pepper* to transfer selectional preferences of *preserve cucumber* to *preserve pepper*. However, because the RNN and Transformer need not encode the similarity between *cucumber* and *pepper* to succeed in the language modeling task used during training, generalization based on the similarity between same-category themes is not guaranteed after training.

Interestingly, the RNN performed much worse than the Transformer in the generalization portion of both experiments. Preliminary follow-up analyses of internal states showed that, whereas the Transformer was able to consistently group experimental themes with control themes of the same category, the RNN was not able to do so. We think this difference is due to the forward-looking bias of the RNN[2] which limits the information the network can use to discover semantic similarity to words that occur after — but not before — a target word. Importantly, in our corpora, the information needed to infer that same-category themes are semantically related is marked only by the set of words that occur *before* each theme in the order in which words are presented to the model. Because the Transformer uses self-attention — the ability to consider word pairs without interference due to intervening items — instead of recurrence, it can leverage markers of semantic similarity that occur both after *and* before a target word.

In contrast, the CTN is guaranteed access to the similarity between *pepper* and *cucumber* via the indirect paths that link the two nodes via VEGETABLE-specific verbs. Consequently, the CTN is able to exploit the relatedness between the parts (*pepper* ↔ *cucumber*) to infer the relatedness between the wholes (*'preserve pepper'* ↔ *'preserve cucumber'*).

Discussion

This work examined the ability of distributional semantic models (DSMs) to perform compositional generalization, the transfer of knowledge about semantic properties of simple expressions to situations not encountered in previous experience with language. We observed that two popular connectionist models achieve lower performance relative to the Constituent Tree Network (CTN), a novel graph-based DSM, and conclude that compositional generalization does not readily emerge in these connectionist models.

There are many potential reasons why the connectionist models did not succeed in the most difficult generalization portion of our task. First, more sophisticated architectures may be needed to more strongly promote the emergence of similarity relations between themes (e.g., see Russin et al., 2020; Gordon, Lopez-Paz, Baroni, & Bouchacourt, 2020). Second, it is possible that training on more naturalistic data would better promote generalization compared to our carefully balanced artificial dataset. Third, a more exhaustive search through hyper-parameter space and random seeds may further improve performance. On the other hand, these concerns do not apply to the CTN, which does not require any hyper-parameter tuning. Compositional generalization is built into the architecture of the CTN even prior to the start of training. The CTN succeeded in the experimental condition due to its representational substrate: its edges represent constituency relations, which explicitly constrain the spreading of activation in accordance with constituent structure. More generally, the CTN cleanly separates structure and function. The formation of the network structure — joining parse-trees — is completely independent of the spreading-activation algorithm used to compute relatedness. This sharp distinction between training and inference is absent in many contemporary neural networks, where the task used during training (e.g. next-word prediction) constrains the kinds of tasks that can be used during inference.

To strengthen our conclusions, future work is needed that (i) includes a larger range of DSMs, (ii) provides a more rigorous examination of the spreading-activation algorithm under more diverse conditions, and (iii) compares the CTN to recent proposals that promise compositional generalization in vector-based and connectionist models.

[2]This follows from the fact that next-word prediction is inherently forward-looking: Next-word predictions forms similar representations of words that predict similar upcoming words. For example, when trained on our corpus, the RNN learns to cluster control themes because they share outcomes (i.e. predict the same set of instruments), but not experimental themes. In contrast, words that share predictors (i.e. all themes occur after the same set of verbs) do not form similar representations. The forward-looking bias first noted by Cleeremans and McClelland (1991) and briefly discussed by Davis and Altmann (2021).

References

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of memory and language*, *62*(1), 67–82.

Carnap, R. (1947). *Meaning and necessity: A study in semantics and modal logic*. University of Chicago Press.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*(3), 235.

Davis, F., & Altmann, G. T. (2021). Finding event structure in time: What recurrent neural networks can tell us about event structure in mind. *Cognition*, *213*, 104651.

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

Elman, J. L. (2014). Systematicity in the lexicon: On having your cake and eating it too. *The architecture of cognition: Rethinking Fodor and Pylyshyn's systematicity challenge*, 115–146.

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

Gordon, J., Lopez-Paz, D., Baroni, M., & Bouchacourt, D. (2020). Permutation equivariant models for compositional generalization in language. In *International conference on learning representations.*

Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. *language*, *39*(2), 170–210.

Kim, N., & Linzen, T. (2020, November). COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 9087–9105). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.731

Lake, B., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of seq2seq recurrent networks. In *Int. conf. on mach. learn.* (pp. 2873–2882).

Mao, S., & Willits, J. (2020). *Graphical vs. spatial models of distributional semantics.* PsyArXiv.

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.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *1st international conference on learning representations.*

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.

Russin, J. L., Jo, J., O'Reilly, R. C., & Bengio, Y. (2020). Systematicity in a recurrent neural network by factorizing syntax and semantics. In *Cogsci.*

Symons, J., & Calvo, P. (2014). Systematicity. In *The architecture of cognition* (p. 3–30). The MIT Press.

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).