COMPOSITIONAL LINGUISTIC GENERALIZATION IN ARTIFICIAL NEURAL NETWORKS

by
Najoung Kim

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
August, 2021

Abstract

Compositionality—the principle that the meaning of a complex expression is built from the meanings of its parts—is considered a central property of human language. This dissertation focuses on *compositional generalization*, a key benefit of compositionality that enables the production and comprehension of novel expressions. Specifically, this dissertation develops a test for compositional generalization for sequence-to-sequence artificial neural networks (ANNs). Before doing so, I start by developing a test for grammatical category abstraction: an important precondition to compositional generalization, because category membership determines the applicability of compositional rules. Then, I construct a test for compositional generalization based on human generalization patterns discussed in existing linguistic and developmental studies. The test takes the form of semantic parsing (translation from natural language expressions to semantic representations) where the training and generalization sets have systematic gaps that can be filled by composing known parts. The generalization cases fall into two broad categories: lexical and structural, depending on whether generalization to novel combinations of known lexical items and known structures is required, or generalization to novel structures is required. The ANNs evaluated on this test exhibit limited degrees of compositional generalization, implying that the inductive biases of the ANNs and human learners differ substantially. An error analysis reveals that all ANNs tested frequently make generalizations that violate faithfulness constraints (e.g., *Emma saw Lina* $\rightsquigarrow$ see′(Emma′, Audrey′) instead of see′(Emma′, Lina′)). Adding a glossing task (word-by-word translation)—a task that requires maximally faithful

input-output mappings—as an auxiliary objective to the Transformer model (Vaswani et al. 2017) greatly improves generalization, demonstrating that a faithfulness bias can be injected through the auxiliary training approach. However, the improvement is limited to lexical generalization; all models struggle with assigning appropriate semantic representations to novel structures regardless of auxiliary training. This difficulty of structural generalization leaves open questions for both ANN and human learners. I discuss promising directions for improving structural generalization in ANNs, and furthermore propose an artificial language learning study for human subjects analogous to the tests posed to ANNs, which will lead to more detailed characterization of the patterns of structural generalization in human learners.

Dissertation Committee

Paul Smolensky, Johns Hopkins University (Primary advisor)
Kyle Rawlins, Johns Hopkins University (Secondary advisor)
Benjamin Van Durme, Johns Hopkins University (Committee chair)
Tal Linzen, New York University
Robert Frank, Yale University

To Cookie, hoping that one day he will understand language.

Acknowledgements

I would like to first thank my dissertation committee: Paul Smolensky, Kyle Rawlins, Ben Van Durme, Tal Linzen, Bob Frank, and alternates Géraldine Legendre and Philipp Koehn for their time, and for the many challenging questions that I will continue to think about after submitting this dissertation. I am extremely grateful to have Paul Smolensky and Kyle Rawlins as advisors—they have guided me through every step of my PhD and I hope one day I can be like them. I also thank Ben Van Durme, Tal Linzen, and Ellie Pavlick whose mentorship has greatly influenced my early career path. I did not know much about anything when I started, and still can't say that I do, but the small number of things that I do know now are mostly due to the people thanked in this paragraph.

I am grateful for the friends and colleagues I met during my PhD. I cannot name everyone that I've crossed paths with, but special thanks goes to Aaron, Ayushi, Celia, Donald, Emalie, Emory, Eric, Giulia, Grusha, Jane, Karen, Karl, Laura, Little Paul, Natalia, Rashi, Sadhwi, Suhas, and Tom, and their fuzzy children I was lucky enough to spend time with: Harry, Zuri, Linus, Layla, Mr. Darcy, Theo, Bud, McRib, and Pepper (one day I hope to meet Pookie, too). I am happy that a lot of you are still going to be near me after leaving Hopkins, at least for the next couple of years. Friends back home, I will not name you all here but one day we will meet again.

I thank Sarah, Sue, and Peggy whose help was indispensable in submitting my Dissertation Improvement Grant. Some parts of the work presented in this dissertation were funded by this grant from the NSF (BCS-2041221). I would also like to acknowl-

edge the generous help and advice from Jane Lutken, Marty van Schijndel, Rachel Rudinger, Eleanor Chodroff, Alexis Ross, Allyson Ettinger, and Adina Williams during the time I was on the job market. Géraldine Legendre, the master negotiator, must be singled out and thanked in a separate sentence—I will try hard not to forget how to make miracles. Another special thanks goes to Alex Warstadt, Sebastian Schuster, Song Feng, Ian Tenney, and Nathan Schneider, who either helped or agreed to help me when I reached out for Bureaucracy Favors. I also thank past and present collaborators that I was lucky to have worked with during my PhD: Deepak Ramachandran, Burcu Karagol Ayan, Colin Wilson, Song Feng, Chulaka Gunasekara, Luis Lastras, and the 2018 JSALT team—I've learned so much from you all.

I just remembered the time when N read my message on the radio that I sent on a whim a few days before I was about to fly out to the US. I also remembered the time that I told T that I'll make something like AlphaGo and name it after them. These times were only 5 years ago and it's absurd how many things have gone out of relevance since then. I am sure I'm going to feel the same way in the few times in my life that I read this again.

Finally, I thank my family: mom, dad, and Cookie, for their love and support (and not maliciously intended anti-support, mostly from one family member). Keeping my fingers crossed that a full family reunion is possible in the near future, and that a long, healthy, happy, extravagant, and prosperous life awaits Cookie!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Compositionality—the principle that the meaning of a complex expression is built from the meanings of its constituent parts—is considered to be a central property of human language and is at the heart of many formal theories of meaning. The key benefit of compositionality is *compositional generalization*, which enables the production and comprehension of a potentially infinite number of novel expressions. Closely related to this capacity is the intrinsic relation between linguistic expressions that share the same combinatorial mechanism in constructing their meanings from their atomic building blocks (*primitives*). For instance, a speaker that knows the meaning of *John loves Mary* is necessarily able to understand *Mary loves John*, even if they have not heard or uttered this sentence before (Fodor and Pylyshyn 1988). This is possible because the expressions *John loves Mary* and *Mary loves John* have shared primitives and shared combinatorial rules that construct their meanings—these shared compositional machinery enables compositional generalization. Note that the notion of grammatical categories and the ability to infer them from context are key prerequisites for compositional generalization, because grammatical category membership plays a crucial role in determining the applicability of linguistic rules involved in meaning composition.

How (or whether) compositional generalization, a process that can be intuitively described in terms of algebraic symbolic manipulation, could emerge in a system that

operates on the basis of parallel, distributed computation is an open question. The *systematicity debate* (Fodor and Pylyshyn 1988; Fodor and McLaughlin 1990; Chalmers 1993; Niklasson and Van Gelder 1994; Aizawa 1997; Smolensky 1988, 1991, *i.a.*) is centered around a relevant issue—the connection of this debate to compositional generalization is elaborated in Chapter 2. In this dissertation, I investigate compositional generalization in artificial neural networks (ANNs) (or more recently, 'deep learning models'), which has led to unforeseen breakthroughs in the field of Artificial Intelligence in the past decade. Contemporary ANNs typically consist of multiple layers of neurons (hence 'deep') that take on real-numbered activation values, with weighted connections between them. The learning process involves adjusting the weights via gradient descent (or other optimization algorithms). One of the several possible ways to guide the learning process is to get signals about how to adjust the weights from the error measured by comparing the desired output value to the value predicted by the network. Recently, these networks have paralleled human performance on many benchmark language tasks such as open-domain question answering and textual inference (Storks et al. 2019), and have also been shown to capture human behavioral and neural data to promising degrees (e.g., recent progress on the acceptability dataset of Warstadt et al. 2019; Toneva and Wehbe 2019). Since these networks generally do not use explicit symbolic representations during computation (e.g., lexical items are represented as vectors; no categorical labels such as 'noun' or 'verb' explicitly represented), how or whether they can represent and manipulate meaningful linguistic structures important for compositional generalization is unclear.

The primary contribution of this dissertation is the development of a test based on the poverty of the stimulus method (Wilson 2006) that can be used to investigate the compositional generalization patterns of ANNs and their inductive biases that give rise to those patterns. The target generalizations in the test are (idealized) generalization patterns of human learners, constructed based on discussions in the developmental and

theoretical literature. Why would we want to investigate compositional generalization in ANNs? There are two possible outcomes to the investigation: (1) human and ANN generalization patterns align, or (2) they do not align. In the case of (1), we gain access to a model that replicates certain human generalization patterns—first, this is useful as a demonstration of how a human-like compositional capacity can be implemented in a distributed system. Second, such a model is an intelligent system displaying a human-like behavior that we can study with tools from cognitive science, with higher degrees of freedom in what experimental manipulations can be applied compared to human subjects. Finally, a model that replicates human compositional generalization patterns has applied uses for Artificial Intelligence. In the case of (2), it must indicate that the inductive bias of human and ANNs tested differ. Then, a new question emerges: what factors contribute to changing the inductive bias of the tested ANNs, to more closely match the inductive bias of human learners (i.e., better match the generalization patterns)? If we can identify such factors, they can help us better characterize the inductive biases that give rise to compositional generalization in an intelligent system.

Either case, new hypotheses for human subject studies may emerge through the practice of making the target generalizations explicit, and analyzing the generalization patterns of ANN subjects. I am not making the claim here that—even if some ANNs achieve (1)— the ways in which they achieve compositional generalization would be analogous to how it is implemented in the human brain. The view of ANNs I take is close to the 'animal models' analogy from McCloskey (1991). What we can gain in the modeling process are novel insights and ideas for future experiments involving human subjects. Independently from the connection to studies of human cognition, this dissertation can be viewed as a study of *machine cognition*, which I believe is a meaningful research enterprise of its own, especially in light of recent developments in machine learning and the surging interest in interpretation work.

1.1 Structure of the Dissertation

This dissertation consists of three main parts, corresponding to Chapters 3 to 5. Before the main parts, important terminology, related works in the literature, and the motivation behind using ANN models to study compositional generalization are discussed in Chapter 2. The content of the main chapters are summarized as follows:

Part 1 (Chapter 3) proposes a test for grammatical category abstraction and detection, a critical precondition to compositional generalization. The test is applied to a Transformer model (Vaswani et al. 2017)—a broadly adopted model in Natural Language Processing—that has been trained on large amounts of language data. More specifically, this model's ability to detect linguistic categories (parts-of-speech) is investigated. The model displayed a nontrivial degree of abstraction (although with several limitations), showing a promise that the precondition for compositional generalization could be met.

Part 2 (Chapter 4) describes the development of a test that evaluates the compositional generalization capacity of sequence-to-sequence ANN models. The test cases are based on generalization patterns discussed in linguistic and developmental studies in the literature. The test is in the format of semantic parsing—translation of natural language expressions into semantic representation language. We adopt a lambda calculus-based logical language as our semantic representation language. The models are trained on a dataset of ⟨sentence, semantic representation⟩ pairs that differ systematically from the generalization set. For example, one generalization tested in the dataset is, knowing how to translate *The cat saw the hedgehog* and *The cat saw the rat* into the semantic representation language, can one also translate *The hedgehog saw the rat* even if *hedgehog* has never been observed in the subject noun phrase?. The ANNs tested (Long Short-Term Memory (Hochreiter and Schmidhuber 1997) and Transformer models) only achieved partial success in compositional generalization,

finding generalizations that require translations of structurally novel sentences especially challenging (i.e., *structural generalization*; for instance, translating a modifier that modifies the subject noun phrase, when the modifier has only been observed modifying the object noun phrase. *The girl saw a cat on the mat* → *The girl on the mat saw a cat*).

Part 3 (Chapter 5) investigates promising factors for changing the inductive bias of the models towards preferring compositional generalization, based on the experimental results of Part 2 that demonstrate the limited generalization capacity of ANN models. I show that adding an auxiliary training objective of glossing (word-by-word translation) to Transformer models significantly improves generalization, although structural generalization remains challenging. In contrast, I report mixed results regarding the benefit of structurally informative category prediction (CCG supertagging: Bangalore and Joshi 1999) and word prediction in context (language modeling). In light of the success of the glossing objective, I discuss its connection (in terms of description, not implementation) to faithfulness constraints in Optimality Theory (Prince and Smolensky 2002)/Harmonic Grammar (Legendre et al. 1990), and how the auxiliary training promotes more faithful translations.

Finally, I end with a proposal (and a pilot experiment) for a follow-up human subject study. Parts 2–3 show that ANNs struggle the most with generalizations that require translating novel structures. As well as highlighting a major weakness in the ANN models tested, the structural generalization tests also highlight an important gap in the human experimental literature. That is, these particular structural generalizations are yet to be tested experimentally in human learners, although the theoretical prediction is that human learners are able to generalize to novel structures tested. As a follow-up study, I propose an artificial language learning experiment in which human subjects are tested in an analogous setup to the task given to ANNs in Part 2. I expect this effort to lead to a more detailed characterization of structural generalization patterns

in human learners, and a potential refinement of the ANN evaluation developed in Part 2.

Chapter 2

Background

In this chapter, I first define the terminology used in this dissertation, since terms like *compositionality*, *systematicity*, *productivity*, and *generalization*, and the relation between the concepts described by these terms are often used in relevant literature, but frequently in different ways across different works. After defining the terminology, I review relevant works in the literature that describe the linguistic generalization patterns of human and artificial neural network (ANN) learners to situate our work in this discourse. The literature suggests that machine learning models, including contemporary ANNs that have achieved great degrees of success in solving various natural language tasks, has yet to achieve strong compositional generalization capacity akin to human learners. Considering this failure, I review several promising ideas that could be applied to modify the learning biases of ANN models to make their generalization patterns match those of humans more closely.[1] Finally, I discuss what can be gained by the effort to model the compositional generalization patterns of human learners using ANN models.

[1]It is not always desirable for machine learning models to mimic the behaviors or beliefs of human learners—we would not want models in production to make generalizations based on pervasive social biases, even if many of us as humans do make such generalizations.

2.1 Terminology

2.1.1 Linguistic Compositionality

Many formal theories of meaning critically depend on the principle of *compositionality*—informally put, the idea that the meaning of a complex linguistic expression is a function of the meanings of its constituent parts. Compositionality is often offered as an explanation to how productivity in language is achieved (but see Section 2.1.3 for caveats). Here is how the argument generally proceeds: it is clear that those who have a command of language are able to produce complex expressions that they have not encountered before, and one way this can be achieved is by putting together a limited number of primitive linguistic units that they already know. While this is not the only way an unencountered expression can be produced (e.g., one can create random novel words), in order for others who have a command of the same language (who also have not encountered the produced expression) to be able to understand the expression, the meanings of the primitive linguistic units and how they are put together to construct the meaning of the complex expression must be a part of the common linguistic knowledge the interlocutors share.

However, the exact definition of compositionality has been a subject of a lengthy discussion in the literature (Pelletier 1994; Zadrozny 1994; Westerståhl 1998; Pagin and Westerståhl 2010; Pagin 2012, *i.a.*). Perhaps a good starting point is a definition from Partee (1984, p. 281) (1):

(1) The meaning of an expression is a function of the meanings of its parts and of the way they are syntactically combined. ('broad construal')

I will refer to (1) as the 'broad construal' of the compositionality principle (Partee 1984 refers to this formulation as "the most general form" of the compositionality principle). This construal is clearly insufficient, especially as a principle to abide to,

because this definition places no real constraints on what the syntax, semantics, and the relation between them could be. Stronger versions of the broad construal principle have also been proposed in the literature, one influential idea from Montague (1970) being that there exists a homomorphism between syntactic and semantic algebra (although Van Benthem 1986 and Zadrozny 1994 have argued that this formulation also does not constrain semantics in a meaningful way).

The definition (1) is taken from Partee (1984, p. 281), but immediately after this definition is given in the paper, Partee moves on to discuss what I call the 'narrow construal' (2) of the compositionality principle:

> But the principle can be made precise only in conjunction with an explicit theory of meaning and of syntax, together with a fuller specification of what is required by the relation "is a function of".

A similar idea is also expressed in Partee (1995, p. 313):

> There are several key words in the Principle of Compositionality which on closer examination can be seen to stand for theory-dependent concepts. Sharpening the Principle of Compositionality requires a theory of syntax, to specify the nature of the relevant part-whole structure, and a theory of what meanings are and by what kinds of functions they are combined.

In practice, when definitions like (1) are invoked outside of the debate about the definition of compositionality itself, it is almost always the case that additional theory-driven constraints (that are much stronger than the Montagovian homomorphism assumption) are presupposed. That is, terms like "parts", "function", and "syntactically combined" have meanings reflecting specific theories of syntax, semantics and the syntax-semantics mapping (and these specificities prevent the narrow contrual

compositionality from being vacuous[2]). This theory-dependence of compositionality is taken to be the standard view in formal semantics (Groenendijk and Stokhof 2004, p. 84: "[. . .] an assessment of the status of the compositionality principle is intimately tied to whatever view we have of both structure and meaning"). This can be viewed as a 'narrow construal' of the broad definition:

(2) The meaning of a whole is a function of the meanings of the parts and of the way they are syntactically combined according to a theory of language that defines the syntactic and semantic machinery, the set of primitives, and the syntax-semantics mapping. ('narrow construal')

The natural consequence of the standard view is that the constraints concerning the *what* and the *how* of composition must be (at least at the beginning of theorizing) motivated independently from the broad construal principle of compositionality itself. For example, a system that adopts Frege's Conjecture and restricts the combinatorial operator for semantic composition to function application clearly adheres to the broad construal of compositionality. However, the broad construal compositionality does not speak to why the permitted operation should be restricted to function application (i.e., this restriction is not motivated by the broad version of compositionality). Dowty (2007, p. 24) also shares several examples of semantic rules that are implausible due to reasons external to compositionality:

> "if the maximum depth of embedding in the sentence is less than seven, interpret the whole sentence as negative; if it is seven or more, interpret it as affirmative". [...] "If the number of words in the sentence is odd, interpret the scope of quantificational NPs from left to right; if it is even, interpret

[2]For example, see the discussion of Groenendijk and Stokhof (1991) which illustrates how positing a compositional truth-conditional semantics leads to the rejection of the theory that meaning *only* consists of truth conditions. Then, compositionality according to truth-conditional semantics clearly is falsifiable and therefore not vacuous.

> scope from right to left". Such rules as these are ways of "determining (a part of) the meaning of a sentence from its words and how they are combined syntactically", but no linguist would entertain rules like them for a moment. The UNCONSTRAINED COMPOSITIONALITY that a broadly stated "Frege's Principle" encompasses is most likely not what linguists really have in mind when they question whether language is or is not 'compositional'.

Thus, compositionality is mostly taken to be a methodological principle in building theories (Partee 1995; Janssen 1997)—if some phenomenon X cannot be derived compositionally under a specific theory (that defines the *what* and the *how* of composition), the theory must be revised so we can have a compositional account of X.[3]

The view adopted in this dissertation is in line with the standard view—I assume that there are broad (1) and narrow (2) construals of compositionality, where narrower construals are specific theory-dependent instantiations of the broad compositionality principle. While the broad construal can be subject to the vacuity criticism, the narrower ones need not be, depending on the nature of the qualification. However, it is true that all narrower instantiations adhere to the compositionality principle broadly construed.

2.1.2 Systematicity

The notion of systematicity is usually attributed to Fodor and Pylyshyn (1988, p. 37) (F&P),[4] where they define systematicity (or what it means for a linguistic capacity to

[3] In this regard, the algebraic findings that any semantics can be encoded as a compositional one (which are typically used to argue for the vacuity of compositionality) is in fact a welcome result, since it means that a compositional account of X can always be found (Janssen 1997). However, it has been contested whether such formal results (e.g., Zadrozny 1994; Van Benthem 1986; Janssen 1986) have any significance to natural language semantics due to their largely ad hoc assumptions about syntax and semantics (Westerståhl 1998).

[4] Note that the goal of F&P was to claim the systematicity of thought, and the systematicity of linguistic capacities was used as an illustration for this claim.

be systematic) as follows:

> What we mean when we say that linguistic capacities are systematic is that the ability to produce/understand some sentences is *intrinsically* connected to the abiliity to produce/understand certain others. [...] systematicity is a property of the mastery of syntax of a language, not of its lexicon.

According to F&P, complex expressions in language are connected to other complex expressions based on their constituent structure—since *John loves Mary* and *Mary loves John* share constituent analyses and the primitives, it is impossible for speakers of English to accept one without accepting the other. The same cannot be said about the lexicon which consists of atomic units; understanding *rabbit* does not imply understanding other items in the lexicon such as *tree*. While the above quote only phrases systematicity in terms of "mastery of syntax", F&P's overall appeal to "understanding" implies that systematicity not only concerns well-formedness judgments but also the meanings of the expressions that are intrinsically connected. These intrinsic connections are attributed to the shared structural analyses and combinatorial mechanisms, meaning that a narrow construal of compositionality (Section 2.1.1) is presupposed in linguistic systematicity as presented in F&P.[5]

The Systematicity Debate: Systematicity was proposed as a counterargument to the claim that connectionist models can serve as adequate models of the mind. The argument is that they cannot serve as adequate models of the mind because exhibition of systematicity does not follow from connectionist models (Fodor and McLaughlin 1990; McLaughlin 1993, *i.a.*). The core argument is not about whether connectionst models can exhibit systematicity—even those who reject connectionist models as models of the mind admit that it is possible to find connectionist models that display systematicity (e.g., one can build a connectionist model that implements

[5]Note that the reverse does not hold; compositionality does not presuppose systematicity as presented in F&P.

a classical symbolic system). Rather, the argument is that, it is the guarantee of systematicity (or *nomological necessity*) that connectionist models fail to provide, even if they can somehow be successfully trained to display systematic behavior under specific conditions. This view has been criticized my many. One counterargument is that the guarantee of systematicity does not in fact differentiate symbolic and connectionist models, since symbolic models also only have the guarantee under specific assumptions about the relevant rules that derive the systematic properties in question (Hadley 1997). For instance, it is clear that the systematic relation between *John loves Mary* and *Mary loves John* is not achieved for free in a symbolic model by virtue of the model being symbolic—it is a product of additional assumptions that the two sentences receive the same structural analysis, and that meaning computation makes use of these shared constituent structures. It is difficult to claim that these additional nontrivial assumptions are necessary properties of a symbolic architecture.

Smolensky (1995) also presents a criticism, specifically that the argument laid out in Fodor and McLaughlin (1990) that properties such as systematicity is entailed by the classical symbolic account results from conflating assumptions and entailments—systematicity only follows from the symbolic account because of their own (nontrivial) assumptions about the symbolic account (see Fodor and McLaughlin 1990, pp. 202–203 and Smolensky 1995, pp. 285–286 & Footnote 36 for the exact discussion). Moving beyond this particular criticism, Smolensky (1995) takes a stance that certain kinds of connectionist models can in fact explain systematicity. In brief, the claim advanced is that the symbolic accounts are a high level description of the mental structure (*levels* in the sense of (Marr 1982), although the use of levels here follow the revision proposed in Smolensky 2006), and the specifics of the lower levels (which are subject to more constraints) have causal implications at the higher levels. Regarding systematicity in particular, it is argued that if the intermediate level—below the level of symbolic description—is a connectionist model that uses Tensor Product

Representations (Smolensky 1990), properties like systematicity at the higher level can be explained as a consequence of the principles at the lower level that realize the symbol structures.[6]

This discussion leads to two follow-up thoughts. First, if neither classical nor connectionist models meet the 'guarantee of systematicity' criterion, perhaps a more fruitful formulation of the question regarding this criterion would be, how can models be constrained such that they display a guarantee of systematicity under those constraints? As discussed before, in symbolic models, defining the syntactic and semantic machinery, a set of primitives, and a syntax-semantics mapping would suffice. In connectionist models, it is still an open question what the constraints would be, and several contemporary works have inherited this question in the context of modern deep learning models. A review of this line of work is continued in Section 2.2. Second, how does the 'levels of analysis' discussion connect to the contemporary ANN models? I revisit this second thought in Section 2.4.

Systematicity and linguistic categoricity: F&P's conception of systematicity has been criticized as imprecise (Niklasson and Van Gelder 1994), and many attempts at a more precise conceptualization have been made. One such attempt can be found in Johnson (2004), where a formulation based on *intracategorial mutual substitutability*[7] is proposed: there are sets of linguistic expressions that are characterized by full subtitutability of any pairs of expressions within that set without losing grammaticality. This formulation is not without its problem (and Johnson himself does not endorse

[6]This is not the *only* account that could derive systematicity at the symbolic level from lower-level principles. If the implementational level consists of symbolic computation instead of connectionist computation using Tensor Product Representation, the account would have explanatory power that the Fodorean view based on assumptions within the functional level would not (Smolensky, p.c.), and a sketch of the 'symbolic computation all the way down' view is laid out in Smolensky (2006). Nevertheless, this view of the implicational relation between different levels of analysis, and the discussion of a particular connectionist model that is capable of deriving desired properties at the symbolic level, counters the view that connectionist models cannot have any meaningful role in understanding the mind.

[7]This term itself was used by Pullum and Scholz (2007) in their discussion of Johnson (2004).

adopting this definition), especially in the context of natural language where identifying a set with full intracategorial substitutability itself is a challenging enterprise.[8] However, it is also true that such a set, however fine-grained, exists,[9] and the substitutability there must be accounted for (Pullum and Scholz 2007). I do not engage in the discussion of whether this is possible—rather, the point to be made is that the existence of a systematic connection between two complex expressions depends on the *category membership*[10] of the constituents under the substitutability view (a similar observation is made in Werning 2005).

One could adopt a slightly more general view of what *intrinsic connection* by systematicity means (as I have been implicitly doing so in the beginning of this section and also at the beginning of this dissertation in Chapter 1), such as connection via shared underlying structure and mapping rules. Under this view, *The hedgehog loves Giulia* and *Jane loves Giulia* would have a systematic connection, assuming syntactic parses like [[The hedgehog]$_{NP}$ [loves [Giulia]$_{NP}$]$_{VP}$]$_S$ and [[Jane]$_{NP}$ [loves [Giulia]$_{NP}$]$_{VP}$]$_S$, and semantic representations like $love'(\iota x.hedgehog'(x), Giulia')$ and $love'(Jane', Giulia')$ for the two sentences, respectively. The two sentences share a common syntactic parse (at the level I have bracketed; they do not share the exact same parse if the internal structures of the noun phrases are also considered), and a common mapping rule that the semantic representation of the subject noun phrase (NP) saturates the first argument slot of the predicate, and the semantic representation of the object NP saturates the second. This means that there are common computations involved in how the semantic representations of these two sentences are derived—these shared computations are what I view to be the systematic connection between different complex expressions. Note that these two sentences do not satisfy the definition of

[8] See Pullum and Scholz (2007) for an extensive list of examples for why this is not a trivial problem.

[9] For example, a small subset of proper names like {Sam, Shane} is probably fully substitutable without loss of grammaticality.

[10] Specifically, the membership of a set that satisfies the substitutability criterion.

systematicity based on intracategorial substitutability—for instance, *Jane* cannot be substituted by *the hedgehog* in the vocative use of *Jane* in *Look at this, Jane!*. This view of systematicity depends on a broader set of linguistic categories than categories defined on the basis of intracategorial substitutability. Still, the observation that systematicity critically relies on the category membership of the constituents still holds, since category membership plays a critical role in determining the syntactic parses, which in turn affect the application of the mapping rules. This interpretation of systematicity has a possibility to be tied to compositionality in a way that systematicity-as-substitutability cannot, *if* the theory one assumes are compositional. I adopt this rule-governed (or shared computation-based) view of systematicity, as do Lake and Baroni (2018). This is reflected in the design of the task in Chapter 4. Furthermore, I separately explore the issue of detecting linguistic category membership in Chapter 3 since it plays an indispensable role in the generalization facilitated by systematicity under this view.

2.1.3 Productivity

The argument from productivity is often presented as evidence for the compositionality of language. This argument is commonly introduced by the following quote from Frege (1963, p. 1) (the original idea usually attributed to Frege 1892):

> It is astonishing what language can do. With a few syllables it can express an incalculable number of thoughts, so that even a thought grasped by a terrestrial being for the very first time can be put into a form of words which will be understood by somebody to whom the thought is entirely new. This would be impossible, were we not able to distinguish parts in the thought corresponding to the parts of a sentence, so that the structure of the sentence serves as an image of the structure of the thought.

Note that this quote concerns not just the ability to produce novel expressions, but also the possibility for others to understand them. Indeed, productivity from a pure expressibilty perspective has been argued to be weak (Pagin and Westerståhl 2010): for a speaker to express unlimited number of sentences or propositions from a limited number of primitives, they can simply resort to concatenation without assuming any internal structure or assuming that the primitives are meaningful.[11]

Thus, the argument from productivity as it is commonly understood must include a "convergence of interpretation" (Pagin and Westerståhl 2010) component that those who share a common language must be able to arrive at a similar interpretation, given an expression that they have not encountered before. Still, productivity including the convergence of interpretation aspect does not require compositionality as a precondition (it only requires semantics to be computable, not compositional: Pagin 2020), although compositionality could be an inference to the best explanation.

However, there exist some narrow construals of compositionality (2) from which productivity is a natural byproduct. For example, if we assume a context-free grammar that contains at least one recursive rule and a homomorphism between syntax and semantics, productivity follows from this narrow construal of compositionality.

2.1.4 Compositional Linguistic Generalization

To summarize the discussions in the previous sections, I have adopted a theory-centric view of compositionality, and have argued that properties like systematicity and productivity can follow from certain narrow construals of compositionality. But they are by no means the only possible construals of compositionality; hence it is incorrect to claim that properties such as systematicity and productivity are characteristic of compositional systems in general. In this dissertation, I adopt the view of compositionality as a reasonable working hypothesis for how production and

[11] Also see Werning (2005) for an example analysis of quotations that is productive but not compositional.

interpretation of novel complex expressions are possible. And I refer to the production and interpretation of novel complex expressions as predicted by an underlying set of linguistic rules *compositional (linguistic) generalization*.

2.2 Compositional Generalization in Human and Machine Learners

The most relevant empirical studies of compositional linguistic generalization in human learners come from the developmental literature studying the formation of abstract linguistic rules. The ability to form and infer linguistic categories[12] is also related, since category formation is a critical precondition to compositional generalization as discussed in previous sections—category membership determines the applicability of linguistic rules. In this regard,[13] generalizations conforming to the posited grammatical categories and linguistic rules (*abstractions*) are presented as evidence for their existence (e.g., Tomasello and Olguin 1993, Gertner et al. 2006). Moreover, the lack of generalization is presented as evidence for the lack of certain abstractions (Olguin and Tomasello 1993), and overgeneralization as evidence for non-adult-like abstractions (Bowerman 1982; Brooks et al. 1999). These developmental studies suggest that human learners exhibit compositional generalization from a very young age: children can generalize nouns in different semantic roles[14] than they were observed in (Tomasello and Olguin 1993), generalize to passive verbs only observed in active constructions (production: Brooks and Tomasello 1999, comprehension: Messenger and Fisher 2018), generalize to different transitivity of verbs (production: Kline and Demuth 2014, comprehension: Scott and Fisher 2009), and generalize verbs

[12] I use the word *categories* in a broad sense that encompasses both the general lexical categories (e.g., noun, verb) and specific subcategorization frames (e.g., Levin classes; Levin 1993) here.

[13] A lot of the discussions in this line of work center around nativist versus empiricist views about the formation of linguistic abstractions, but neither view denies that abstractions going beyond individual lexical items are an important part of adult linguistic knowledge.

[14] The generalized uses differed in syntactic configurations as well as semantic roles.

only presented in one dative construction to the alternate construction (production: Conwell and Demuth 2007, comprehension: Rowland and Noble 2010).

It is usually accepted without extensive experimental investigation that adults also are capable of similar generalizations, and also that in general they can produce and interpret novel well-formed expressions in their native language. One relevant study is Lake et al. (2019), which tested compositional generalization more explicitly through artificial grammar learning. In their experiment, adult English speakers were taught instructions in an artificial language, where some lexical items denoted arguments and some denoted functions that take arguments (e.g., repeating the argument three times). The majority of the participants adopted a compositional strategy in interpreting novel instructions, showing generalization behavior that was in accordance with the interpretation in which the learned functions were composed in a novel way.

Compositionality has also been a continuous topic of interest in the context of Artificial Intelligence. In particular, much attention is being drawn towards compositionality in the context of the recent advances based on deep learning. Deep learning models generalize extremely well to test examples drawn from the same distribution as their training, demonstrated by impressive performance on many benchmark tasks that primarily adopt this kind of *in-distribution* generalization-based metrics (e.g., Wang et al. 2019b; He et al. 2015). However, their generalization behaviors on examples outside of the training distribution often do not align with the "desirable direction" (Geirhos et al. 2020); for example, models of vision are often not robust to distribution shifts like rotation or color shift that humans are robust to, and models of language are not robust to the meaning shifts like those triggered by negation (Ettinger 2020; Kim et al. 2019b). There seems to be a shared belief in the field that compositionality is how models may achieve generalization (of the desirable kind; for instance, for language, generalizations consistent with the rules of the language) beyond their training, and there are active efforts to endow deep learning models with a compositional capacity

(Schlag et al. 2019; Herzig et al. 2021; Lake 2019; Chen et al. 2020b, *i.a.*).[15]

Some part of this effort has been spent on assessing the ways in which deep learning models generalize, and comparing them to the generalization capacity of humans (of course, in tasks where human generalization patterns are deemed to be "desirable"). This is by no means a new question that only emerged in the era of deep learning; it has a long history in the context of the systematicity debate (Section 2.1.2). Early works in this area have attempted to propose a finer-grained categorization of different generalization behaviors in terms of the degree of systematicity they display; for instance, Hadley (1994) proposed weak, quasi- and strong systematicity based on whether models can process/interpret expressions with a lexical item in unseen complex contexts when the model has been exposed to that lexical item in all possible syntactic configurations, unseen complex contexts that may vary in their syntactic configurations but not in terms of the grammatical roles they occupy, and unseen complex contexts with the lexical item occurring in a held-out syntactic configuration (cf. Christiansen and Chater 1994 for a further refinement of the terminology). Hadley (1994) also offers a review of the modeling efforts, concluding that no model that has been proposed in the literature at the point of publication displayed strong systematicity (which is the degree of systematicity that we expect from human learners). While Hadley and Cardei (1999) claim to have developed a model that achieves strong systematicity in a network that performs the task of semantic parsing, this has been later argued to be a classical system rather than connectionsit, because of the explicit structured representations the models were provided with (Frank et al. 2009). Later works that involve various degrees of success in achieving systematic behavior include Hadley

[15]The issue of broad versus narrow construal of compositionality (Section 2.1.1) seems relevant here, although this is not often explicitly discussed. In virtue of how deep learning models are designed, the broad construal of compositionality is trivially satisfied—the output of a deep learning model is a function of its inputs. It is unclear whether there is one 'compositionality' that is relevant to the various generalization problems for which compositionality is considered to be the right solution. A more precise characterization of the narrow construal of compositionality may help understand the space of the problems better.

et al. (2001); Bodén (2004); Frank (2006); Frank et al. (2009).

Turning to the more recent literature, Lake and Baroni (2018) have revisited the issue of systematicity in neural network models and proposed a simple compositional generalization task using navigation commands called SCAN, and demonstrated that Simple Recurrent Neural Networks (Elman 1990), Long Short-Term Memory (Hochreiter and Schmidhuber 1997) and Gated Recurrent Units (Cho et al. 2014) did not generalize systematically. This work has influenced the creation of numerous other benchmarks that test for compositional generalization, including an earlier version of the work presented in Chapter 4 of this dissertation (Kim and Linzen 2020). As direct follow-ups of Lake and Baroni (2018), Loula et al. (2018) have proposed additional training/generalization splits of SCAN, and Ruis et al. (2020) have proposed gSCAN which is SCAN grounded in a grid world—an agent has to process its surrounding environments in following the navigation commands. Outside of the SCAN task, Keysers et al. (2020) have created the Compositional Freebase Questions (CFQ) dataset, where the training and generalization sets contain SQL queries with similar primitive distribution but different distribution of the composed forms. Hupkes et al. (2020) have also proposed novel synthetic datasets for testing various aspects of compositionality, such as productivity, overgeneralization and substitutability. Overall, the evaluation results on these tests suggest that contemporary neural architectures show only a limited degree of compositional generalization. They are easily able to generalize to novel examples that require composing parts of known examples, if the training and evaluation sets are not distributionally distinct. But if there are distributional differences between training and evaluation sets (e.g., the target output is longer than any of the examples the model has encountered during training), compositional generalization to novel examples poses significant challenges. I refer the readers to Baroni (2020) for a more detailed overview of compositionality in the context of contemporary modeling developments.

2.3 Learning Bias for Compositional Linguistic Generalization

This dissertation focuses on particular generalization patterns exhibited by human learners that can be credited to linguistic compositionality (Section 2.2). We take these patterns (more precisely, an idealized version of these patterns) as the "desirable" generalization patterns. Without defining such a reference point, it is difficult to assess different generalization behaviors, because there are always multiple plausible generalization patterns that are consistent with observed data. Let's look at an example, assuming that a learner must make generalizations based on past observations only. The learner is shown input-output pairs (1, 1), (2, 2), (3, 3). Based on these observations, what should the learner output, given the input 4? The answer would be 4 if the generalization hypothesis is that the input and output values are equal. But the answer 2 is equally consistent with the observations (i.e., there are no contradictory observations), if the hypothesis is that an odd number maps onto itself and even numbers map onto 2. The answer 3 is equally consistent if the hypothesis is that numbers 1 and 2 map onto themselves and everything else maps onto 3. If the only criterion to assess these generalization hypotheses is consistency with past observations, we can create infinitely many hypotheses of arbitrary complexity that are all equally plausible. If a learner ends up choosing one hypothesis (or probabilistically picking a hypothesis among a set of hypotheses) over all others, that must mean that the learner has additional criteria that makes that hypothesis more plausible to the learner than others. These additional criteria comprise the learner's *inductive bias* (or *learning bias*), which determines how the learner generalizes to novel inputs (Mitchell 1980).

Thus, compositional generalization, which requires generalization to novel linguistic expressions, is also an issue concerning the inductive bias of learners. Now we can unpack the underlying assumptions behind the phrase *compositional linguis-*

tic generalization a little more: first, there must be some instantiation of linguistic compositionality (2) that distinguishes compositional and noncompositional ways of constructing a meaning representation given a complex expression. Given a novel complex expression, a competent speaker can make a generalization about the expression via the compositional way of constructing its meaning. Here is a toy example, under an instantiation of compositionality under theory T that assumes a simple phrase-structure grammar that parses transitive sentences as [NP [V_{tr} NP]$_{VP}$]$_S$ and a mapping to a semantic representation that treats transitive verbs as two-place predicates saturated by the semantic representation of the first NP and the semantic representation of the second NP (i.e., T: $NP_1\ V_{tr}\ NP_2 \rightsquigarrow V'_{tr}(rep(NP_1), rep(NP_2))$)[16]:

(3) Observations

Najoung loves Cookie $\rightsquigarrow$ love′(N, C)

Najoung loves Linus $\rightsquigarrow$ love′(N, L)

Cookie loves Najoung $\rightsquigarrow$ love′(C, N)

Cookie loves Linus $\rightsquigarrow$ love′(C, L)

Generalization 1 (Compositional according to T)

Linus loves Cookie $\rightsquigarrow$ love′(L, C)

Generalization 2 (Not compositional according to T)[17]

Linus loves Cookie $\rightsquigarrow$ love′(C, C)

Without T, Generalizations 1 and 2 will both be equally possible, and one could

[16] I will use a single character as the semantic representation of NPs here for simplicity. For example, *rep*(Cookie) = C, *rep*(the wall) = w

[17] While Generalization 2 seems quite unacceptable to us (the readers might not even have entertained the thought that Generalization 2 is a possibility), we will see in later parts of this dissertation that some ANN learners do in fact make similar generalizations, suggesting that they have a very distinct inductive bias from human learners. This is of course not claiming that the ANNs are generalizing according to T′; T′ is just an illustrative example.

potentially argue that both generalizations are compositional under the broad construal, because one can come up with an alternative theory T′ that allows Generalization 2 to be compositional under T′[18]. However, in the presence of T, only Generalization 1 is compositional according to the permitted mechanisms—there is no way of constructing a meaning representation like love′(C, C) from *Linus loves Cookie* without a substantial revision of T.[19]

Therefore, examining the generalization patterns of a learner given a set of observations as in (3) can give insights into what T is, or more generally, into their inductive bias (there is no guarantee that the generalizations are compositional or straightfowardly describable in terms of linguistic rules, depending on the learner). In this dissertation, I develop a test for compositional linguistic generalization by constructing target generalization patterns based on the what we know or expect

[18]For example, imagine a T′ that is equivalent to T, but only differs in that the first argument of $V'_{tr}(arg_1, arg_2)$ is saturated by C if the semantic representation of the subject is not in the set {N, T, P, C} (i.e., T′: $NP_1\ V_{tr}\ NP_2 = V'_{tr}(C, rep(NP_2))$ if $rep(NP_1) \notin$ {N, T, P, C}, else $V'_{tr}(rep(NP_1), rep(NP_2))$). T′ adheres to the broad construal of compositionality because the meaning of the subject is involved in determining the meaning of the whole expression. If we adopt T′ instead of T, the compositional generalization would be generalization 2.

[19]A more plausible example to human readers where different compositional theories predict different generalizations would be generalizations involving idiomatic expressions. Depending on whether the idiomatic mappings are part of the theory or not (T_1 vs. T_2), compositional generalization yields different predictions for idiomatic verb phrases combined with different subjects. Under T_1, it is impossible to derive the 'die' sense for an unobserved combination of a subject noun phrase and *kicked the bucket* compositionally.

(i) Observations
John kicked the man ⇝ kick′(J, m)
John kicked the wall ⇝ kick′(J, w)
Sam kicked the man ⇝ kick′(S, m)
Sam kicked the bucket ⇝ die′(S)

T_1: $NP_1\ V_{tr}\ NP_2 \rightsquigarrow V'_{tr}(rep(NP_1), rep(NP_2))$
T_2: $NP_1\ V_{tr}\ NP_2 \rightsquigarrow V'_{tr}(rep(NP_1), rep(NP_2))$,
but if V_{tr} = kicked and NP_2 = the bucket, ⇝ die′($rep(NP_1)$) is possible

Generalization 1 (Compositional according to T_1)
John kicked the bucket ⇝ kick′(J, b)

Generalization 2 (Compositional according to T_2)
John kicked the bucket ⇝ kick′(J, b)
or ⇝ die′(J)

about generalization patterns of human learners (Section 2.2). In this test, the subject is provided some observations and is then asked to make generalizations about expressions not in the set of the observations, as in (3). Using this test, I ask the following two questions for ANN learners (specifically, Transformer (Vaswani et al. 2017) and Long Short-Term Memory (Hochreiter and Schmidhuber 1997)): (1) Do their generalization patterns align with the target patterns? and (2) If not, what changes made to the learners make them prefer the target generalization over others?

2.3.1 Review of Ideas

Prior work (Section 2.2) and published versions of the chapters in this dissertation (Kim and Linzen 2020; Kim and Smolensky 2021) suggest that the answer to question (1) is negative. This must mean that the inductive biases of the models tested are different from human learners. Then, what would be the answer to question (2)? In other words, what factors contribute to forming an inductive bias in a computational model such that the models will show more similar patterns of generalization to human learners? I review some promising ideas here, although not all of the ideas are explored in this dissertation.

Linguistic theory-based priors: As discussed previously, being able to recognize and manipulate abstract linguistic structures is critical to compositional linguistic generalization. For example, in order to make a generalization like (3)—assigning a correct semantic representation to a sentence that contains a noun phrase that was only observed as a grammatical object in the subject position—the learner must be aware of what noun phrases are and what syntactic positions noun phrases can occupy. If sensitivity to linguistic structures is important, changes to the model that facilitates the assignment of linguistic structures to the input/output may facilitate compositional generalization that relies on those structures. There are multiple ways that this may be achieved:

- By using model architectures that are able to (or are encouraged to) represent tree structure, the models can be provided with information about the structure of the inputs analyzed according to certain theoretical frameworks (works that use phrase structure grammar: Wang et al. 2019c; Harer et al. 2019; Tai et al. 2015; Nguyen et al. 2020; dependency grammar: Tai et al. 2015; Strubell et al. 2018).

- Structural analyses can be encoded in the distributed representations that the models process (Zanzotto et al. 2020; Sachan et al. 2020).

- Linguistic features (e.g., part-of-speech, case) can be provided as a part of the input (Sundararaman et al. 2019).

- Structure detection can be used as an auxiliary training objective. For instance, parsing (Vilares et al. 2020), CCG Supertagging (Bangalore and Joshi 1999; Hockenmaier and Steedman 2007), and dependency distance prediction (Xu et al. 2020) have been proposed in the literature. CCG Supertagging in particular has been shown to be helpful for tasks that require structure sensitivity, such as end-of-sentence detection (Kim et al. 2019a; Pruksachatkun et al. 2020).

Domain-general regularity priors: The methods listed above all carry specific theoretical assumptions about what the right linguistic structures are. On the other hand, there are domain-general regularities that are purported to be connected to regularities that underlie linguistic structures. Here are ways in which preference for a more domain-general regularity (i.e., not specific to language) may contribute to an inductive bias for compositional generalization:

- A general perceptual mechanism of invariance detection has been hypothesized to underlie the detection of linguistic regularities (Gogate and Hollich 2010).

Translation-invariant properties of convolutional encoder-decoders (Gehring et al. 2017) could be useful in this respect.

- Geometric regularities in vector space that correspond to linguistic regularities have been shown to emerge in unsupervised neural network models trained to predict words in context (Mikolov et al. 2013). Furthermore, Kim and Linzen (2019) have found that consistent offsets between constituent parts of a compositional phrase are correlated with better generalization. Hence, a mechanism that promotes such a geometric regularity (e.g., a loss function that explicitly incentivizes consistent offsets) may facilitate the representation of structural regularities, leading to better compositional generalization.

- Tensor Product Representations (Smolensky 1990) provide a general technique for representing symbolic structures in distributed space through tensor product binding of fillers and roles. This explicit separation of fillers and roles may work as a general bias towards representing the structural regularities in the observed data—for instance, Chen et al. (2020a) have shown that Tensor Product Representations help improve performance on a math question-answering task that requires learning a natural language-to-formal language mapping.

Stochastic factors: One important consideration in analyzing the findings from modern deep learning based methods is that every aspect of the model specification potentially has an impact on generalization behavior. For instance, even within the same class of models (e.g., Long Short-Term Memory; Hochreiter and Schmidhuber 1997), the generalization pattern may be modulated by factors such as input length and number of input examples, and may also be affected by hyperparameters such as embedding sizes and dropout probability (Kharitonov and Chaabouni 2020). Thus, the questions of models' inductive bias should be discussed in the context of a holistic model configuration rather than specific to a certain class of model architecture (which

would be an overgeneralization of the findings).

2.4 Why Artificial Neural Networks?

This dissertation examines generalization patterns of contemporary varieties of ANNs that are commonly used to model language in the field of Natural Language Processing. But what is the utility of using ANNs as models of human language? Broadly, my view aligns with that of McCloskey (1991), both about the utility and the limitations of this approach. McCloskey (1991) draws an analogy to the role of animal models in cognitive science: under the assumption that there are some shared underlying properties between an animal system and a human system, a better understanding of the structure and function of an animal system may provide insights that advance the understanding of a human system. Using animal models allows for manipulations that are not necessarily possible with human subjects, such as lesion studies. ANN models have the same benefits, except that they may allow for even finer-grained levels of control over environmental manipulations and better replicability. However, as is the case with animal models, ANN models leave open a question of how exactly the insights gained from the experiments can transfer to a human system. For example, teasing apart which aspects of the model design are relevant to the broader cognitive question and which aspects are idiosyncratic to the specific model used is a challenging issue. The following quote from Frank et al. (2009, p. 373) expresses a similar sentiment regarding the arbitrariness of ANN model specifications:

> Traditionally, connectionist solutions to the problem of systematicity are sought in architectural constraints, combined with specifics of training data. Such an approach is unlikely to succeed in our opinion, because the specifics of the architectures and training procedures appear to be chosen to achieve the desired results rather than being independently motivated.

This concern is valid even outside of the systematicity debate, and applies equally to contemporary ANNs. A lot of the implementational idiosyncracies that do not have immediately obvious cognitive motivations (e.g., choice of optimizer, early stopping patience, initialization function) modulate model behavior, and we should be aware of this issue. However, my view is that independently motivated factors and controlled experiments that only manipulate those factors may still contribute something valuable, especially if the experiments show that they can significantly constrain the learning space in comparison to experiments conducted without the target manipulations. For instance, identifying the factors that enable us to train a set of models that reliably display systematic behavior could help address a softer version of the 'guarantee of systematicity' question (Section 2.1.2).[20]

The discussion above focused on how the studies presented in this dissertation can potentially be useful to studies of human cognition. However, I note that independently from this discussion, I consider this dissertation to be a study of machine cognition, taking models that display some sort of intelligent linguistic behavior (even if they do not exactly align with human behavior) as experimental subjects.[21] Also relevant here is the question I have raised in Section 2.1.2 regarding the role of contemporary ANNs in different levels of analysis to understand an intelligent system. My view is that each instantiation of an ANN model provides access to intermediate- (the specifications and components of the model) and lower-level (the actual numerical operations involved) descriptions of a system. The frequent reference to ANN models as 'black boxes'

[20]The goal of the dissertation is not to provide a counterargument to the classicist side of the systematicity debate (i.e., "do away with the symbol level of analysis": Fodor and Pylyshyn 1988). Rather, the goal is first to explore whether we can train a set of models that generalize compositionally without building in or teaching them explicit rules of symbol manipulation that we assume to underlie those generalizations, and then identify factors or model components that contribute to constraining the learning space. The outcome could plausibly be that the set of models that successfully generalizes internally implements a mechanism that is analogous to symbolic manipulation (which would then not be a counterargument to classical systematicity), and this would very much be a welcome result.

[21]The connection to human behavior from the viewpoint of machine cognition is based on the observation that compositional generalization patterns of human learners are desirable but not currently exhibited by the ANN models.

arise from the lack of higher level symbolic descriptions. To be more precise, there are definitely symbolic representations that correspond to representations that are processed by these models (e.g., in many ANN models, words are mapped to vectors at the embedding layer, and the output vector sequences can be interpreted as a sequence of words)—it is the lack of functional descriptions in terms of these symbolic representations that give rise to this impression. The recent literature on 'probing' ANN models used for Natural Language Processing can be thought of as nascent attempts at higher level symbolic descriptions, based on the literature on symbolic, functional characterizations of human language (i.e., a large portion of theoretical linguistics). The tests developed in this dissertation can be understood as a part of this effort. Furthermore, under the view of Smolensky (1995) regarding levels of analysis, properties of the lower level have causal implications at the higher level.[22] While this is not within the scope of this dissertation, this signals the possibility of identifying novel constraints for descriptions at the symbolic level that arise from the lower level properties of the models. This approach would be especially interesting if we can identify a set of models that successfully replicates complex human generalization patterns.

[22]This view differs from the classical Marr's levels that the levels are largely independent from each other.

Chapter 3

A Test for Grammatical Category Abstraction

3.1 Motivation

The notion of grammatical categories is fundamental to human language. Humans can abstract over individual lexical items to form grammatical categories, such as nouns and verbs in English, and this abstraction is important because category membership (rather than individual lexical identity) determines the applicability of linguistic rules. For instance (with lots of caveats), 'Determiners combine with nouns' rather than '*the* combines with *robot* and *cookie*'. *Category membership inference*, therefore, is a critical precondition to compositional linguistic generalization as discussed in Sections 2.1.2, 2.1.4, and 2.2. For this reason, we start with developing a test that evaluates basic category membership inference in ANNs before discussing their capacity to make compositional generalization in the subsequent chapters.

The category membership of a new word that a learner encounters can be rapidly inferred from their linguistic environment: if a speaker of English hears *I saw a blick*, it is immediately clear that *blick* is a noun. If it is known that *blick* is a noun, this knowledge about the grammatical category of the word furthermore facilitates the

*An earlier version of this chapter has been published as Kim and Smolensky (2021) in the 2021 Proceedings of the Society for Computation in Linguistics.

speakers to produce or judge as grammatical new sentences such as *We like the blick* and *The blick jumped*, even though these new sentences have no lexical overlap (other than the word *blick* itself) with the context that *blick* was first observed in. Hence, the identification of a grammatical category allows application of rules that operate over that category, allowing for generalization outside of the context that the novel word has been observed in (Gómez and Gerken 2000). The capacity to infer linguistic categories and make generalizations based on category information has been actively explored in studies of child language (Gelman and Taylor 1984; Gerken et al. 2005; Kemp et al. 2005; Meylan et al. 2017; Olguin and Tomasello 1993; Skipp et al. 2002; Tomasello and Olguin 1993, *i.a.*), and studies using artificial language learning (Mintz 2002; Reeder et al. 2013; Aslin and Newport 2014; Reeder et al. 2017, *i.a.*).

In this chapter, we propose a test that targets the question of category membership inference inspired by methods used in human developmental studies. Our method, described in more detail in Section 3.2, is based on language modeling probability (i.e., predicting word probability given a context: $P(w|context)$—this is analogous to *cloze probability* (Taylor 1953; Bloom and Fischler 1980)). Specifically, our test targets models that can assign a probability distribution over the vocabulary space for arbitrary positions (and not just the terminal position) in a given context. For example, a model that can produce an answer to the following question can be tested through our method: what is the probability that each word in the vocabulary of the model appears in the place of ___ in *I went to a* ___ *with my friend*?

The empirical question asked through this new test is: can we find evidence of abstract grammatical categories and category-based generalization in an ANN model that is trained on text data on the objective of predicting a word in context? In addition to sharing the general motivations described in Section 2.4, the experiments described in this chapter has the following potential contributions. From the perspective of cognitive science, finding evidence of category abstraction in ANN models can provide

an argument against the need for an innate bias towards categorization and/or pre-specification of the set of lexical categories for learners of language (Chomsky 1965; Grimshaw 1981; Pinker 1984, *i.a.*). If a model with neither priors successfully shows category abstraction, it would provide converging evidence to studies such as Cartwright and Brent (1997); Redington et al. (1993, 1998); Mintz et al. (2002); Mintz (2003) that show word distribution in naturally occurring English data (usually child-directed speech in CHILDES: MacWhinney 1995a,b; MacWhinney and Snow 1985) can inform the distinction of syntactic categories. Furthermore, the ANN study would be helpful in investigating the role and necessity of non-distributional (perceptual) cues that have been claimed to help category abstraction (e.g., phonological: Gerken et al. 2005; Monaghan et al. 2005, perceptual similarity: Gómez and Lakusta 2004) but are not reliably available in natural language. In particular, the ANN model that we use does not represent phonological information,[1] and therefore could help gauge the degree to which category abstraction can be performed in the absence of phonological cues.

From the perspective of Natural Language Processing (NLP), it is known that contemporary ANNs perform well (near 98% accuracy) on benchmarks[2] for English part-of-speech tagging (Bohnet et al. 2018; He and Choi 2019), and that diagnostic classifiers for 'probing' pretrained[3] models' knowledge of part-of-speech also achieve similarly high performance (about 97%) (Tenney et al. 2019). However, it still remains an open question whether pretrained models can make category-based generalizations with limited exposure to novel words, and without being explicitly trained to perform

[1] But it does have an impoverished, potentially inconsistent representation of morphology based on SentencePiece subword segmentation (Kudo and Richardson 2018).

[2] Typical benchmark datasets include the Wall Street Journal corpus from the Penn TreeBank (Marcus et al. 1993) and OntoNotes (Weischedel et al. 2013).

[3] *Pretraining* in the narrow sense refers to training a model (typically with an objective of language modeling, the objective of predicting a word in context) using a large amount of data to obtain a 'general purpose' model of language. This model is then *finetuned* to perform a specific target task (e.g., question-answering, natural language inference, machine translation...). Pretraining in the broader sense can also refer to any kind of training before the model is trained on the target task.

categorization. Our proposed method has a benefit in that it does not require training a separate classifier on top of the model being evaluated. This lets us bypass the methodological questions raised in the recent literature on the validity of using diagnostic classifiers as probes (Hewitt and Liang 2019; Voita and Titov 2020, *i.a.*). More generally, this work is in line with the current efforts to analyze the linguistic capacities of (pretrained) ANN models (Alishahi et al. 2019).

3.2 Test Format

Our method is inspired by the experimental design of Höhle et al. (2004), in which infants were familiarized to contexts containing novel words, and were tested with new sentences that either obeyed or violated category-based co-occurrence restrictions using a head-turn preference procedure (Jusczyk and Aslin 1995; Kemler Nelson et al. 1995). They selected two monosyllabic nonce words[4] *glamm* and *pronk*,[5] and first exposed the infants to these words in contexts that either signaled that the nonce word is a noun (1-a) or a verb (1-b). In the noun-signaling context, the nonce word was combined with the indefinite article *ein*, suggesting that the nonce word is a (masculine or neuter) noun. In the verb-signaling context, the nonce word was combined with the pronoun *sie*, suggesting that the nonce word is a verb. After this first exposure (the *familiarization phase*), the infants were exposed to two passages each per nonce word, where in one passage the nonce word was consistently used as a noun and in the other, it was used as a verb. In these passages, the nonce words did not appear with the words that were used as signals in the familiarization phase (i.e., *ein* and *sie*). The perceived difference between the passages was quantified in terms of listening times, measured using the head-turn preference procedure. The infants were either assigned to the noun signaling familiarization condition or the verb signaling familiarization condition.

[4]This was to remove the potential phonological and/or prosodic cues to category inference.
[5]The experiments were conducted in German.

The prediction is, if the infants inferred the grammatical category of the nonce words during the familiarization phase, there would be a difference in listening times of the two passages where the distribution of the nonce word matches or mismatches the distribution of the inferred category.[6]

(1) a. NOUN-SIGNALING: *ein Glamm*/*ein Pronk* ('a glamm'/'a pronk')

b. VERB-SIGNALING: *sie glamm*/*sie pronk* ('she glamms'/'she pronks')

We reformulate this method to one that uses cloze probability as the difference measure instead of listening time, in order to make it applicable to the model we plan to test (details about the model that we test is discussed in Section 3.4.1). Specifically, this reformulated version of the test of Höhle et al. (2004) is applicable to a class of models that assign a probability distribution over the space of the vocabulary given a context. If such a model makes a valid category inference about a novel word from the context that it appears in in the familiarization phase, the model should be able to assign a higher probability to that word in new contexts that obeys the co-occurrence restriction for that category, over a word of a different category. For example, if the model is exposed to unseen words *blick* and *dax* in contexts that signal distinct category membership (2), they should be able to make a generalization that in (3-a), the word that fills the blank is more likely to be *blick* than *dax*. On the other hand, in (3-b), *dax* should be more likely than *blick*. To summarize, we expect the cloze probability to follow the pattern in (4) if the model correctly inferred that *blick* is a noun and *dax* is a verb.

[6] I refer the readers to the original paper of Höhle et al. (2004) regarding their results and interpretation. When there was a significant effect of the familiarization condition, the infants listened *longer* to the passage that disobeyed the co-occurrence restriction of the category inferrable from the familiarization phase (i.e., a novelty effect). Their findings about differences regarding 12–13 month old children and 14–16 month old children are very interesting, but here we focus more on their methodology. The assumption is that adult speakers would be able to infer the grammatical categories of nonce words given signaling examples like (1). In English, that would be, given *a glamm*, speakers will be able to infer that *glamm* is a noun, and given *we glamm*, speakers will be able to infer that *glamm* is a verb. Though, see Section 3.5 for a more nuanced discussion of the expected behavior of adult speakers.

(2) a. NOUN-SIGNALING: *the blick*

b. VERB-SIGNALING: *they dax*

(3) a. NOUN-EXPECTING: *I went to a* ___.

b. VERB-EXPECTING: *I* ___ *with some friends.*

(4) a. $P(\textit{blick} \mid \textit{I went to a \_}) > P(\textit{dax} \mid \textit{I went to a \_})$

b. $P(\textit{dax} \mid \textit{I \_ with some friends}) > P(\textit{blick} \mid \textit{I \_ with some friends})$

Note that our adaptation differs from the original methodology of Höhle et al. (2004) in that a single model is familiarized to two novel words pertaining to distinct grammatical categories, whereas in the original experiment the infants were randomly assigned a familiarization condition of either noun or verb, so each infant was only exposed to signal contexts of a single category.

3.3 Dataset Generation

We constructed the signal contexts (like (2)) and test contexts (like (3)) from sentences in the Multi-Genre Natural Languange Inference (MNLI; Williams et al. 2018) dataset. We selected MNLI because (1) we wanted to ensure that the contexts had different sources from the data that the model we planned to test (Section 3.4.1) was already exposed to (i.e., English Wikipedia and BooksCorpus (Zhu et al. 2015) data), and (2) the MNLI dataset contains pre-generated constituency parses (from the parser of Klein and Manning 2003) that is useful for a first-pass identification of grammatical categories.[7]

Signal contexts: Two sentences with one novel word each (let's call them w_1 and w_2)—each sentence providing a context that unambiguously signals the novel word's

[7]We used MNLI here, but as long as the first criterion is satisfied, other corpora would be equally valid choices.

grammatical category—constituted a set of signal contexts. The two contexts matched in the number of words and the linear position of the unseen word from both left and right. For example:

(5) a. NOUN-SIGNALING: *A w_1 needs two people.*

b. VERB-SIGNALING: *She w_2 at the group.*

Test contexts: For testing, we sampled sentences from MNLI (excluding the genres that were transcriptions of speech) that contained a word in the same grammatical category as w_1 and w_2, respectively, and replaced the word with a placeholder as in (6). The test context, like the signal contexts, were sentences in which the category of the placeholder is unambiguous, but they did not have any lexical overlap with either of the signal contexts. This was to prevent the models from using lexical overlap as a superficial heuristic to assign higher probability to the novel words without relying on grammatical category membership. Furthermore, we only selected sentences that contained a different number of subword tokens from the signal contexts—this way, the placeholders in the test contexts always appeared in different linear positions from the novel words in the signaling contexts, both from left and right ends of the sentence.

(6) a. NOUN-EXPECTING: *Keep everyone else company by sitting in the ___.*

b. VERB-EXPECTING: *The colonel ___ us to a hotel.*

By applying the above generation method, we created six English datasets (i.e., a set of signal and test contexts) that test for the binary distinguishability between four open-class grammatical categories: noun, verb, adjective, and adverb[8] ($_4C_2 = 6$). Each dataset included 2 signal contexts and 400 test contexts.[9] Since we used

[8]We used the Stanford parser tags 'NN' for noun, 'VB' and 'VBD' for verb, 'JJ' for adjective, and 'RB' for adverb to identify the grammatical categories.

[9]This number of signal contexts is substantially larger than the two six-sentence passages used in

the automatically generated syntactic parses provided in MNLI to determine the grammatical categories, we manually verified the sentences after generation to rule out parser errors and contexts that were ambiguous between the two categories being compared. For instance, (7) gives an example of a context that would be ruled out in a verb/adjective distinction dataset due to its category ambiguity.

(7) a. *Cookie the cat is* ___.

b. VERB CONTINUATION: *Cookie the cat is dancing.*

c. ADJECTIVE CONTINUATION: *Cookie the cat is healthy.*

Proposed metric: As discussed in Section 3.2, the model's ability to infer the grammatical categories of the novel words can be evaluated by comparing the probabilities of the two novel words in multiple test contexts like (6). We considered the model's category inference to be accurate if the model assigns higher probability to the novel word in the test context that obeys the co-occurrence restriction of the signaled category (e.g., higher probability for w_1 over w_2 in (6-a), and vice versa in (6-b)), and calculated the accuracy by dividing the total number of correct test contexts by the number of total test contexts.

3.4 Experiment

3.4.1 Model

We tested the BERT-large model (Devlin et al. 2019), which is a Transformer model (Vaswani et al. 2017) pretrained on English Wikipedia data (~2500M words) using the masked language modeling objective (MLM).[10] MLM is a predictive objective, where a model is required to predict the probability distribution over the token '[MASK]' in

Höhle et al. (2004). We expanded the number of signal contexts to reduce potential semantic effects of individual contexts affecting the results.

[10]BERT models are also pretrained on the next sentence prediction objective on BooksCorpus data, but this objective is less relevant for the discussion in this section.

a given context, as in (8):

(8) *The cat sat on the* [MASK].

Here, the model may assign a very high probability value to the word *mat* based on frequent collocations, and may also assign decent probability to other words that denote things a cat can plausibly sit on, like *cushion* or *human*.

We used a version of the model that uses whole word masking instead of random subword token masking as described in the original paper by Devlin et al. (2019).[11] The BERT model uses a segmentation unit of subwords, which may not always correspond to valid morpheme boundaries. Hence, if trained on the MLM objective with subword token masking, they would be trained on predicting fragmented masked tokens in sentences like (9-b), because BERT's tokenizer splits the word *kitten* into subwords *kit* and *##ten*.[12] Whole word masking adopts a masking scheme that masks all of the subword tokens that constitute a single word as in (9-c).[13]

(9) a. ORIGINAL: *The kitten is very cute.*

b. MASKED: *The* [MASK] *##ten is very cute.*

c. WHOLE WORD MASKED: *The* [MASK] [MASK] *is very cute.*

3.4.2 Training

We used the two signal contexts (5) to first familiarize the BERT-large model to these contexts. Then, we used the test contexts (6) to evaluate whether expected differences

[11] We used model checkpoints and code provided by Wolf et al. (2020): https://github.com/huggingface/transformers.

[12] ## denotes a subword boundary. See the repository description in https://github.com/google-research/bert for more examples.

[13] However, the model is still required to predict the masked tokens in sentences with multiple masked tokens independently even under the whole word masking setup. This means that although the prediction would not be conditioned on false segmentations like (9-b), but the model would still be required to predict *kit* and *##ten* separately given (9-c). In our dataset generation, we only selected words that consist of single subword tokens to avoid the need to make multiple predictions like this in any part of the experiment.

were observed in the probability assignment of the two novel words shown to the models during the familiarization phase. Half of the 400 test examples were used as the development set, and half were used for conducting the final evaluation.

For the signal contexts to function as intended, we need them to contain words that the model has never been exposed to before. To represent these novel words, we used *unused tokens* (called '[unused$n$]') in the vocabulary of BERT-large.[14] These tokens are special tokens that never appear in the data that the model was pretrained on. So w_1 and w_2 in (5) were replaced with these unused tokens, as in (10).

(10) a. NOUN-SIGNALING: *A* [unused1] *needs two people.*

b. VERB-SIGNALING: *She* [unused2] *at the group.*

For the test contexts, we placed the mask token '[MASK]' in the position of the placeholder in (6), which is the position for which we want the model to assign a probability distribution. (11) shows some examples:

(11) a. NOUN-EXPECTING: *Keep everyone else company by sitting in the* [MASK].

b. VERB-EXPECTING: *The colonel* [MASK] *us to a hotel.*

Before familiarizing the model to the signal contexts, we froze the entire model at the state of its release—that is, after it has been pretrained on Wikipedia using the MLM objective—except for the embeddings of the two unused tokens the models would be exposed to during the familiarization phase (10).[15] We trained[16] this frozen-except-two-words model on the two signal contexts for 70 epochs (i.e., 70 exposures to each of the signal contexts), using the same MLM objective as its pretraining, checkpointing

[14] https://huggingface.co/bert-large-uncased-whole-word-masking/blob/main/vocab.txt

[15] This amounts to asking the question: can the novel words be placed in a pre-constructed space that enables category-based generalization?

[16] This is the step typically known as *finetuning* (see Footnote 3) in the NLP literature. Though here, our finetuning slightly differs from usual finetuning in that most of the model parameters were frozen.

after each epoch. Then, we selected the checkpoint to evaluate based on performance (measured by the accuracy metric described in Section 3.3) on the development set.

Category 1	Category 2	Accuracy	Accuracy (1 > 2)	Accuracy (2 > 1)
N	V	88.1 (±6.6)	87.2	89.0
N	Adj.	83.1 (±5.7)	86.2	80.0
N	Adv.	67.3 (±7.8)	63.0	71.6
V	Adj.	87.3 (±5.0)	88.4	86.2
V	Adv.	78.7 (±13.9)	80.2	77.2
Adj.	Adv.	71.2 (±10.2)	60.6	81.8

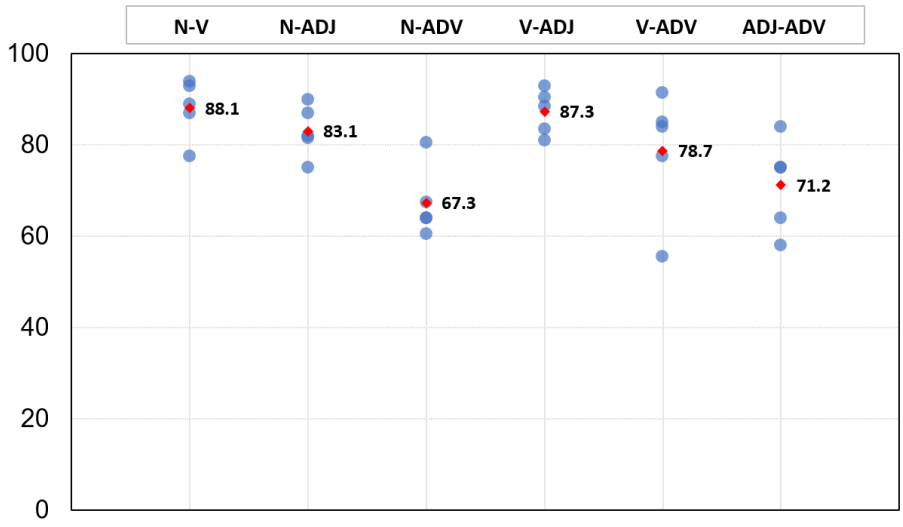


Table 3-1. Accuracy (%) of distinguishing two grammatical categories in novel contexts, averaged over five random seeds (individual accuracies are shown in the bottom figure). 'Accuracy(1 > 2)' denotes the accuracy on the set of sentences where Category 1 should be preferred over Category 2 (e.g., assigning higher probability to a noun in a noun-expecting context for row 1), and vice versa. Column 'Accuracy' lists the aggregate accuracy.

3.4.3 Results

The third column of Table 3-1 shows the generalization accuracy for each category pair in the six binary comparisons of four open-class grammatical categories (noun, verb, adjective, and adverb). The final two columns show the division of total accuracy by expected category (i.e., in a noun versus verb distinction test, the columns denote accuracy on noun-expecting contexts and accuracy on verb-expecting contexts, respectively). If the models were guessing randomly, the values for all three columns

would be 50%. If the models were blindly preferring one nonce word over another regardless of context, the overall accuracy would still be 50%, but one of the by-expected category columns would display 100% accuracy and the other, 0%, depending on which nonce word was preferred. A model with perfect category distinction would show a value of 100% for all three columns. As for the BERT-large model we tested, all accuracy was significantly above chance ($p < .05$, one proportion z-test, with 50% as the null hypothesis). This suggests that this model showed nontrivial generalization to novel contexts based on grammatical categories, even in the absence of phonological cues, and without being provided any prior information about the grammatical categories tested.

Category 1	Category 2	Acc. w/o training (epoch 0)	Acc. after 1 epoch
N	V	45.5	59.7 (±13.2)
N	Adj.	61.0	46.0 (±24.4)
N	Adv.	41.0	55.5 (±9.3)
V	Adj.	26.5	55.1 (±13.8)
V	Adv.	53.5	51.2 (±3.0)
Adj.	Adv.	37.0	50.5 (±2.4)

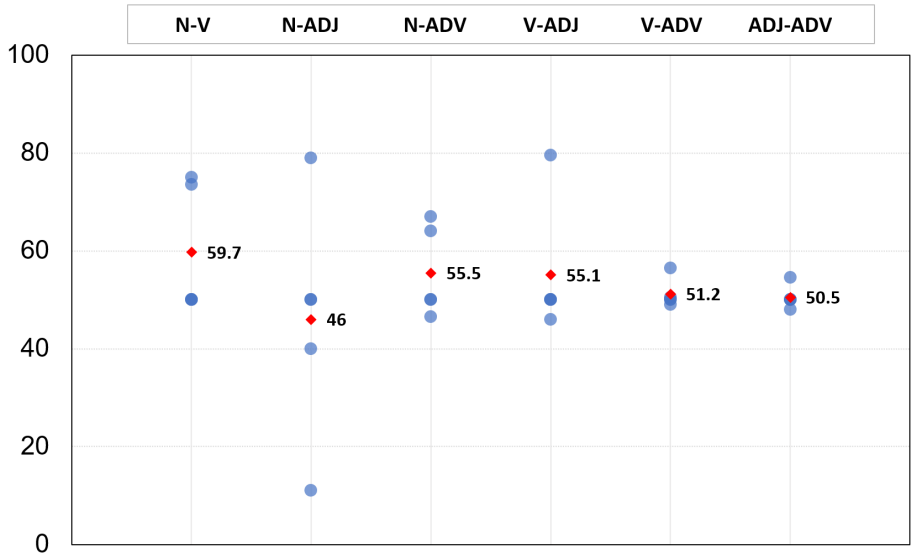


Table 3-2. Accuracy of distinguishing two grammatical categories in novel contexts after a single exposure to signal contexts.

However, the model's generalization capacity was still limited in several aspects.

First, while the accuracy values were all above chance, some distinctions were still much weaker than others (e.g., noun vs. adverb). Second, the model failed to display rapid category inference as competent English speakers often can from even a single exposure (e.g., many adult readers of this work would have inferred that *blick* in *A blick danced* is a noun by reading it just once). On the other hand, the reported accuracy in Table 3-1 was only reached after many iterations of the familiarization examples—on average 51 epochs with an initial learning rate as high as 5.[17] While developmental studies also typically expose subjects to familiarization examples multiple times, the number of iterations is rarely as high as 51. In light of this observation, we additionally show the category inference accuracy of the model after a single exposure to the familiarization examples in Table 3-2 for comparison. We can see that the models are on average unable to make accurate category distinctions after a single exposure, with performance at around chance on all distinctions except for the noun versus verb distinction.[18] Finally, there was high variation in accuracy over random seeds, which is in line with the general observation of high degree of instability over random seeds, especially in tasks that require generalization outside of the training distribution (Kim and Linzen 2020; McCoy et al. 2019, *i.a.*).

3.4.4 Effect of Embedding Initialization

In obtaining the results shown in Figure 3-1, we had randomly selected two tokens from the 1000 '[unused$n$]' tokens available in the vocabulary of BERT to represent the novel words being learned (specifically, '[unused89]' and '[unused11]'). In the model, the embeddings of these unused tokens are randomly initialized and remain unchanged during pretraining. Still, the actual randomly initialized values of different

[17] We used the AdamW optimizer (Loshchilov and Hutter 2019) with a constant schedule.

[18] Some outlier random seeds did achieve nontrivial category inference, with around 80% distinction accuracy after only a single exposure. Nevertheless, most runs were at exactly 50% accuracy, following the 'blind preference' pattern mentioned in Section 3.4.3. This suggests that during the early iterations of the familiarization phase, the models tend to show a strong preference for one novel word over another regardless of context, and category-based preference emerges in the later iterations.

'[unused$n$]' tokens differ, and the choice of the particular n may be a source of substantial variability. To investigate the effect of the particular choice of the tokens representing the novel words (i.e., the initialization of the embeddings), we reran the whole experiment four additional times, selecting different '[unused$n$]' tokens for the novel words each time. The variation in the mean accuracy depending on the random selection of unused tokens is shown in Figure 3-1.

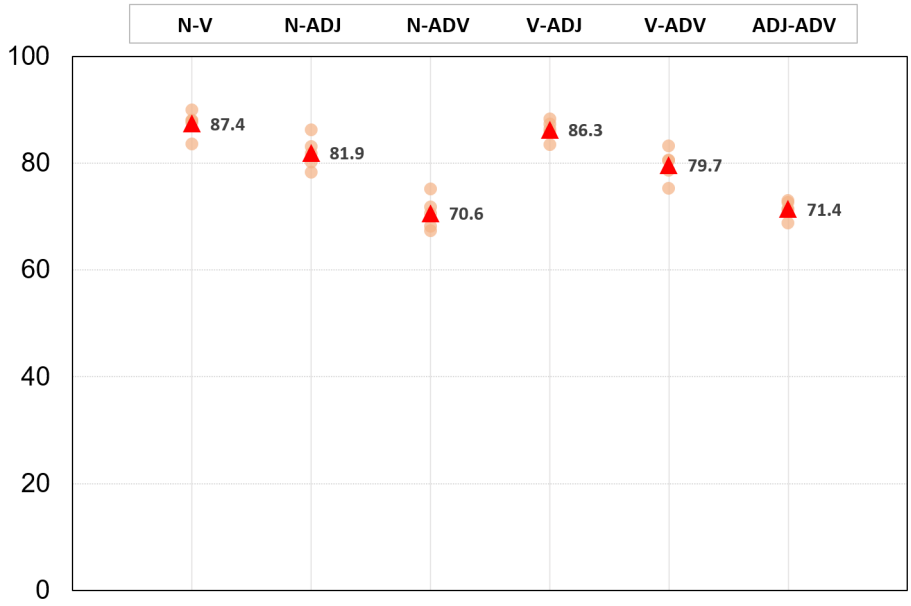


Figure 3-1. Variation in the mean accuracy across five experiment reruns with different '[unused$n$]' tokens to initialize the two novel words being learned. Each dot represents the mean accuracy of an experiment. That is, a single dot corresponds to a single number under the 'Accuracy' column in Table 3-1, which itself is an average over five random seeds).

3.5 Conclusion and Remaining Questions

We proposed a method for testing category-based generalization in models that predict cloze probabilities, and tested BERT-large with this method. Our results show that this model achieves nontrivial success on making such generalizations, suggesting that there is promise for contemporary ANN models to meet the precondition of grammatical category inference for compositional generalization (the test for which is discussed in the next chapter). However, the model's generalization capacity was

not without limitations: in addition to the weak distinguishability of some open-class categories, BERT did not display rapid category inference. It was only able to achieve the reported performance after many repeated exposures to the familiarization examples. How this gap could be closed, we leave to future work. One remaining comment based on the learning patterns we saw in this experiment is that, gradient descent-based optimization might not be the most appropriate way to approach novel word learning. Especially if our desideratum is for learning and category inference to be rapid, developing an informative initialization algorithm for the initial representation of novel words may be a promising direction.

Additional limitations of the methodology and remaining questions are discussed below.

Is the model's success actually driven by abstraction? While our method does test the generalization capacity of BERT regarding the usage of novel words outside of the contexts that they were observed in, further analysis is needed in terms of how the generalization is achieved. This will elucidate whether the current success is actually driven by abstraction. In its current form, good performance on our test is a necessary condition for abstraction-based generalization but not sufficient. That is, a model that has category abstraction (for the categories we tested) would achieve high accuracy on our test, but it is too strong to claim that all models that achieve high accuracy on our test did so by the means of category abstraction. For instance, we could imagine a scenario where a model achieves success on generalization without abstraction by analogy to a single exemplar that is not part of any subspace representative of the relevant grammatical category. One way we could tease apart a true case of category-based abstraction would be by examining whether there exists a subspace (rather than a single point in space) of embeddings that gives rise to similar degrees of success on our generalization test.

Status of grammatical categories tested (and need for human subject experiments): In our test, we used the categories noun, verb, adjective, and adverb. While this set of categories is a good starting point, we note that the status of these categories is part of an ongoing debate. For example, whether adverb is a standalone category, or adjectives and adverbs should be considered a single category has been contested (Payne et al. 2010; Baker 2003). Furthermore, even within a single category, there are various subcategories within which the members are more distributionally similar to each other (e.g., Levin classes for verbs (Levin 1993), adverbs that occupy the specifier positions of different functional projections (Cinque 1999)). Whether we expect a generalization pattern like (4) within/across subcategories or not is an open question that should be addressed through human subject experiments. To the best of our knowledge, an experiment like ours (or like Höhle et al. (2004)'s) has not been conducted with adult human subjects, and also has not been used to compare the distinguishability of categories other than noun versus verb. We believe an analogous experiment to the one used to test ANN models in this work[19] would further our understanding of the nature of grammatical categories in human learners, as well as better contextualizing the ANN results presented here.

[19]We would need to use a different preference measure than cloze probability, since we cannot directly obtain a probability distribution from human subjects unlike the ANN model we tested. One possible method would be a binary forced-choice between w_1 and w_2 for blanks in test contexts (6).

Chapter 4

A Test for Compositional Linguistic Generalization

4.1 Motivation

A widely accepted view about natural language is that it allows for production and interpretation of novel complex expressions through composition of their constituent parts (Section 2.1.1). Hence, one way to test whether an intelligent system has a command of a certain language is to test whether it can make valid generalizations predicted by the compositional rules in the language (Section 2.1.4). Here, we propose a methodology to test such *compositional linguistic generalizations* that can be applied to evaluate any system or model that takes a sequence of symbols as its input and produces a sequence of symbols as an output. We use a setup where a model is asked to assign a semantic representation to a given sentence (i.e., semantic parsing).

After we propose the test, we use the test to evaluate several configurations of two contemporary artificial neural networks: Long Short-Term Memory (Hochreiter and Schmidhuber 1997) in a sequence-to-sequence setup[1] (Sutskever et al. 2014) and Transformer (Vaswani et al. 2017). These models have driven the recent progress made

*An earlier version of this chapter has been published as Kim and Linzen (2020) in the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.

[1]This setup can be understood as performing sequence transformation; a model takes as input a sequence of symbols and produces as output a sequence of symbols.

in the field of Natural Language Processing (NLP), and have been shown to display nontrivial generalization patterns that rely on the command of hierarchical linguistic structures (e.g., Gulordava et al. 2018; Goldberg 2019). The choice of models was partly due to this empirical promise (see the discussion in Chapter 1 regarding the utility of models that match the generalization patterns of human learners. However, we also intend the first set of experiments presented here to serve as a reference point so that the effect of modifications or augmentations can be estimated. This is why, for instance, we are not starting the experiments with models that are already trained on a lot of data, even though such models have even greater empirical promise. Since the additional training adds another source of inductive bias on top of the architecture and other model configurations, having a set of reference results first would be helpful in teasing apart the effect of the additional layer of training. In addition, it should be emphasized that the set of experiments we present here cannot be fully representative of the architectures—there is a wide space of possible configurations within the class of models that fall under the label LSTM and Transformer, and it remains to be seen whether the our findings will generalize more broadly.[2] See Section 2.4 for a more detailed discussion of the broad motivation behind testing the generalization capacity of artificial neural networks.

4.2 Test Format

Our compositional generalization test is in the form of semantic parsing—a task of translating natural language expressions into a semantic representation language. For example, in (1), given the sentence on the left hand side as input, the test subject must output its corresponding semantic representation: the logical formula on the right hand side ($\rightsquigarrow$ means 'translates to'). The specifics of the semantic representation

[2]Thus, any mention of 'LSTM and Transformer models' in the context of our experimental results must be prefaced with this disclaimer (or read as shorthand for 'the particular configuration of LSTM and Transformer models we tested').

language is discussed in detail in Section 4.4.2).

(1) A cat saw a dog. $\rightsquigarrow$ $cat(x_1)$ *AND* $see.agent(x_2, x_1)$ *AND* $see.theme(x_2, x_4)$ *AND* $dog(x_4)$

The dataset for the task consists of examples (⟨sentence, semantic representation⟩ pairs) divided into training and generalization sets. The set of examples used for training contains various systematic gaps that, during evaluation, must be filled via compositional generalization (the various systematic gaps are discussed in Section 4.3). All generalization examples require the models to translate sentences that they have never encountered during training into the semantic representation language.

There are two broad categories of generalizations included in our test: *lexical* and *structural*. Lexical generalizations concern translating expressions with known primitives in syntactic configurations that they have not been observed in. Neither the primitives nor the syntactic configurations are novel—they can all be found as parts of the training examples. What is novel is the *combination* of the primitive and the syntactic configuration it appears in—there is no example in the training set that provides a translation for such a combination. Structural generalizations concern translating expressions with novel syntactic structures, where all of the sub-parts of the structure that are necessary to translate such a structure are contained in the training data.

The way that the models can achieve generalization, in the abstract, is to deduce the compositional rules of the language and build a semantic representation from those. However, reconstructing the exact compositional rules and performing each composition step starting from the primitive meanings might not be necessary to arrive at the desired generalization. Instead, for example, the models may observe that some parts of the input correspond to some parts of the output, and this correspondence relation systematically holds across all training examples. Then, by piecing together parts that they have observed the input-output correspondence of, they can arrive at

the correct semantic representation.

Metric: For individual examples, we use exact string match accuracy between the target output and the output produced by the model tested to determine whether the model's output is correct. The aggregate metric reported for each model as a measure of generalization performance is the number of examples that had perfect string match over the total number of examples. We additionally report separate accuracy values for lexical and structural generalization.

4.3 Generalizations Tested

There are various generalization patterns that human speakers display that can be attributed to the compositional rules of the language (Section 2.2). We describe below the generalizations that are included in our test.

4.3.1 Novel Combination of Familiar Primitives and Grammatical Roles

Competent speakers of English can easily interpret a familiar primitive in a grammatical role that is different from the one in which it was previously observed. For example, a noun that have only been observed as a part of a subject noun phrase (NP) can easily be interpreted in a direct object NP. This generalization capacity has been attested in children as young as 20 months old (Tomasello and Olguin 1993). We ensure that in the training set some lexical items only appear in the subject position (2-a), and some only appear in the direct object position. In the generalization set, these lexical items appear in the alternate grammatical role (2-b).

(2) a. TRAINING: A **hedgehog** ate the cake.

b. GENERALIZATION: Alex saw a **hedgehog**.[3]

[3]Note that all of the non-bolded primitives (*Alex, saw, a*) are also part of the training data, although other examples containing them are not shown here to highlight only the targeted systematic

We test for generalization to the targeted grammatical roles not only in simple sentences, but also *embedded* clauses; this form of generalization is a defining criterion of *strong systematicity* (Hadley 1994). For instance, a noun that only appeared in the subject of a simple sentence in training (3-a) may occur in the object of an embedded clause in the generalization set (3-b).

(3) a. TRAINING: A **hedgehog** ate the cake.

b. GENERALIZATION: A girl said that Emma called the **hedgehog**.

While some primitives appear in the training set in the context of a sentence as above, others only occur in isolation. The semantic representations of common nouns are unary predicates (*shark* $\rightsquigarrow \lambda x.\text{shark}(x)$), proper nouns are represented as constants (*Emma* $\rightsquigarrow$ Emma), and verbs are represented as n-ary predicates with thematic role specifications (*like* $\rightsquigarrow \lambda x.\lambda y.\lambda e.\text{like.agent}(e, y)$ AND $\text{like.theme}(e, x)$). The training set contains these primitives as isolated words (4-a), but not as a part of a sentence; by contrast, the generalization set includes examples that require translating these primitives in context (4-b).

(4) a. TRAINING: **shark**

b. GENERALIZATION: The **shark** smiled.

4.3.2 Novel Combination of Modified Phrases and Grammatical Roles

Phrases with a modifier, such as an NP modified by a prepositional phrase (PP), can occupy the same grammatical roles as unmodified phrases. For example, just like

differences between training and generalization. That is, the training set would contain other examples like *Alex danced, The butterfly saw a flower* in addition to *A hedgehog ate the cake*, so that all the information necessary to generalize is provided in the training set. However, the distribution of these primitives do not systematically vary across training/generalization: for example, *Alex* may appear as both subject and object in training. This footnote applies to all other illustrative examples in this section.

$[the\ cat]_{NP}$, the phrase $[[the\ cat]_{NP}\ [on\ the\ mat]_{PP}]_{NP}$ is an NP, and can occupy the same syntactic positions. Speakers of English are most likely not exposed to modifiers in every possible syntactic position that the modified element may occur. Yet, the phrasal modification rule is considered to be general (e.g., NP → NP PP) rather than specific to different grammatical roles (e.g., NP_{obj} → NP_{obj} PP, NP_{subj} → NP_{subj} PP). To test for generalization to modifiers in an unseen grammatical role, our training set includes only examples with PP modifiers of object NPs (5-a), and the generalization set contains PP modifiers of subject NPs (5-b). We note that this is a simplification of the generalization problem that human learners may encounter; see Section 4.7.2 for a further discussion.

(5) a. TRAINING: The cat saw **the rat on the mat**.

b. GENERALIZATION: **The cat on the mat** saw the girl.

4.3.3 Deeper Recursion

Recursion is considered to be an important property of human linguistic competence (Hauser et al. 2002). Human language achieves this property by allowing certain phrase types to be nested within a phrase of the same type. For example, in English, complementizer phrases (CPs) can be nested: for instance, [*Mary knows* [*that John knows* [*that Emma cooks*]$_{CP}$]$_{CP}$]$_{CP}$. Our dataset includes two types of recursive constructions that are usually viewed as allowing arbitrary depths of nesting: sentential complements (nested CPs; (6)) and nominal PP modifiers (nested PPs; (7)). The training set contains nestings of depth 0–2, where depth 0 is a phrase without nesting. The generalization set contains nestings of strictly greater depths (3–12).

(6) a. TRAINING: Emma said **that** Noah **knew** that the cat danced.

b. GENERALIZATION: Emma said **that** Noah knew **that** Lucas saw **that** the cat danced.

(7) a. TRAINING: Ava saw the ball **in the bottle on the table**.

b. GENERALIZATION: Ava saw the ball **in the bottle on the table on the floor**.

Whether unbounded recursion should be considered as an intrinsic part of the linguistic machinery is a debated issue, the evidence against being the significant human processing limitations on center-embedding constructions (see Christiansen and MacDonald 2009 for an overview and an alternative account). In our dataset, we only included structures that are traditionally thought of as recursive (i.e., left- or right-recursion), but could be implemented by iteration and can be processed by a Finite State Machine (Reich 1969; Christiansen 1992).

4.3.4 Verb Argument Structure Alternation

Many English verbs participate in argument structure alternations (Levin 1993). For instance, *break* can be used both as a transitive verb (*John broke the window*), and as an unaccusative verb, with its theme in the subject position (*The window broke*). Likewise, agent-patient verbs can passivize; *John broke the window* can be passivized to *The window was broken*, or with an optional agent *by*-phrase, *The window was broken by John*. These alternation patterns are not strictly restricted to particular lexical items (although there are constraints), and speakers of English often extend such alternation patterns to verbs that have only been observed in one of the alternate forms. To illustrate, if I told the readers that *I floosed the cat* means "I fed the cat twice", hopefully, the interpretation of *The cat was floosed* would also be immediately available (though see Section 4.7.1 for a caveat).

The dataset contains alternation patterns that humans have been shown experimentally to generalize to nonce verbs: active-passive (Brooks and Tomasello 1999), transitive-intransitive (unaccusative and object-omitted transitives: Ono and Budwig

2006; Hu et al. 2007; Kline and Demuth 2014), and the alternation between double object datives and prepositional phrase datives (Conwell and Demuth 2007). For each of these alternation patterns, we include only one of the alternating forms (e.g., active) in the training set, and only the other form (e.g., passive) in the generalization set (8).

(8) a. TRAINING: The crocodile **blessed** William.

b. GENERALIZATION: The muffin **was blessed**.

4.3.5 Verb Class

In English, the semantic role of the argument of a verb with a single argument depends on the property of the individual verbs; the surface syntax of the sentence is not enough to determine its interpretation. For example, *froze* in the sentence *The lake froze* is an unaccusative verb, which takes a theme (or patient) as its grammatical subject, whereas in *The dog smiled*, *smiled* is an unergative verb that takes an agent as its grammatical subject. Inspired by this property, we include in our generalization set combinations of verbs and NPs, which all occur separately in the training set, but such that the NPs never appear as the thematic role specified by the verb in the training set. For instance, the training set contains a sentence with *cobra* as an agent subject (9-a), and sentences with unaccusative verbs (9-b), and the generalization set contains examples in which *cobra* and *freeze* appear together (10). Correctly mapping *cobra* to the theme, even though it only appears in the training set as an agent, requires sensitivity to the argument structure of *freeze*.

(9) TRAINING

a. A cobra helped a dog. →

cobra(x_1) AND help.agent(x_2,x_1) AND help.theme(x_2,x_4) AND dog(x_4)

b. The drink froze. →

*drink(x_1) AND freeze.theme(x_2,x_1)

(10) GENERALIZATION

The cobra froze. →

*cobra(x_1) AND freeze.theme(x_2,x_1)

Table 4-1 summarizes all of the generalization patterns tested in our dataset.

4.4 Dataset Generation

In this section, we discuss the details of the dataset construction. Each example in the dataset is a ⟨sentence, semantic representation⟩ pair. The sentences are sampled from a hand-designed grammar (Section 4.4.1) and automatically mapped onto the corresponding semantic representation (Section 4.4.2).

4.4.1 Grammar

We generated the constructions described in Section 4.3 using a Probabilistic Context-Free Grammar (PCFG).[4] The types of constructions covered by our PCFG are as follows (the names of the construction types are taken from Roland et al. 2007).

- Simple Intransitive
- *To* Infinitive Verb Phrase
- Sentential Complement[5]
- Simple Transitive
- Ditransitive
- Passive

[4]It is debatable whether all expressions in English can be captured by a context-free grammar. For example, phrasal reduplication seems to be allowed in English: *I was thinking about this dissertation yesterday but not thinking about it thinking about it.* But for our purposes, a context-free grammar will suffice.

[5]We only included sentential complements with a complementizer *that*.

Case	Training	Generalization
Section 4.3.1. Novel Combination of Familiar Primitives and Grammatical Roles		
Subject → Object (common noun)	A **hedgehog** ate the cake.	The baby liked the **hedgehog**.
Subject → Object (proper noun)	**Lina** gave the cake to Olivia.	A hero shortened **Lina**.
Object → Subject (common noun)	Henry liked a **cockroach**.	The **cockroach** ate the bat.
Object → Subject (proper noun)	The creature grew **Charlie**.	**Charlie** worshipped the cake.
Primitive noun → Subject (common noun)	**shark**	A **shark** examined the child.
Primitive noun → Subject (proper noun)	**Paula**	**Paula** sketched William.
Primitive noun → Object (common noun)	**shark**	A chief heard the **shark**.
Primitive noun → Object (proper noun)	**Paula**	The child helped **Paula**.
Primitive verb → Infinitival argument	**crawl**	A baby planned to **crawl**.
Section 4.3.2. Novel Combination Modified Phrases and Grammatical Roles		
Object modification → Subject modification	Noah ate **the cake on the plate.**	**The cake on the table** burned.
Section 4.3.3. Deeper Recursion		
Depth generalization: Sentential complements	Emma said **that** Noah knew **that** the cat danced.	Emma said **that** Noah knew **that** Lucas saw **that** the cat danced.
Depth generalization: PP modifiers	Ava saw the ball **in the bottle on the table**.	Ava saw the ball **in the bottle on the table on the floor**.
Section 4.3.4. Verb Argument Structure Alternation		
Active → Passive	The crocodile **blessed** William.	A muffin **was blessed**.
Passive → Active	The book **was squeezed**.	The girl **squeezed** the strawberry.
Object-omitted transitive → Transitive	Emily **baked**.	The giraffe **baked a cake**.
Unaccusative → Transitive	The glass **shattered**.	Liam **shattered** the jigsaw.
Double object dative → PP dative	The girl **teleported** Liam the cookie.	Benjamin **teleported** the cake **to** Isabella.
PP dative → Double Object Dative	Jane shipped the cake to John.	Jane shipped John the cake.
Section 4.3.5. Verb Class		
Agent NP → Unaccusative subject	The **cobra** helped a dog.	The cobra **froze**.
Theme NP → Object-omitted transitive subject	The hippo **decomposed**.	The hippo **painted**.
Theme NP → Unergative subject	The hippo **decomposed**.	The hippo **giggled**.

Table 4-1. A full list of generalization cases. Each sentence in the table represents a ⟨sentence, semantic representation⟩ pair. For instance, the sentence *A hedgehog ate the cake* represents the following input-output mapping: *A hedgehog ate the cake* $\rightsquigarrow$ *cake(x_4) ; hedgehog(x_1) AND eat.agent(x_2,x_1) AND eat.theme(x_2,x_4). "Subject" and "Object" include subjects and objects of both simple and embedded sentences.

The PCFG assigns uniform probability (about 5%) to each frame (e.g., transitive verb with both subject and object, transitive verb with only subject, passivized transitive with subject only, passivized transitive with subject and agent *by*-phrase...) except for CP embedding constructions, whose probability was increased to about 8% to match their distribution in natural corpora.[6]

Distributional cues to verb subcategorization: There are several distributional cues in the dataset that allow syntactically ambiguous verb subcategories to be distinguished. Object-omitted transitives always have its transitive counterpart included in the dataset, whereas unergatives do not alternate and thus do not appear in any other frames. Unaccusative verbs appear with both animate and inanimate subjects, whereas unergatives and object-omitted transitives only appear with animate subjects (this is not always true in English). Verbs in different subcategories also have distinct primitive representations, some (but not all) of which were provided as part of the training set.

Selection of lexical items: We selected the 403 common nouns in our lexical inventory from the MacArthur-Bates Communicative Development Inventories (Fenson et al. 2007) and the British National Corpus (Leech et al. 2001). 100 proper nouns were selected from top baby names of 2019 in the United States according to the United States Social Security Administration. In selecting the verbs, we referred to Levin (1993) and Kipper-Schuler (2005). There were 113 unique verbs and 6 verb types, with some overlapping verbs across verb types (e.g., *like* with NP and CP arguments). The list of verb types are as follows:

- Verbs that take NP arguments that allow direct object omission (e.g., *eat*)
- Verbs that take NP arguments that do not allow direct object omission (e.g.,

[6]The assigned probabilities did not necessarily translate into the proportion in the generated dataset, since there were post-generation filtering mechanisms such as removing duplicate entries.

find)

- Subject control verbs that take infinitival arguments (e.g., *try*)
- Verbs that take CP arguments (e.g., *say*)
- Unaccusative verbs (e.g., *freeze*)
- Unergative verbs (e.g., *sleep*)
- Dative verbs (e.g., *give*)

5 common nouns, 3 proper nouns and 7 verbs used as primitive exposure examples (Section 4.4.3) were selected at the discretion of the author.

Lexical probability: The grammar assigns Zipfian probability distribution (inverse rank-frequency distribution) over lexical items in each noun and verb subcategory.[7] This was done in order to ensure that all possible grammatical patterns that a lexical item could appear in were sampled by the PCFG and included in our dataset, for at least the most frequent items in the class (e.g., both forms of the object omission alternation are sampled for the most frequent verb).

PP attachment ambiguity: Our grammar does not generate VP-modifying PPs (the only PP verbal dependents are recipient *to*-phrases, which are always arguments rather than modifiers). Therefore, all PP modifiers in our dataset should strictly have an NP-attachment reading, although for human readers VP-attachment readings could sometimes be more prominent based on the lexical content of the sentences. All modifications are nested rather than sequential: *The cat ate [the cookie [on the mat [beside the table]]]* rather than *The cat ate [the cookie [on the mat] [beside the table]]*.

Selectional preference: Predicates show selectional preference—a tendency to semantically constrain the arguments they take. For instance, verbs such as *sing,*

[7] This is a simplification, since not all synctactic categories or category subtypes are expected to follow a Zipfian frequency distribution (Piantadosi 2014).

walk are likely to take animate subjects. Our grammar only implements a simplified version of selectional preference: namely the animacy of the NP arguments based on verb type (e.g., subjects of unergatives are animate). In reality, selectional preference is much more complex and highly predicate-specific; for instance the theme of *eat* should be something that is edible. The simplification of selectional preference results in semantic infelicity in some of the generated sentences. This should not create any inherent difficulty in assigning a valid semantic representation if models are trained from scratch,[8] but may be adversarial to models with prior training on real language data.

4.4.2 Semantic Representation Language

The semantic representation language that we use is based on a Neo-Davidsonian view of verbal arguments (Parsons 1990), in which verbs specify an event,[9] and thematic roles link non-event dependents to the event. In event semantics, the sentence *John ate the cookie* would typically be associated with an eating event with *John* and *the cookie* as the event participants (11):

(11) John ate the cookie$\rightsquigarrow \exists e.eat(e, J, c)$

Associating predicates with events is a characteristic of the Davidsonian view (Davidson 1967). This view has an advantage of providing a more intuitive way to represent adverbial modification compared to alternatives such as classical predicate logic without events (12):

[8]There is also an argument that all syntactically well-formed sentences are meaningful (though it can be truth-valueless or fail denotationally), even if that sentence contains *category mistakes* like violations of selectional preference as discussed here: see Magidor (2009).

[9]Different views exist regarding whether all verbs introduce an event; for instance, Kratzer (1995) argues that individual-level predicates (i.e., predicates that typically express non-transient properties) such as being a dancer in *Manon is a dancer* do not specify events. Our dataset currently does not include such predicates, so the semantic representations of all examples in the dataset excluding the primitives contain events.

(12) a. John ate slowly $\rightsquigarrow \exists e.eat(J, c) \wedge slow(J)$ (unclear whether *John ate slowly* entails *John is slow*)

b. John ate slowly $\rightsquigarrow \exists e.eat(e, J, c) \wedge slow(e)$ (it is the eating event that is slow, rather than John)

The Neo-Davidsonian view is characterized by an additional assumption—namely that the roles played by the event participants (i.e., thematic roles) are predicate-general. For instance, representing event participants as separate conjuncts (13) (as opposed to via additional arity to the semantic representation of the predicate as in (11)) reflects the Neo-Davidsonian view. Such a representation carries the assumption that the roles *John* and *the cookie* play in the eating event (typically considered to be *agent* and *theme*, respectively) are not specific to the predicate *eat* only, but are generalizable to different predicates. To illustrate, it seems quite intuitive that there are shared properties between the ways in which *John* participates in the eating event in *John ate the cookie* and the ways in which *John* participates in petting event in *John pet the cat*, such as *John* being the initiator of the event.

(13) $\exists e.eat(e) \wedge agent(e, J) \wedge theme(e, c)$

Still, identifying a set of thematic roles that generalizes well across all predicates is a challenging problem (e.g., Dowty 1991). An alternative is assuming predicate-specific thematic roles without the generality assumption—for instance, *John* as the eater and *the cookie* as the eaten—but this view suffers from the opposite issue of missed generalizations. The representation language we use takes an intermediate view between these two options (14). This formulation carries a slightly weakened assumption about the generality of thematic roles by making the thematic roles predicate-specific. Still, this is closer to the Neo-Davidsonian view in that it makes use of predicate-general roles such as *agent* and *theme*.

(14) $\exists e.eat(e) \wedge eat.agent(e, J) \wedge eat.theme(e, c)$

In practice, the representation of thematic roles in (14) has the effect of letting the user decide which view to adopt through alternative tokenization. In our dataset, *predicate.role* (e.g., *eat.agent*) is by default treated as three separate tokens, '*predicate*', '.', and '*role*'. This treatment allows the models to capture the cross-predicate generality of *role* through the role token. On the other hand, the dataset can be easily modified so that *predicate.role* is treated as a single token. Under this alternative tokenization, it would be equivalent to assuming that thematic roles are predicate-specific, since every *predicate.role* would be unique. The tokenization scheme adopted by our dataset is discussed in more detail in Section 4.4.3.

In the context of NLP, several popular formalisms share similar assumptions regarding events and thematic roles as discussed above (e.g., semantic representation language used in PropBank (Palmer et al. 2005), Abstract Meaning Representation (Banarescu et al. 2013), Discourse Representation Theory (Kamp and Reyle 1993)-based representation language used in Groningen Meaning Bank (Bos et al. 2017)). We did not adopt these particular formalisms because they require manual translation that requires significant amount of annotator training.[10] Common alternative representation languages in NLP that do not share similar assumptions about events and thematic roles are query languages like SQL—see Section 4.6 for more discussions regarding the limitations of query languages for the questions asked in this work.

Detailed description of the semantic representation language: The semantic representation language that we adopt is a modification of the simplified logical formalism of Reddy et al. (2017),[11] which is a conjunctive, skolemized Neo-Davidsonian representation language. See Table 4-2 for a comparison between a traditional Neo-

[10] There are automatic methods but they are not guaranteed to be correct.

[11] https://github.com/sivareddyg/UDepLambda/blob/master/doc/SimplifiedLogicForm.md provides a description of this formalism.

Davidsonian representation, the simplified representation of Reddy et al. (2017) and ours.

Expression: *John ate the cookie.*	
Neo-Davidsonian	$\exists e.eat(e) \wedge agent(e, John) \wedge theme(e, \iota x.cookie(x))$
Reddy et al. (2017)	['arg0(3:e, 3:cookie)', 'eat.arg1(1:e, 0:m.John)', 'eat.arg2(1:e, 3:cookie)']
Ours	*cookie(x_3) ; eat.agent(x_1, John) AND eat.theme(x_1, x_3)

Table 4-2. Comparison of semantic representations corresponding to the expression *John ate the cookie.*

The changes we made to the original formalism are as follows:

- Skolem constants are named x_i instead of i, where i is the 0-based index of the head of the phrase represented by the constant.
- Our semantic representation conjoins each conjunct with AND, whereas the original representation is a list of conjuncts. When conjoining, we sort the conjuncts by the subscript of the Skolem constants so that the order of the conjuncts is deterministic.
- Event predicates associated with nominals are removed for simplicity.
- Definite and indefinite descriptions are formally distinguished (more details follow).

Primitive tranlsations: Primitives in our dataset take the following forms.

- Common noun: *shark* $\rightsquigarrow \lambda a.$shark($a$)
- Proper noun: *Emma* $\rightsquigarrow$ Emma
- Verb: *like* $\rightsquigarrow \lambda a.\lambda b.\lambda e.$like.agent($e, a$) AND like.theme($e, b$)

In the actual dataset, λ is written as LAMBDA.[12] Primitive meanings are not skolemized because they are not existentially quantified. We used the letters *e, a, b* for the bound variables so they do not overlap with x used for skolem constants (x_n). Verbs that are compatible with agents specify an agent as an argument in their primitive meanings for simplicity, rather than following the external argument analysis of Kratzer (1996).

Naming of Skolem constants: Both Reddy et al. (2017)'s and our representation languages use indexed skolem constants that express the existence of an entity or an event specified by the predicate. For example, in (15), x_1 expresses the existence of an entity that is both a cat and an agent of a smiling event; x_2 expresses the existence of an event that is a smiling event.

(15) A cat smiled $\rightsquigarrow$

$\text{cat}(x_1)$ AND $\text{smile.agent}(x_2, x_1)$

Index-based constant naming lets us avoid the need to select arbitrary constant names (e.g, x, y, z, ...) as the number of entities and events in the expression grows. The constants are named after indices of the phrasal head in the original sentence; in (15), the noun *cat* is in position 1, so the corresponding constant is x_1.

Definite descriptions: Definite descriptions that are not proper names are marked with an asterisk, standing in place of the standard ι notation. The asterisk expressions appear to the leftmost of the semantic representation to avoid nesting of predicated expressions. They are not conjoined with AND but separated with a ;, because ι expressions are of type e rather than t. Semantic representations containing asterisk expressions (e.g., The cat ran $\rightsquigarrow$ $*\text{cat}(x_1)$; $\text{run.agent}(x_2, x_1)$) should be equivalent to those containing nested ι expressions ($\exists e.\ \text{run.agent}(e, \iota x.\text{cat}(x))$, if ι is scopally inert. This may not necessarily be the case for definite descriptions in intensional semantics;

[12] I keep the λ notations in the text for readability.

for instance, under modals. See the discussion of Kaplan (1989) in Wolter (2019) for more details.

Mapping sentences to semantic representations: The semantic representation of a sentence follows deterministically from the PCFG rules, which directly encode disambiguating information for ambiguous syntactic structures (Section 4.3.5). For example (probabilities omitted):

$$\text{S} \rightarrow \text{NP}_{animate}\ \text{VP}_{unerg} \mid \text{VP}_{unacc}$$
$$\text{VP}_{unerg} \rightarrow \text{V}_{unerg}$$
$$\text{VP}_{unacc} \rightarrow \text{NP}_{inanimate}\ \text{V}_{unacc}$$

Sentences and their gold syntactic parses are first mapped to the simplified logical formalism of Reddy et al. (2017) using their codebase,[13] and then passed through several postprocessing steps laid out in the beginning of this section to obtain semantic representations in our modified formalism.

4.4.3 Structure of Dataset

Splits: We sampled 30,000 distinct sentences from our PCFG, excluding ones with duplicate nominals (e.g., *The **cat** saw a **cat***). These sentences were divided into training ($80\%; n = 24{,}000$), development ($10\%; n = 3000$), and test ($10\%; n = 3000$) sets. We then added to the training set examples that specify the primitive meanings of 80 verbs and 60 nouns (including common and proper nouns). Separately, we generated exposure examples ($n = 15$, details to follow in the next paragraph) to add to the training set. The resulting training set consists of 24,155 examples. The examples in the generalization set were sampled from separate PCFGs, each of which generates examples pertaining to a particular generalization case. For the Subject → Object generalization (Section 4.3.1), for example, we generated sentences with

[13] https://github.com/sivareddyg/udeplambda

hedgehog as a part of the object noun phrase. We sampled 1000 examples of each of the 21 cases (each row of Table 4-1), for a total of 21,000 examples.

Exposure examples: Many generalization cases crucially rely on particular examples being available in the training data. For instance, to apply the Subject → Object generalization to the noun *hedgehog*, at least one example containing *hedgehog* within the subject NP must be included in the training set. Human learners only need to observe an item in a small number of distinct contexts before they can generalize to new contexts. For example, children of age 2 years and 11 months were able to produce in a passive construction a nonce verb they have only heard in an active transitive construction, after being exposed to 8 distinct usages of the construction (Brooks et al. 1999). Borovsky et al. (2010, 2012) further suggest that human learners are even capable of single-shot learning of word meaning in context. We included in our training set a single example to generalize from ('exposure example') per generalization case that requires it.[14] We additionally created a version of the dataset that contains 100 exposure examples to test the effect of the number of distinct exposure examples.

Tokenization: A whitespace was placed between each word, and also before and after any special characters except when the special character appears in the initial position. Additionally, subscripts were written as _ followed by the index. Following this rule, the semantic representation of the example shown in Table 4-2 (repeated as (16-a)) would be (16-b) in the actual dataset. Note that depending on the tokenization policy adopted in individual experiments, the whitespace boundaries in (16-b) may not necessarily match the actual token boundaries as processed by the models.

(16) a. *cookie(x_3) ; eat.agent(x_1, John) AND eat.theme(x_1, x_3)

b. * cookie (x _ 3) ; eat . agent (x _ 1 , John) AND eat . theme (x _ 1

[14]The structural generalization cases do not have a restricted number of exposure examples. For instance, we did not impose an upper bound on how many depth 0-2 PP embeddings are sampled.

, x_3)

For readability, we maintain the notation in (16-a) in the remaining discussion.

4.5 Experiments

We next analyze the performance of two ANN models: Long Short-Term Memory (LSTM; Hochreiter and Schmidhuber 1997) and Transformer (Vaswani et al. 2017), both in an encoder-decoder setup (Sutskever et al. 2014). Transformers have been quickly adopted in practical NLP systems (Storks et al. 2019), but the literature has reported mixed results on the benefit of Transformers over LSTMs in terms of linguistic generalization (Hupkes et al. 2020; van Schijndel et al. 2019). The questions asked through these experiments are as follows: (1) whether these models are equipped with the compositional generalization abilities required by our test, and (2) whether there exist substantial differences across the models we test (i.e., different inductive biases), when the number of trainable parameters is controlled for.

4.5.1 Model and Training

We trained LSTM and Transformer models on our training set only without any training beforehand, and evaluated them on the generalization set. They were additionally evaluated on the test set, which contains distinct examples from the training set but is not systematically different (i.e., in-distribution; Section 4.4.3). The inputs (left hand side of (1)) were presented to the model as vectors corresponding to each token, without providing any explicit structure such as parse trees. We used cross-entropy loss, a batch size of 128, and early stopping when loss on the development set did not improve for five steps (step size = 500). All experiments were run five times with different random seeds, which determined the initial weights (sampled using Xavier initialization: Glorot and Bengio 2010) and the order of the training examples.

Models were implemented and trained using OpenNMT-py[15] (Klein et al. 2017). The input and output were both tokenized at whitespace boundaries as marked in the dataset (see Section 4.4.3 and Example (16-b)). The data and code to replicate the experiments in this section are publicly available.[16]

LSTM: We used a 2-layer LSTM encoder-decoder with global attention and a dot-product score function. The decoder followed an input-feeding approach (Luong et al. 2015). We tested both unidirectional (LSTM) and bidirectional (BiLSTM) encoders. We used inputs of dimension 512 and two hidden layers of dimension 512 (256 for model with bidirectional encoders so that the input dimension of the decoder stays constant across models after concatenating forward and backward states, and the number of parameters in the models remains comparable). A dropout of 0.1 was applied after the embedding layer and after each hidden layer except for the last. Following Lake and Baroni (2018), we used the Adam optimizer, and clipped gradients with a norm larger than 5.0. The training time for each model was around 3 to 4 hours on a single NVIDIA K80 GPU.

Transformer: Our Transformer model had 2 encoder and decoder layers, 4 attention heads, and a feedforward dimension of 512. Other hyperparameter settings not discussed here are the settings that replicate the performance[17] of Vaswani et al. (2017), according to OpenNMT-py documents.[18] The training time for each model was around 1 to 2 hours on a single NVIDIA K80 GPU. The Transformer had a comparable number of parameters to the LSTMs (Transformer: 9.5M; BiLSTM: 10M; LSTM: 11M).

[15] https://github.com/OpenNMT/OpenNMT-py

[16] https://github.com/najoungkim/COGS

[17] The Transformer model we used in our main experimental results is smaller, so they might not necessarily replicate the performance of the Vaswani et al. (2017). The model that we use in the additional experiment that examines the effect of model size might be comparable to Vaswani et al. (2017).

[18] https://opennmt.net/OpenNMT-py/FAQ.html

4.5.2 Results

Table 4-3 shows the generalization accuracy of the models tested, and additionally the development and test set accuracy. The poor generalization accuracy suggests that the models overall struggled with making compositional generalization (i.e., their generalization differed from target generalizations) in the presence of systematic gaps between training and evaluation. The near-perfect accuracy on development and test sets show that the models are able to assign correct semantic representations to novel sentences, but only if they are not distributionally distinct from sentences shown during training. Furthermore, the results on the generalization set were highly variable across identical model configurations with different random seeds, whereas the development/test set performance was stably high.

Model	Dev.	Test	Gen.
Transformer	0.96	0.96	**0.35** (± 0.06)
LSTM (Bi)	0.99	0.99	0.16 (± 0.08)
LSTM (Uni)	0.99	0.99	0.32 (± 0.06)

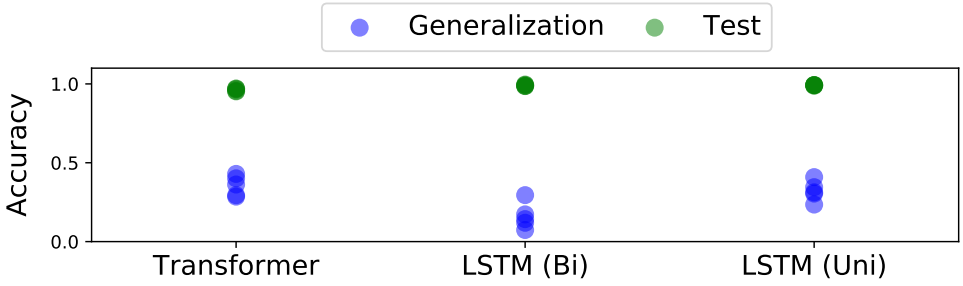


Table 4-3. Average accuracy of tested models. An output sequence is considered correct only if it exactly matches the target sequence. Only standard deviation greater than 0.01 is shown in the table (above). Each dot in the plot (below) represents a model trained with a different random seed. Five green dots are overlayed on top of each other in all models.

Accuracy by broad generalization type: Looking at the results divided by the broad category of generalization required (lexical or structural: Section 4.2), we can see that almost all of the successful generalizations are cases of lexical generalization

(Figure 4-1). For structural generalization, all models were almost completely unsuccessful, with unidirectional LSTMs being the most successful (though still very marginal) achieving around 1% structural generalization accuracy (3% on the shorter depth generalization subset: see Table 4-6).

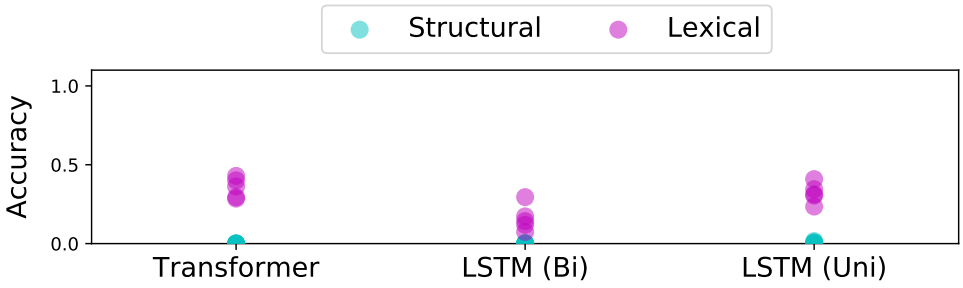


Figure 4-1. Accuracy by generalization type (lexical or structural). Five cyan dots are overlayed on top of each other in all models.

Accuracy by generalization case: Accuracy on each generalization case greatly fluctuated across different runs of the same model, except for the cases where accuracy was close to zero (see selected cases visualized in Table 4-4, and see Table 4-5 for full results by each generalization case). The only exception to the trend of high variance was the Active → Passive generalization (but not vice versa) in the Transformer model, where all runs of the model achieved close to 100% accuracy. The majority of the LSTMs' predictions were structurally correct even when they did not exactly match the expected output, suggesting that Active → Passive is one of the least challenging cases in our generalization set (further error analysis follows in Section 4.5.5.2).

4.5.3 General Error Patterns

One noticeable difference between LSTMs and Transformers was that, LSTMs' erroneous outputs were followed a more homogeneous pattern and were closer to the correct outputs in terms of token edit distance. The average token edit distance between errors and correct answers across all generalization cases, only considering error cases,

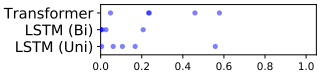
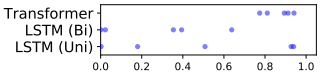
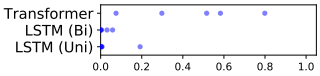
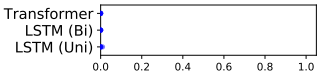
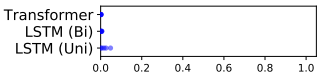
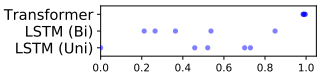
Case	Training	Generalization	Accuracy Distribution
Subject → Object (common noun)	*Subject* A **hedgehog** ate the cake.	*Object* The baby liked the **hedgehog**.	
Object → Subject (common noun)	*Object* Henry liked a **cockroach**.	*Subject* The **cockroach** ate the bat.	
Object → Subject (proper noun)	*Object* Mary saw **Charlie**.	*Subject* **Charlie** ate a donut.	
Primitive → Object (proper noun)	*Primitive* **Paula**	*Object* The child helped **Paula**.	
Depth generalization: PP modifiers	*Depth 2* Ava saw the ball **in the bottle on the table**.	*Depth 3* Ava saw the ball **in the bottle on the table on the floor**.	
Active → Passive	*Active* Emma **blessed** William.	*Passive* A child **was blessed**.	

Table 4-4. Accuracy by generalization case. Each dot represents a single run of the model.

were 11 and 14 tokens for bidirectional and unidirectional LSTMs, compared to 42 tokens for Transformers. Furthermore, Transformers frequently produced ill-formed logical forms; for example, they often failed to close the final parenthesis (17). In fact, ending the logical form with anything other than a right parenthesis is ill-formed (18). This type of error accounted for 12% of all Transformer errors, while only 0.5% of bidirectional and unidirectional LSTM errors were ill-formed in this simple way.

(17) Paula packed. ⇝

TARGET: pack.agent(x_1, Paula)

TRANSFORMER: pack.agent(x_1, Paula

(18) Emma appreciated the hedgehog. ⇝

TARGET: *hedgehog(x_3) ; appreciate.agent(x_1,Emma) AND appreciate.theme(x_1,x_3)

TRANSFORMER: *

4.5.4 Challenges with Structural Generalization

As briefly discussed in Section 4.4.3, some of the generalization cases require *lexical* generalization: a primitive needs to be translated in a structure which, while not itself novel, did not occur with that primitive in training. This is the case for Object → Subject generalization: the training set contains many examples of the structure [NP [V NP]$_{VP}$]$_S$ (Figure 4-2a), and the generalization concerns the particular NP that has never been observed in the first NP position. This contrasts with cases requiring *structural* generalization, where the structure of the sentence is itself novel. This is the case, for instance, for the structure [[NP PP]$_{NP}$ [V NP]$_{VP}$]$_S$—a PP modifier on the subject—which appears in the generalization set but not in training (Figure 4-2b). Note that structural novelty as I discuss here can only be determined latently, since the hierarchical phrasal structure is not directly provided as part of the input. For example, we can only determine that the sentences *A cat in its carriage purred* and *The rat danced on the table* (with the same number of words and share no lexical items) are structurally different based on the grammatical knowledge we have. Thus, in order to be able to generalize from training data, shared structural properties of the training and generalization examples (e.g., what parts in a given sentence correspond to a PP and how the phrase maps onto the semantic representation) must be deduced by the models.

The depth generalizations and the generalization of modifiers across grammatical roles require structural generalization; all such cases had zero or near-zero accuracies. This suggests that structural generalization is more challenging to the models we tested, compared to lexical generalization where the they achieved partial success. We investigate where the challenges in structural generalization may come from through analyzing the models' predictions.

Successful depth generalization cases: Depth generalization with PP modifiers

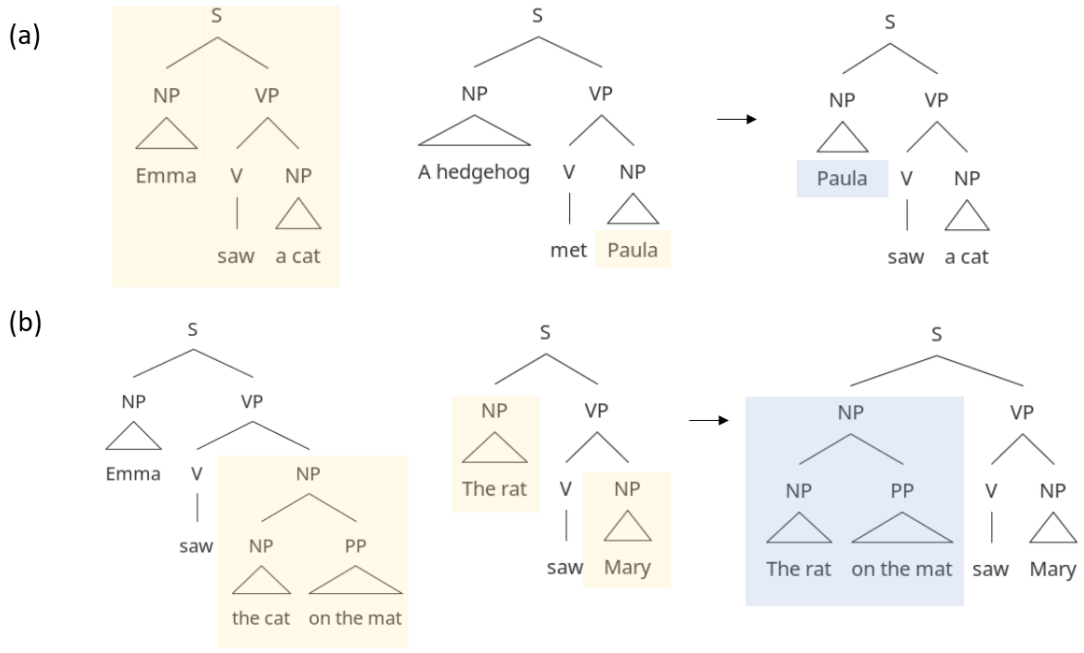


Figure 4-2. (a) Lexical generalization requires generalization to a novel combination of a familiar primitive and a familiar structure. (b) Structural generalization requires generalization to novel structures.

was the only case of structural generalization on which some models managed to achieve non-zero accuracy. All of the successful examples were cases of depth 3, the smallest unseen depth tested. The success cases also had shorter output lengths, with a maximum length of 120 tokens. This was within the range of output lengths seen during training (the longest training example included 153 tokens), which may account for the somewhat higher accuracy on the deep-but-shorter examples. The models we tested seem to be strongly bounded by the length of the observed examples—this aligns with the poor performance of LSTMs and Gated Recurrent Units reported in Lake and Baroni (2018) on a generalization task requiring translation of commands that are strictly longer than commands seen during training.[19]

[19]This length generalization task is very challenging; most proposed improvements on SCAN, a larger dataset that the length generalization is a part of, still perform poorly on this particular split (Furrer et al. 2020; Li et al. 2019; Lake 2019, *i.a.*). A notable exception is Chen et al. (2020b) which achieves perfect accuracy on the length split: I discuss the implication of the findings of this work in more detail in Section 4.6.

Failure to generalize structurally or failure to produce novel labels? It is known that contemporary ANNs find it challenging to produce labels they have not seen during training (Gandhi and Lake 2020). Handling this problem is a necessary part of solving depth generalization, since outputs of the depth generalization examples, such as (19-b) below, contain more constants than the training examples (19-a):

(19) a. Depth 1: The **cat liked** that the **dog saw** the **mouse**. *(5 index-taking items)*

b. Depth 3: The **cat liked** that the **dog liked** that the **mouse liked** that the **girl saw** the **rat**. *(9 index-taking items)*

As discussed in Section 4.4, we used index-based labels for constants precisely to help models with this issue of producing novel elements, by tying the labels to the indices. Specifically, the 5 index-taking items in (19-a) are labeled x_1, x_2, x_5, x_6 and x_8 instead of being assigned arbitrary labels such as $x, y, z \ldots$. However, even with such non-arbitrary labels, the model still needs to learn that a word at index i relates to the output string 'i'. This is an independent issue from producing a structurally well-formed semantic representation—the logical formula may be fully accurate except for the indexing of the constants. We try to tease apart this potential source of error through an error analysis.

Actual error patterns in structural generalization: While the problem of novel symbols is indeed an issue that the models need to handle during depth generalization, the pattern of errors suggest that the low accuracy is not purely due to this issue. In fact, only 0.5% of all depth generalization errors were cases where the structural form of the outputs were correct with only the indices being incorrect. More frequently, the models produced an end-of-sentence token too early (90.3% of all depth generalization errors), or produced sequences that were superfluously long (3% of errors contained more than 1000 tokens—more than twice as long as the maximum target output

length: 480). This implies that models struggle with handling longer and deeper sequences than those observed during training, independently of their inability to produce novel labels. While output length likely contributed to the difficulty of our depth generalization cases—even in the in-domain test set, the average length of correct answers was 43 tokens, compared to 83 for incorrect answers—deeply nested structures imposed additional challenges. On the long in-distribution test set examples (output length greater than 95), LSTM models and Transformer models still managed to achieve 68% and 13% accuracy, respectively. Their PP modifier depth generalization accuracy (which concern examples that are also longer in length) was much lower (LSTM: 2%; BiLSTM and Transformer: near 0%). Hence, depth imposes additional challenges to length.

Levels of embedding: Our depth generalization set contains examples with embedding depths 3–12. However, it is likely that humans would also find deeply embedded structures difficult to interpret (e.g., imagine reading a a CP embedding of depth 10!). Given this potential difficulty for humans, is our depth generalization a fair challenge to pose? Comprehensibility of 3–5 degrees of embeddings like the ones in our dataset is attested in the literature; Blaubergs and Braine (1974) showed that humans can understand 3–5 levels of right-branching CP embedding, and Karlsson (2010) observed that 3–5 levels of right-branching PP and CP embeddings do occur in corpora. In the case of the models we tested, they almost completely failed on generalization to any levels of embedding, including depths 3–5 that humans should be able easily understand (Table 4-6). I discuss the issue of generalization to depths greater than 5 in Section 4.7.3.

4.5.5 Errors in Lexical Generalization

4.5.5.1 Single Lexical Retrieval Error

A prominent pattern of error commonly observed in all models was what we name a *single lexical retrieval error*. Single lexical errors are errors in which the lexical part of the semantic representation (e.g., shark in (20-b)) associated with a single word in the *input* is incorrectly produced (20), while the output structure is overall correct.[20] Since the a single word in the input could be associated with multiple tokens in the output (e.g., a single word *shipped* in (21-a) corresponding to multiple occurrences of the token *ship* in (21-b)), single lexical retrieval errors include cases in which there are multiple output misproductions.

(20) a. INPUT: A shark burned Sophia .

b. TARGET OUTPUT: shark(x_1) AND burn.agent (x_2, x_1) AND burn.theme (x_2, Sophia)

c. SINGLE LEXICAL ERROR: director(x_1) AND burn.agent (x_2, x_1) AND burn.theme (x_2, Sophia)

(21) a. INPUT: The cat shipped Emma the donut.

b. TARGET OUTPUT: *cat(x_1) ; *donut (x_5) ; ship.agent(x_2, x_1) AND ship.recipient(x_2, Emma) AND ship.theme(x_2, x_5)

c. SINGLE LEXICAL ERROR (MULTIPLE MISPRODUCTIONS): *cat(x_1) ; *donut (x_5) ; admire.agent(x_2, x_1) AND admire.recipient(x_2, Emma) AND admire.theme(x_2, x_5)

d. SINGLE LEXICAL ERROR (SINGLE MISPRODUCTION): *cat(x_1) ; *donut

[20] Correct output structure means that the token identity and the position of the non-lexical tokens (parentheses, conjunctions, asterisks, periods, skolem constant prefixes x and _) were all identical to the target output, and there were no cross-category errors for lexical tokens. There are two categories of lexical tokens: (1) those that represent the lexical meanings like *balloon, see*, and (2) thematic role tokens like *agent, theme*). An output qualifies as structurally correct only if the lexical errors were within the category of (1) or within the category of (2). Indexing errors (e.g., x_3 instead of x_2) counts as a lexical error, so it is possible to have a structurally correct output with indexing errors.

(x_5) ; ship.agent(x_2, x_1) AND ship.recipient(x_2, Emma)

AND admire.theme(x_2, x_5)

All models we tested made this type of error very frequently: Transformer models made this type of error for 17% of all examples in the generalization dataset, LSTM models for 43% of the dataset, and BiLSTM models for 56% of the dataset. The frequency of these errors reflects the lack of sufficient bias that makes the models prefer outputs that are *faithful* to the given input. In other words, many errors were unfaithful to the input: there is nothing in the input that corresponds to the translation *director* in (20) and *admire* in (21)—the models are hallucinating these translations, likely due to frequently observing these tokens in the output during training. We follow up on this issue in Section 5.2.1.2.

4.5.5.2 Active → Passive: Error Patterns in LSTMs and Transformers

As mentioned in Section 4.5.2, the Active → Passive generalization was a case in which Transformers performed almost perfectly, whereas LSTMs did not. However, an error analysis reveals that the errors made by LSTMs were more homogeneous than those of Transformers. The majority of LSTM errors were structurally correct: only 0.3% (7/2591) of the unidirectional LSTM errors and 0.5% (14/2773) of the bidirectional LSTM errors had a different structure from the target output. LSTMs often replaced the target passive verb with a different one (22), misused a thematic role (23), or misused an index (24). These types of errors have equivalent structure to the correct output, and have the same number of tokens as the correct output.

(22) A balloon was blessed. ⇝

TARGET: balloon(x_1) AND bless.theme(x_3,x_1)

LSTM: balloon(x_1) AND inflate.theme(x_3,x_1)

(23) The book was blessed by a girl. ⇝

TARGET: *book(x_1) ; bless.theme(x_3,x_1) AND bless.agent(x_3,x_6) AND girl(x_6)

LSTM: *book(x_1) ; bless.theme(x_3,x_1) AND send.recipient(x_3,x_6) AND girl(x_6)

(24) A rose was blessed by the baby. ⇝

TARGET: *baby(x_6) ; rose(x_1) AND bless.theme(x_3,x_1) AND bless.agent(x_3,x_6)

LSTM: *baby(x_5) ; rose(x_1) AND bless.theme(x_3,x_1) AND bless.agent(x_3,x_6)

By contrast, the Transformer's errors in the Active → Passive generalization, despite being much fewer in number, mostly had incorrect structure (79.6% of all errors; 39/49. This is also numerically higher than the structural errors of LSTMs). The pattern in the total of 49 errors made by Transformer models in aggregate included omission of whole conjunct, spurious indices, not producing an output, using a numbered constant in place of a proper noun, and more. The following example shows a Transformer output with multiple errors—the model misinterpreted *tool* as a binary predicate and misindexed one of the constants:

(25) The tool was blessed by the girl. ⇝

TARGET: *tool(x_1) ; *girl(x_6) ; bless.theme(x_3,x_1) AND bless.agent(x_3,x_6)

TRANSFORMER: *tool(x_1) ; *girl(x_6) ; tool(x_3,x_1) AND bless.theme(x_3,x_6)

Some Transformer runs produced more homogeneous errors than others, despite having similar accuracy on the Active → Passive generalization. For example, some runs mostly made the error of using the wrong verb as in (22), whereas others made more idiosyncratic errors with mixed patterns like (25).

One possible reason for the high performance on the Active → Passive case is that our training data included both passive constructions with and without the agent

by-phrase (e.g., both *The book was seen* and *The book was seen by Emma*). In these two constructions, the semantic representation of the former is a prefix of the semantic representation of the latter:

(26) The book was seen (by Emma). ⇝

NO BY: ***book**(x_1) **and see.theme**(x_3,x_1)

WITH BY: ***book**(x_1) **and see.theme**(x_3,x_1) AND see.agent(x_3,Emma)

Since these two types of passive constructions were sampled with equal probability, performance on the Active → Passive case may have benefited from more exposures to examples relevant to forming the passive construction. A useful follow-up experiment would address whether this effect is absolute (more exposure to relevant examples always facilitates generalization) or relative to the number of other examples relevant to other generalizations (the number has to be greater than examples that contribute information about other generalization cases). If the latter hypothesis is correct, it would signal the intrinsic difficulty of the models learning *all* generalizations in the dataset.

4.5.5.3 Common vs. Proper Nouns

Table 4-4 shows that even for the same type of targeted generalization (e.g., Object → Subject, Primitive → Object), the variant that used proper nouns (27) was more challenging than the variant using common nouns (28).

(27) TRAINING: A creature grew **Charlie**. ⇝

creature(x_1) AND grow.agent(x_2, x_1) AND grow.theme(x_2, **Charlie**)

GENERALIZATION: **Charlie** ate a cookie. ⇝

eat.agent(x_1,**Charlie**) AND eat.theme(x_1,x_3) AND cookie(x_3)

(28) TRAINING: Henry liked **a cockroach**. ⇝

like.agent(x_1, Henry) AND like.theme(x_1,x_3) AND **cockroach(x_3)**

GENERALIZATION: **A cockroach** ate a bat. $\rightsquigarrow$

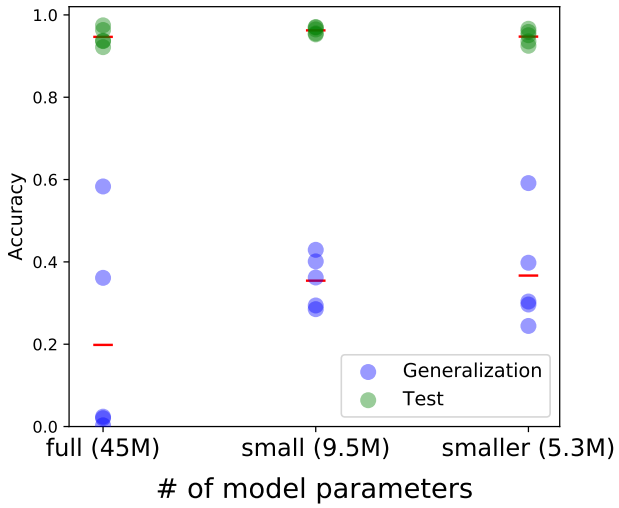
cockroach(x_1) AND eat.agent(x_2,x_1) AND eat.theme(x_2,x_4) AND bat(x_4)

What is the source of this discrepancy? As can be seen from the above examples, common and proper nouns are formally distinct in both the sentence and the semantic representation. Translating a common noun requires conjoining a unary predicate (cockroach(x_n)), and placing the predicated constant (x_n) in appropriate event predicates. On the other hand, translating a proper noun requires placing a constant (Charlie) inside appropriate event predicates. Given the lower complexity of (symbolic) steps required for translating proper nouns, the lower accuracy is surprising. While we do not have a definite explanation for this discrepancy, one possibility is that it is due to a frequency effect; our dataset overall contained more common nouns than proper nouns, in terms of both type and token frequency.

The discrepancy in accuracy between common and proper nouns indicates that performance is sensitive to seemingly minor formal differences in cases that require the same type of generalization, echoing the discrepancy between the *jump* and *turn left* primitive splits of SCAN that were originally observed by Lake and Baroni (2018).

4.5.6 Effect of Model Size

The results we reported are from models that have around 10M learnable parameters. How does the number of parameters affect the models' success in generalization, given that it is a general trend for model performance to scale with size? We conducted a follow-up experiment varying the number of learnable parameters in the Transformer model. Figure 4-3 compares the performance of three Transformer models of varying size (large: 45M; small (main result repeated): 9.5M; smaller: 4.5M). The number of parameters did not a have large impact on test set accuracy; all runs of all models



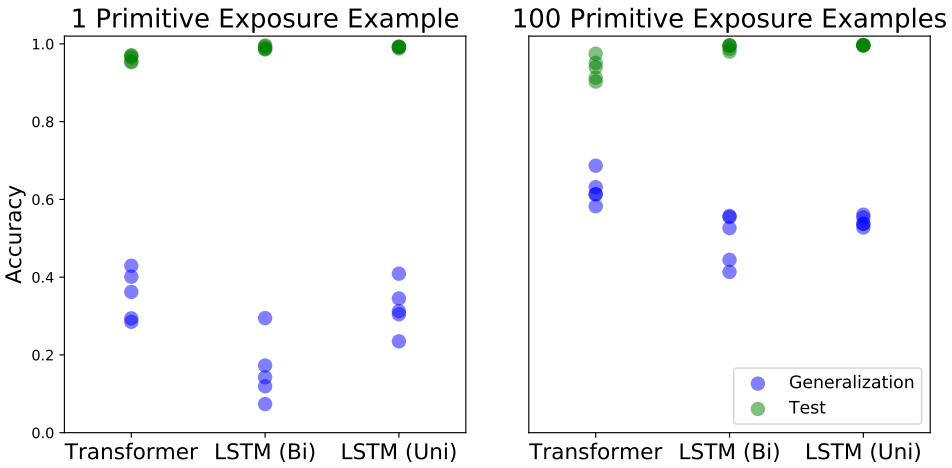
# Params.	Dev.	Test	Gen.
45M	0.95	0.95	0.20 (± 0.26)
9.5M	0.96	0.96	0.35 (± 0.06)
5.3M	0.95	0.95	0.37 (± 0.14)

Figure 4-3. The effect of Transformer model size on generalization and test set accuracy.

consistently achieved higher than 90% accuracy. On the other hand, model size did affect generalization. Perhaps surprisingly, the average across 5 runs of the large model that had an almost five-fold increase in size was lower than those of smaller models; however, this average result is hard to interpret given the very high variance in accuracy across runs of the the largest Transformer.

4.5.7 Effect of Number of Distinct Exposure Examples per Primitive

As discussed in Section 4.4.3, our training dataset includes a single exposure example for each primitive. To test whether a larger number of exposure examples help generalization (i.e., observing the primitives in various distinct contexts, while still maintaining the systematic gap for the evaluation), we repeated our experiments with a version of the training dataset in which the number of exposure examples was increased to 100. All models benefited from the greater number of exposure examples



Model	# Exposure examples	Dev.	Test	Gen.
Transformer	1	0.96	0.96	**0.35**
	100	0.94	0.94	**0.63**
LSTM (Bi)	1	0.99	0.99	0.16
	100	0.99	0.99	0.50
LSTM (Uni)	1	0.99	0.99	0.32
	100	1.00	1.00	0.54

Figure 4-4. Accuracy with a different number of exposure examples. Left figure ("1 Primitive Exposure Example") is a repeat of the main results in Figure 4-3.

(Figure 4-4). Note that the structural generalization cases, such as Object-Modifying PP → Subject-Modifying PP, did not require primitive exposure examples, and are therefore identical across the 1-shot and 100-shot settings (for a detailed breakdown by case, see Table 4-5). Thus, it is expected that only the lexical generalizations benefit from this increase in exposure examples.

4.6 Related Work

Semantic parsing: Our dataset takes the form of semantic parsing (mapping natural language expressions onto their corresponding semantic representations). Many existing semantic parsing datasets adopt query languages as the semantic representation

language (SQL: Yu et al. 2018; Zhong et al. 2017; Finegan-Dollak et al. 2018, SPARQL: Keysers et al. 2020, *i.a.*). Datasets adopting query language as the representation language often envision clear downstream applications such as question-answering and flight reservation—the queries in the dataset are means to retrieving information from a database.

On the other hand, more general-purpose semantic representation language (i.e., not specific to database queries) have also been adopted in semantic parsing datasets. These include lambda calculus-based representation languages (Hemphill et al. 1990; Dahl et al. 1994; Zettlemoyer and Collins 2005; Reddy et al. 2016, 2017), languages using predicate logic (with some higher-order augmentations) (Zelle and Mooney 1996), Abstract Meaning Representation (AMR) (Banarescu et al. 2013) (influenced by the semantic representation language used in PropBank (Palmer et al. 2005)), and semantic dependency graphs (e.g., Flickinger 2000; Miyao 2006; Hajič et al. 2012[21]).

The representation language we adopt is closer to the latter tradition. In particular, we adopted the lambda calculus-based semantic representation language of Reddy et al. (2016) and Reddy et al. (2017)—the relative ease of mapping between syntax and semantics was behind this choice (compared to alternatives presented here; for instance, mapping natural language expressions to PropBank-style semantic representation language requires expert annotators). One alternative approach to generating data is to select one of the existing resources and subsample from the corpus. We did not take this approach because some of the target phenomena involved constructions that go beyond what is expected to be attested in corpora (e.g., embedding depths greater than 5).

Motivation: Our work shares motivation with Finegan-Dollak et al. (2018), who propose an evaluation setup that uses distributionally distinct training/test sets to

[21]See Oepen et al. (2015) for a discussion of the differences between these different graph representations.

test generalization in semantic parsing datasets. The most similar in motivation to our work seems to be Compositional Freebase Questions (CFQ; Keysers et al. 2020), a synthetic dataset designed to test for compositional generalization in SQL (specifically, SPARQL) parsing. Despite this shared motivation, CFQ and our test diverge in many ways. First, because CFQ adopts a query language as the representation language, their dataset mainly consists of questions (e.g., *Who directed the movie Blue Velvet?*) and imperatives (*Find the director of the movie Blue Velvet*). Second, in CFQ, the generalization examples are selected from queries with a similar primitive distribution but different distributions of the composed forms ("compound divergence"). This can lead to a training/generalization split that is not characterized by any principled linguistic difference (in the authors' words: the difference between training and generalization sets is not *easily observable "with the naked eye"* (p. 18)). Here are some examples from their dataset:

Training

- What was founded by a costume designer, founded by Forgotten Silver's star, and founded by Jamie Selkirk?
- Which male person influenced and was influenced by William Dean Howells?

Generalization

- What sequel of Paranormal Activity 2 was edited by and written by a film director?
- Was Luke Larkin Music an art director's employer?

In contrast, the two broad criteria in our test that distinguishes training/generalization are: (1) generalization examples that contain novel combinations (observed zero times in training) of known primitives and known syntactic configurations, and (2)

generalization examples with novel complex structures, the parts of which occur in the training set. This means that the distribution of primitives across training and generalization systematically vary, unlike CFQ where the distribution of primitives are kept as close as possible across training/generalization sets. Although both tests are under the banner of compositional generalization, it is not entirely clear whether the two tests are evaluating similar capacities.

More broadly, our work is connected to work that tests ANN models for syntactic generalization (Gulordava et al. 2018; Marvin and Linzen 2018, *i.a.*), but our test concerns assignment of semantic representation rather than production. This line of work tests for syntactic generalization by comparing language modeling probabilities (i.e., production probabilities) given preceding context, similarly to cloze tasks.

Findings: Our aggregate results in Table 4-3 are in line with recent work that has documented a significant discrepancy between neural models' excellent performance within distribution and their degraded performance out of distribution (Johnson et al. 2017; Lake and Baroni 2018; Hupkes et al. 2020). In particular, our finding of poor generalization to deeper nested structures aligns with the results of Hupkes et al. (2020).

Given that deeper structures also tend to be longer than shallower ones, this finding also relates to the difficulty of generalization to longer sequences. One illustrative example is the poor performance of LSTMs on a SCAN split that requires generalizing from shorter to longer sequences. While several models have made significant improvements over other SCAN splits, progress on the length split remains minimal (Li et al. 2019; Lake 2019; Gordon et al. 2020). A notable exception is Chen et al. (2020b), where full generalization accuracy is achieved on the length split. The core components of their model are a neural stack machine that provides an explicit mechanism of symbolic manipulation (most importantly, supporting recursion) and a latent category

predictor. Both of these components align with the intuition regarding mechanisms that would help achieve compositional generalization in our task (i.e., recursion and category abstraction),[22] suggesting a promising future direction.

4.7 Limitations and Future Work

4.7.1 Constraints on Generalization

To reach full adult linguistic competence, human learners not only need to be able to make generalizations, but also need to learn how to constrain them. For example, the verb *donate* takes a recipient *to*-PP (*Emma donated the book to the museum*) but does not allow double-object alternation (**Emma donated the museum the book*). How constraints as such could be learned has been discussed in linguistics under the banner of the projection problem (Baker 1979). We focused on evaluating computational models' ability to make compositional generalizations, but not on evaluating the ability to constrain them. For this reason, we only included examples to which generalizations are applicable (e.g., dative verbs that alternate). This is a simplification; in natural language, generalizations are not applicable across-the-board, and are modulated by a multitude of morphophonological, syntactic and semantic factors. In the case of the dative alternation, properties such as animacy and definiteness are involved (Bresnan and Ford 2010). Thus, evaluating constraints on generalization requires a detailed characterization of factors that govern individual generalization cases, as well as a formalism capable of expressing these factors, which we leave to future work.

4.7.2 Generalization of PP Modification

Effect of PP fragments: The PP modifier generalization set (Section 4.3.2) requires generalizing PPs that modify NPs in the object position to NPs in the subject position,

[22]However, whether models that explicitly introduces/mimics symbolic computation would be useful in the context of more cognitive questions posed in Section 2.4 is unclear.

without having seen any subject modification. We note that this may be a stronger generalization problem than what humans may actually encounter based on the following two observations. First, it is true that PP modifiers in the subject position are much less frequent than PP modifiers in the object position in child-directed speech, but subject-modifying PPs are not absent from it: according to our analysis of the Epochs corpus of Perfors et al. (2011), PP modification on the subject of a declarative sentence occurred only 13 times whereas PP modification on the object occurred over 100 times. Second, there exist many [NP PP] fragments that are not full sentences (e.g., *a disk from a game*) in the corpus. It is still likely that PP modification does not occur in all possible syntactic positions that can be occupied by an NP—for instance, in the subject position of a depth 2 embedded CP—and to interpret such sentences structural generalization would be required. Nevertheless, whether humans would be able to generalize modifiers in one syntactic position in the total absence of observing modifiers in other syntactic positions (or as fragments) remains to be tested, and is part of our future work.

Effect of exposure to multiple modification contexts: Perfors et al. (2011) point to a possibility that exposure to modification in multiple syntactic contexts may be key to learning a more general PP modification rule rather than ones specific to certain grammatical roles. Hence, rather than testing generalization from object PP modification to subject PP modification, exposing models to both and testing generalization to modification in subjects and objects of embedded CPs may yield better generalization results. Increasing the diversity of syntactic positions may have an analogous effect to how increasing the number of exposure examples helped lexical generalization (Section 4.5.7).

4.7.3 Comparing Depth Generalization with Human Learners

Our depth generalization sets test generalization to 3–12 degrees of embedding in right-branching structures. However, human processing of embedded structures degrades over levels of embedding (Blaubergs and Braine 1974) and attestation of embeddings greater than depth 5 is rare (Karlsson 2010). Given this limitation in human processing, should the inability to handle generalization to our generalization set, and furthermore arbitrary depths of embedding be viewed as a limitation of the system? This question, which we leave to future work, calls for an analysis of the memory constraints that the models are subject to. If a model is not restricted by the same memory limitations as humans, they should not fail to process such sentences. In such a model, any such failure would be diagnostic of a discrepancy between what the model has learned and the correct way to perform the task, as defined by our grammar and the mapping rules. Combined with the analysis of memory constraints, an analysis of the effect of memory on ANN models' performance and generalization patterns on the depth subset would be an interesting follow-up investigation.

4.7.4 Semantic Representation Language

The semantic representation language that we adopt only accounts for a fragment of English. For instance, it is purely extensional and does not capture important scope-taking meanings such as universal quantification and negation. Incorporating these properties poses a substantial challenge to the evaluation setup. In the case of incorporating scope, expressions containing multiple scope-taking elements often have multiple valid mappings (e.g., *Every cat ate a cookie* has two interpretations: (1) every cat each ate one cookie of their own and (2) every cat ate a single cookie) due to scope ambiguity. How to formulate the mapping task for such sentences, including whether it would be reconcilable with the current evaluation setup, requires further

thought.

Moreover, an inevitable limitation of our experimental setup (and others that adopt any sort of representation language) is that the choice of the representation language itself may impact the results independently from the phenomena being tested. One way to move forward is simultaneously testing different formalisms to tease apart the effect of particular formal assumptions, as has been done in Guo et al. (2020) and Yanaka et al. (2021).

4.8 Conclusion

We have proposed an evaluation method to test the capacity for compositional linguistic generalization that uses a synthetic sentence-to-semantic representation mapping task that captures a fragment of English. The performance on our task is designed to represent the degree to which the models' generalization patterns match the expected patterns of generalization of human learners based on the literature. When evaluated using this test, the Transformer and LSTM models that we tested performed poorly on the generalization set, with high variability across difference random restarts of the network, despite their performance on the in-domain test set being consistently near-perfect. In particular, all models we tested failed almost completely on structural generalization. Among the models tested, unidirectional LSTMs seemed to provide the most helpful inductive bias for structural generalization in comparison to others—they were the most successful (though, still a very marginal success) with around 1% generalization accuracy and 3% on the shorter depth generalization subset. Overall, LSTMs also produced errors that were closer to the target, with token edit distance of 11 (bidirectional) and 14 (unidirectional), which is much lower than the edit distance of 42 for Transformer models.

Both Transformer and LSTM models tested achieved partial success on lexical

generalization. When supplemented with a more diverse set of contexts that the target lexical items appeared in, the models' lexical generalization improved further. Nevertheless, all models' generalization patterns substantially deviated from the target patterns, calling for further investigation into the inductive bias of the models. The next chapter examines this question, exploring ways to modify the models' inductive bias.

# Exposure Contexts	Case	Transformer	LSTM (Bi)	LSTM (Uni)
1	Subject → Object (common noun)	0.31	0.05	0.18
	Subject → Object (proper noun)	0.30	0.00	0.06
	Object → Subject (common noun)	0.87	0.28	0.51
	Object → Subject (proper noun)	0.45	0.02	0.04
	Primitive noun → Subject (common noun)	0.17	0.02	0.03
	Primitive noun → Subject (proper noun)	0.00	0.00	0.17
	Primitive noun → Object (common noun)	0.06	0.05	0.01
	Primitive noun → Object (proper noun)	0.00	0.00	0.00
	Primitive verb → Infinitival argument	0.00	0.23	0.07
	Object-modifying PP → Subject-modifying PP	0.00	0.00	0.00
	Depth generalization: Sentential complements	0.00	0.00	0.00
	Depth generalization: PP modifiers	0.00	0.00	0.02
	Active → Passive	0.99	0.45	0.48
	Passive → Active	0.61	0.19	0.49
	Object-omitted transitive → Transitive	0.61	0.05	0.60
	Unaccusative → Transitive	0.38	0.03	0.26
	Double object dative → PP dative	0.45	0.16	0.75
	PP dative → Double object dative	0.58	0.07	0.79
	Agent NP → Unaccusative Subject	0.69	0.31	0.56
	Theme NP → Object-omitted transitive Subject	0.45	0.74	0.87
	Theme NP → Unergative subject	0.50	0.74	0.87
100	Subject → Object (common noun)	0.86	0.93	0.91
	Subject → Object NP (proper noun)	0.54	0.60	0.54
	Object → Subject (common noun)	0.86	0.98	0.97
	Object → Subject (proper noun)	0.81	0.30	0.32
	Primitive noun → Subject (common noun)	0.83	0.00	0.00
	Primitive noun → Subject (proper noun)	0.24	0.00	0.00
	Primitive noun → Object (common noun)	0.82	0.05	0.01
	Primitive noun → Object (proper noun)	0.23	0.00	0.00
	Primitive verb → Infinitival argument	0.89	0.18	0.21
	Object-modifying PP → Subject-modifying PP	0.00	0.00	0.00
	Depth generalization: Sentential complements	0.00	0.00	0.00
	Depth generalization: PP modifiers	0.00	0.01	0.02
	Active → Passive	0.99	1.00	1.00
	Passive → Active	0.89	0.45	0.79
	Object-omitted transitive → Transitive	0.73	0.63	0.98
	Unaccusative → Transitive	0.47	0.75	0.94
	Double object dative → PP dative	0.83	0.85	0.99
	PP dative → Double object dative	0.82	0.94	0.96
	Agent NP → Unaccusative Subject	0.84	0.99	0.99
	Theme NP → Object-omitted transitive Subject	0.53	0.86	0.81
	Theme NP → Unergative subject	0.96	0.96	0.98

Table 4-5. Full model accuracy by generalization case, with primitive exposure in 1 context (default) and 100 (increased) distinct contexts. Each result is an average over 5 random seeds.

Model	All	3–5	6–12
Transformer	0.00	0.00	0.00
LSTM (Bi)	0.00	0.01	0.00
LSTM (Uni)	0.01	0.03	0.00

Table 4-6. Accuracy on depths 3–5 and depths 6–12.

Chapter 5

Investigating Factors that Contribute Helpful Learning Biases for Compositional Generalization

5.1 Motivation

In Chapter 4, we saw that the Artificial Neural Network (ANN) models we tested, despite their impressive downstream success, did not display compositional generalization patterns that match the desired patterns. This implies that the inductive bias of humans and ANN models must be different. This motivates the question addressed in this chapter: how can we change the inductive bias of the ANNs tested in Chapter 4 to more closely match that of humans? Identifying factors that lead to such a change would help us characterize better the inductive bias that gives rise to compositional generalization in an intelligent system.

One way to influence the learning bias of ANNs is using an auxiliary objective—an objective that is different from the objective for the task that the model is targeting to solve, but is hypothesized to be helpful in solving the target task. This approach has been successfully applied to various domains of machine learning (Zhang et al. 2014; He et al. 2017; Jaderberg et al. 2017; James et al. 2017; Meyerson and Miikkulainen 2018; Zhang et al. 2020, *i.a.*). In Natural Language Processing (NLP), the currently

dominant paradigm of pretraining-then-finetuning[1] can also be seen as an auxiliary training-based approach: a model is first trained (*pretrained*) on an auxiliary objective before it is trained on the target objective (*finetuned*). While various formulations of language modeling (word prediction in context; see (1)) are the most commonly used pretraining objectives, recent ANN analysis works have shown that different auxiliary objectives (e.g., tagging, translation, sentence prediction...) lead to measurable differences in the syntactic and semantic phenomena the models are sensitive to (Zhang and Bowman 2018; Kim et al. 2019a; Wang et al. 2019a; Pruksachatkun et al. 2020). This suggests that the auxiliary objectives can sufficiently modify the inductive bias of the models to a degree that the change in generalization behavior is reflected in the target task performance.[2]

(1) a. INPUT: How can I make my cat understand human ___

b. OUTPUT: language

The first experiment in this chapter investigates the effect of three different auxiliary objectives: CCG supertagging, glossing, and denoising (a sequence-to-sequence variant of the language modeling objective; compare (2) with (1)), the details and motivations of which are discussed in Section 5.2.1. We add an auxiliary training objective to the ANN models we used in Chapter 4 (model descriptions in Section 4.5.1), before the models are trained on the compositional generalization task.

(2) a. INPUT: How can I $\langle x \rangle$ my cat $\langle y \rangle$ human $\langle z \rangle$?

b. OUTPUT: $\langle x \rangle$ make $\langle y \rangle$ understand $\langle z \rangle$ language

[1] Footnote from Chapter 3 repeated here for readers who are unfamiliar with NLP: *Pretraining* in the narrow sense refers to training a model (typically with an objective of language modeling, the objective of predicting a word in context) using a large amount of data to obtain a 'general purpose' model of language. This model is then *finetuned* to perform a specific target task (e.g., question-answering, natural language inference, machine translation...). Pretraining in the broader sense can also refer to any kind of training before the model is trained on the target task.

[2] I say *suggests* here because most of the works mentioned here measure in-distribution generalization performance.

In the second experiment, we focus on the denoising (language modeling) objective, investigating the effect of the auxiliary training on this objective at scale. To allow investigation at a much larger scale than our computational resources permit, we change the ANN models tested to T5 (Raffel et al. 2020), which is a Transformer-based sequence-to-sequence model pretrained on 34 billion tokens of English data crawled from the web using the denoising objective. This experiment addresses two main questions: first, we ask whether an ANN setup with more empirical success[3] in contemporary NLP—a larger model (compared to the ones used in Chapter 4) pretrained on a large amount of data on the language modeling objective—improves compositional generalization. Second, we investigate the effect that the amount of exposure to language data has on compositional generalization by systematically varying the size of pretraining data.

5.2 Experiment 1: Auxiliary Training Objectives

5.2.1 Auxiliary Training Objectives Compared

We compare the effects of three different auxiliary training objectives on compositional generalization: CCG supertagging, glossing and denoising. The motivation behind each objective and the formats of the tasks are discussed below.

5.2.1.1 CCG Supertagging for Category Cues and Structural Constraints

As we saw in Section 4.5.2, the ANN models we tested almost completely failed to demonstrate structural generalization. The structural generalization cases require an ability to recognize constituent structure—for instance, interpreting a PP modifier in a novel syntactic position requires the models to recognize what parts of the sentence comprises a PP. If so, an auxiliary objective that helps models identify linguistic

[3] In addition to being more effective on benchmark tasks, converging evidence from analysis work suggests that language modeling is also the most helpful objective in terms of endowing models with linguistic knowledge (Zhang and Bowman 2018; Kim et al. 2019a; Wang et al. 2019a).

structures could contribute to structural generalization. One such objective is CCG supertagging (Bangalore and Joshi 1999), which is a task of assigning tags to words. CCG supertagging differs from plain part-of-speech tagging in that the tags are indicative of how adjacent constituents combine with each other. Thus, provided a series of supertags for a given (well-formed) expression, the constituency structure of the expression can be inferred.

CCG supertagging can easily be formulated as a sequence-to-sequence task that is compatible with the models that we have been using. The input would be a sequence of words and the output, a sequence of supertags corresponding to each word in the input, as in (3):

(3) a. INPUT: it settled with a loss of 4.95 cents at $ 1.3210 a pound .

b. OUTPUT: NP (S[dcl]\NP/PP) ((S\NP)\(S\NP))/NP NP[nb]/N N (NP\NP)/NP N/N N PP/NP N/N[num] N[num] (NP\NP)/N N .
(*Example re-cited from Cui and Zhang 2019*)

We expect the CCG supertagging objective to inform the models about the syntactic categories in English, which is a precondition to determining the applicability of compositional rules (Chapter 3). We also hypothesize that the models will be provided with signals regarding the syntactic constraints in the language—that is, what words can combine with each other to form valid constituents—which could restrict the generalization space. Kim et al. (2019a) present a promising result in this regard; they have shown that using CCG supertagging as an auxiliary objective improves the performance of LSTM-based models on the target task of end-of-sentence detection, a task that requires sensitivity to valid constituents.

5.2.1.2 Glossing (Word-by-word Translation) to Promote Faithful Translations

A prominent error pattern that was shared across all models tested in Chapter 4 was a single lexical retrieval error, where in place of the tokens that had a systematically controlled distribution in training, the models produced the wrong token, even though everything else in their predicted output was correct (Section 4.5.2). The wrong output tokens produced were either the tokens corresponding to other lexical items in the input sentence (4-a), or tokens corresponding to other lexical items that were in the training set but were not part of the input (4-b).[4]

(4) a. Emma drew Lina $\rightsquigarrow$ draw′.agent(x_1, Emma′) AND draw′.theme(x_1, Emma′)

b. Emma drew Lina $\rightsquigarrow$ draw′.agent(x_1, Emma′) AND draw′.theme(x_1, Audrey′)

Either error pattern can be described[5] as a violation of *faithfulness constraints* in the terms of Optimality Theory (Prince and Smolensky 2002)/Harmonic Grammar (Legendre et al. 1990). The specific faithfulness constraints that are violated in (4) are best described by the constraints proposed in Correspondence Theory (McCarthy and Prince 1995) which assume a correspondence relation between input and output representations. In the case of (4-a), there is no output token that corresponds to the input token *Lina*—this can be described as a violation of the MAX constraint, which states that every element in the input must have a corresponding element in the output (i.e., anti-deletion). In the case of (4-b), the output again violates the MAX

[4]The actual logical formulae in the generalization dataset do not contain the marker ′ that we use here to distinguish input and output tokens. We added these here to highlight the fact that they are not the same token—the models used in Chapter 4 have separate input and output vocabulary spaces, so they are distinctly represented. This can be thought of as the difference between object language and representation language (e.g., in the object language-to-representation language mapping in John smiled $\rightsquigarrow$ smile(John), the left hand side and the right hand side are different languages (left: English, right: logical representation language), and thus have separate vocabulary spaces).

[5]We emphasize that the Optimality Theory/Harmonic Grammar constraints are invoked here for descriptive purposes—we are not making a claim that the ANN models are actually implementing weighted constraints. We use the constraints to describe the behavior of the models rather than as a formal characterization of the mechanics of the models.

constraint; there is no corresponding token of input *Lina* in the output. However, there is an additional faithfulness constraint violation this time. The constraint DEP, which states that every element in the output must have a corresponding element in the input (i.e., anti-insertion), is also violated, because *Audrey′* in the output does not have a corresponding input element.

The two common error patterns in (4) can be interpreted as a result of insufficient weighting of faithfulness constraints: the models are often outputting translations that violate MAX and DEP constraints, likely in the presence of other constraints that the outputs are satisfying. While we are not certain what exact constraints the outputs are satisfying, *Emma′* and *Audrey′* playing the role of theme in the erroneous translation (4) seems to be a result of preferring translations that have been observed during training. That is, while there are examples in the training data in which *Emma′* and *Audrey′* play the theme role, there are no examples in which *Lina′* plays the theme role. The errors made by the models seem to reflect a preference for observed pattern of semantic roles over translations that are faithful—this preference could roughly be expressed as a BE PROBABLE! constraint. In light of this description of the output patterns, what can we do to make the models prefer more faithful translations?

We propose that adding a *glossing*[6] task (or a word-by-word translation task) as an auxiliary objective can help achieve exactly this. In the glossing task, the input is a sequence of symbols, and the output is also a sequence of symbols, each of which corresponds to a single token in the input sequence (5). There is a one-to-one mapping between the corresponding input and output tokens: an input token a corresponds to an output token a' and nothing else, and the output tokens a' corresponds to the input tokens a and nothing else.

(5) a. INPUT: The cat danced

[6]We name this task 'glossing' after the convention in the linguistic literature of providing word-by-word glosses for examples in languages different from the language of the main text.

b. OUTPUT: The′ cat′ danced′

This is a task in which the input-output pairs contain no violation of MAX or DEP constraints; there are no input elements without a corresponding output element (no deletion), and there are no output elements without a corresponding input element (no insertion). By training the models on this task, we hope to provide the models with a signal that outputs that are faithful are more preferable.[7]

5.2.1.3 Word Prediction in Context (i.e., *Language Modeling*)

Finally, we consider language modeling as an auxiliary objective, which is an objective of predicting a word in context.[8] While there is no clear established connection between the language modeling objective and compositionality (or systematicity) in the literature, Elman (1991) has suggested that the task of predicting the next word can help networks discover linguistic regularities. Specifically, he has noted a possibility that this kind of word prediction guides language acquisition for humans by providing a partial solution to Baker's paradox (or more generally, the difficulty of learning language in the absence of negative evidence). That is, by implicitly having anticipations about the linguistic inputs they encounter, children can use the failed predictions as negative signal during learning.

Recent empirical results (Furrer et al. 2020) on the SCAN task (Lake and Baroni 2018) using the T5 model of Raffel et al. (2020) are also in support of the hypothesis that language modeling as an auxiliary objective helps compositional generalization (although see Section 5.2.3 for a methodological problem in their approach). In the authors' words (Furrer et al. 2020, p. 5):

[7]Note that this glossing task also satisfies other faithfulness constraints such as LINEARITY (the linear ordering of the input should be preserved in the output; 'no reordering') or INTEGRITY (each input element corresponds to not more than one output element), that might not necessarily be helpful for our compositional generalization task.

[8]The model tested for category abstraction in Chapter 3 was also pretrained on this objective.

> [...] the primary benefit provided by pretraining is to improve the model's ability to substitute similar words or word phrases by ensuring they are close to each other in the representation space.

While the connection to compositional generalization is less evident compared to the two other auxiliary objectives proposed, we believe it is important to include language modeling as one of the auxiliary objectives compared because of the large empirical success it has achieved in both downstream language tasks and on linguistically motivated evaluation suites. Outside of our immediate question of compositional generalization, language modeling as an auxiliary objective improved performance on numerous natural language understanding tasks (Wang et al. 2019b), and has achieved nontrivial success on tests of syntactic generalization (Gulordava et al. 2018; Goldberg 2019). Furthermore, language modeling has also been shown to be superior to various alternative auxiliary objectives for both downstream tasks and linguistic evaluation (Kim et al. 2019a; Wang et al. 2019a; Pruksachatkun et al. 2020).

We adopt a version of the language modeling objective that is compatible with the sequence-to-sequence setup that we have used in Chapter 4, named *denoising* by Raffel et al. (2020). As illustrated in (6) (repeated from (2)), the input contains some corruptions[9] marked by special tokens. Given these corrupted sentences as inputs, the model must predict the tokens that correspond to the corrupted part (i.e., denoise the input sentence).

(6) a. INPUT: How can I ⟨x⟩ my cat ⟨y⟩ human ⟨z⟩ ?

b. OUTPUT: ⟨x⟩ make ⟨y⟩ understand ⟨z⟩ language

[9] We followed Raffel et al. (2020) and corrupted 15% of the input.

5.2.1.4 Multiple Auxiliary Objectives

The potential benefits of CCG supertagging (Section 5.2.1.1) and glossing (Section 5.2.1.2) objectives we hypothesized are complementary to each other. We proposed CCG supertagging primarily as a way to improve structural generalization, and glossing as a way to improve generalization errors due to unfaithful single lexical retrieval errors. Then, would a model that is auxiliary-trained on both objectives benefit from both? To address this question, we additionally included a model auxiliary-trained on both CCG supertagging and glossing in the comparison.

5.2.2 Datasets for Auxiliary Training

We used CCGBank (Hockenmaier and Steedman 2007) for the CCG supertagging auxiliary training objective. CCGBank is a CCG-annotated version of the Wall Street Journal corpus in the Penn TreeBank (Marcus et al. 1993). For the other two auxiliary objectives, we used the text of CCGBank to create the input/output pairs to keep the comparison across auxiliary objectives as fair as possible. As an illustration, see (7)-(9) showing what the input/output pairs for each objective would be for the same text 'John buys shares'.

(7) a. INPUT: John buys shares
    b. OUTPUT (CCG SUPERTAGGING): NP (S\NP)/NP NP

(8) a. INPUT: John buys shares
    b. OUTPUT (GLOSSING): John′ buys′ shares′

(9) a. INPUT: John <x> shares
    b. OUTPUT (LANGUAGE MODELING): <x> buys

For CCG supertagging and glossing, the inputs are sentences (n = 38,015). For language modeling, we preprocessed the input to add corruption using the default

setup of Raffel et al. (2020),[10] which segments the input data into 512-token chunks (n = 2321).

5.2.3 Modification to the Compositional Generalization Dataset

We made one important modification to the compositional generalization dataset described in Section 4.4 to maintain the distributional mismatch across training and generalization. In the original version of the dataset, we used real words as the lexical items that only appear in controlled contexts to create the distributional mismatch (e.g., the word *hedgehog* appears only in the subject noun phrase). Using real words is not a problem as long as models are trained from scratch on only this dataset. However, if we introduce auxiliary tasks that brings in additional data outside of our dataset (CCGBank in our case), there is no guarantee that the distributional mismatch is maintained unless we modify the auxiliary task dataset to impose the same distributional control (e.g., the CCGBank data might contain examples in which the word *hedgehog* appears in the object noun phrase, breaking the distributional gap).[11] To avoid this issue, we created a version of the dataset that replaces the lexical items that are involved in constructing the distributional mismatch with special tokens ($[w_n]$)[12] as follows:

(10) a. ORIGINAL: Emma appreciated the hedgehog.

[10] https://github.com/google-research/text-to-text-transfer-transformer

[11] One may argue that since the compositional generalization task is a separate task from the auxiliary tasks, the intended systematic gap between training/generalization is still maintained. Despite this, we believe it is better to have a more strictly controlled experimental setup, because we do not know what the exact effect of having encountered held-out contexts during auxiliary training would be.

[12] Each bolded lexical item in Table 4-1 for the lexical generalizations is substituted by a different $[w_n]$ token. For example, all occurrences of *hedgehog* are substituted with $[w_0]$, and all occurrences of *Lina* are substituted with $[w_1]$. This substitution applies to both the input sentence and the output logical form. If the input token and the corresponding output token are different (e.g., the representation of verbs in the output are always the root form, even if the input is past tense), they are substituted with two different tokens $[w_i]$ and $[w_j]$.

$\rightsquigarrow$ *hedgehog(x_3) ; appreciate.agent(x_1, Emma) AND appreciate.theme(x_1, x_3)

b. MODIFIED: Emma appreciated the [w_0].

$\rightsquigarrow$ *[w_0](x_3) ; appreciate.agent(x_1, Emma) AND appreciate.theme(x_1, x_3)

Since the special tokens do not appear in the auxiliary data (i.e., [w_n] appears 0 times in all of the auxiliary data we use), the intended systematic gap is guaranteed to be maintained. However, even with this modification, the systematic gap in the structural generalization cases cannot be guaranteed if additional data are given to the models through auxiliary training. For example, it is difficult to remove all sentences containing subject PP modification from the auxiliary training data. This issue is harder to resolve under a setup where the model's auxiliary training data is not publicly available or difficult to inspect due to its size (e.g., models trained on all of Wikipedia or models trained on BooksCorpus which is no longer available). For the test of generalization to deeper levels of embedding, it is reasonably likely that embeddings of depth 6 or greater would not occur in the training data (Karlsson 2010), but this prediction is only speculative.[13]

Note that the methodology used in works that leverage auxiliary objective-trained models such as Furrer et al. (2020) (work on SCAN) and Tay et al. (2021) (work on our dataset from Chapter 4) suffer from exactly the problem we describe in this section. They directly take the datasets that have training/generalization distributional mismatch and evaluate models that have been trained on additional data, potentially breaking the distributional mismatch. Hence, we should acknowledge that

[13]One possible way to alleviate the issue of maintaining (pre)training/generalization distributional mismatch is to use language models that are pretrained on languages other than English. This resolves the issue of held-out lexical items, so no modification is needed to the current dataset. While there is still no guarantee of held-out structural generalization examples not appearing in the pretraining data (e.g., depth n embeddings), at least these examples do not share surface forms with the English examples. With the working assumption that knowledge about one language can partially transfer to different languages, a language model pretrained on different languages may provide a helpful inductive bias for compositional generalization.

the improvements claimed to have been achieved through auxiliary training have been obtained in the presence of an important confound. We investigate this issue further in Section 5.3.2.

5.2.4 Model and Training

We used the same set of sequence-to-sequence ANN models (LSTM, bidirectional LSTM (BiLSTM), and Transformer) that we used in Section 4.5. The key difference in the setup is that we train the models on an auxiliary objective before training them on the compositional generalization task. The model specifications are kept identical except for the size of the vocabulary. This change was necessary because the additional training data used for the auxiliary training contained vocabulary items that are not present in the training data of the compositional generalization task. Because the vocabulary size of the model had a nontrivial effect on model performance,[14] we trained a new set of baseline models (i.e., models only trained on the compositional generalization task following the setup in Section 4.5) that have the same vocabulary size as the models trained on the auxiliary objectives. The number of parameters in these models were larger than the models in Section 4.5 because of the increased vocabulary size (Transformer: 54M; LSTM: 56M; BiLSTM: 55M).

We used the same early stopping patience of 5 for the auxiliary training objectives.[15] We trained each ⟨model, auxiliary objective⟩ combination with 10 different random

[14]There are various reasons behind this. First, the default initialization function in the OpenNMT package that we used to implement the models is Xavier initialization (Glorot and Bengio 2010), which is sensitive to the number of incoming/outgoing connections. Since the number of incoming connections to the embedding layer depends on vocabulary size, models with different vocabulary size would have different initializations of the embedding layer even when the random seed is kept constant. Second, the default OpenNMT setup includes label smoothing, which regularizes over all labels (including ones not directly seen in the training data). This means that models with different vocabulary size would have different numbers of labels to smooth over, even when they are trained on exactly the same dataset.

[15]There are other (potentially more effective) ways to conduct auxiliary training that we have not tested. For instance, one may choose to mix in auxiliary training with the target task training, or choose to iterate over the auxiliary training dataset for a fixed number of times rather than using early stopping.

seeds.

Model	Dev.	Test	Gen.
Transformer (baseline)	0.96 (± 0.02)	0.96 (± 0.02)	0.37 (± 0.19)
+ CCG supertagging	0.96 (± 0.03)	0.96 (± 0.03)	0.42 (± 0.11)
+ Glossing	0.98	0.98	**0.62 (± 0.04)**
+ Language modeling	0.96 (± 0.02)	0.96 (± 0.02)	0.37 (± 0.14)
+ Glossing + CCG supertagging	0.98 (± 0.02)	0.97 (± 0.02)	0.58 (± 0.07)

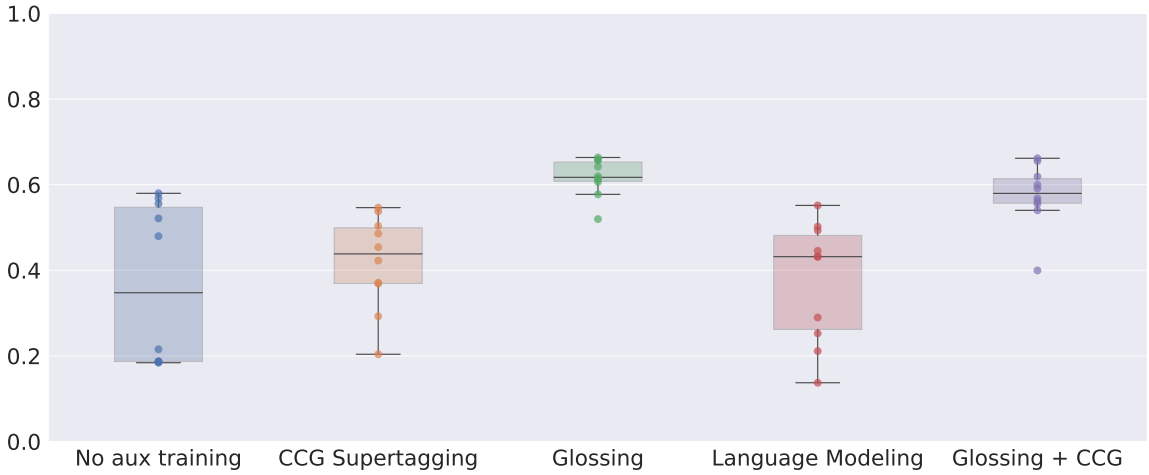


Figure 5-1. Comparison of auxiliary training objectives for the **Transformer** model. Only standard deviations greater than 0.01 are shown.

5.2.5 Results

Figures 5-1 through 5-3 show the generalization accuracy of Transformer, LSTM and BiLSTM models, respectively, with the 3 different auxiliary objectives and 1 multi-objective discussed in Section 5.2.1. The leftmost column (the blue box plot) of each figure denotes the accuracy distribution of the model without any auxiliary training (i.e., randomly initialized and then only trained on the compositional generalization task). This serves as the baseline to examine the effect of auxiliary training. The change in the baseline distributions of generalization accuracy from the distributions in Table 4-3 is due to the aforementioned difference in the base models' vocabulary size (see Section 5.2.4).

The results show that there is an interaction effect of auxiliary training objective and model architecture. Specifically, the glossing task substantially improved the generalization of Transformer models (and also reduced the variance across random seeds),[16] but did not have this effect on LSTM and BiLSTM models' performance. There was a small benefit of the language modeling objective for BiLSTM models, but the objective did not improve generalization in LSTM and Transformer models. Other than these cases, we did not find a noticeable impact of other ⟨model, auxiliary objective⟩ combinations. While adding both glossing and CCG supertagging as auxiliary objectives to the Transformer model also yielded improvements, the improvements were similar to models with just the glossing task as an auxiliary objective, suggesting that most of the improvements are due to the glossing task.

Model	Dev.	Test	Gen.
LSTM (baseline)	0.99	0.99	0.21 (± 0.10)
+ CCG supertagging	0.99	0.99	0.21 (± 0.11)
+ Glossing	0.96 (± 0.06)	0.96 (± 0.05)	0.20 (± 0.14)
+ Language modeling	0.99	0.99	0.20 (± 0.07)
+ Glossing + CCG supertagging	0.99	0.99	0.24 (± 0.13)

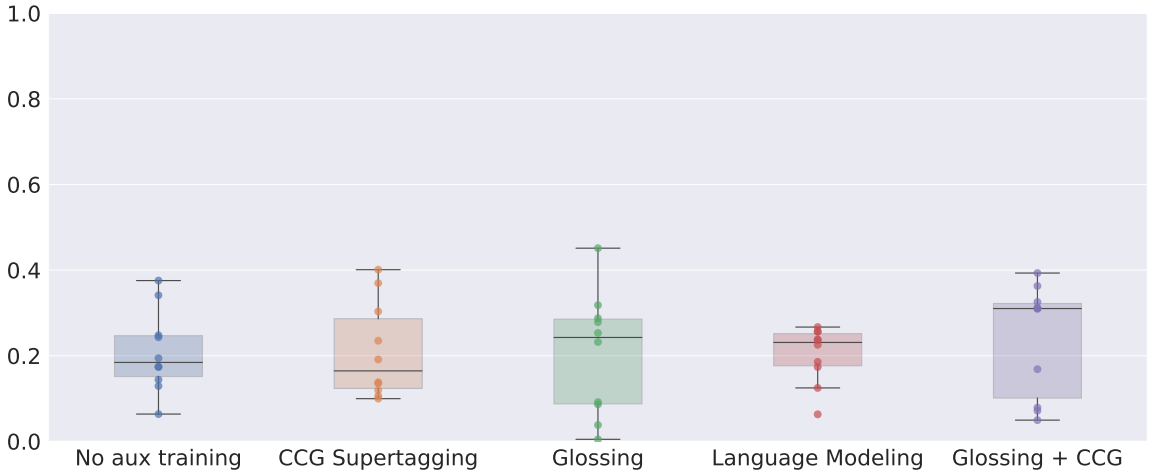


Figure 5-2. Comparison of auxiliary training objectives for the **LSTM** model. Only standard deviations greater than 0.01 are shown.

[16] Adding the glossing task to Transformer models also improved development and test performance.

Model	Dev.	Test	Gen.
BiLSTM (baseline)	0.99	0.99	0.17 (± 0.09)
+ CCG supertagging	0.97 (± 0.04)	0.97 (± 0.04)	0.15 (± 0.12)
+ Glossing	0.99	0.99	0.19 (± 0.09)
+ Language modeling	0.99	0.99	0.24 (± 0.07)
+ Glossing + CCG supertagging	0.99	0.99	0.20 (± 0.09)

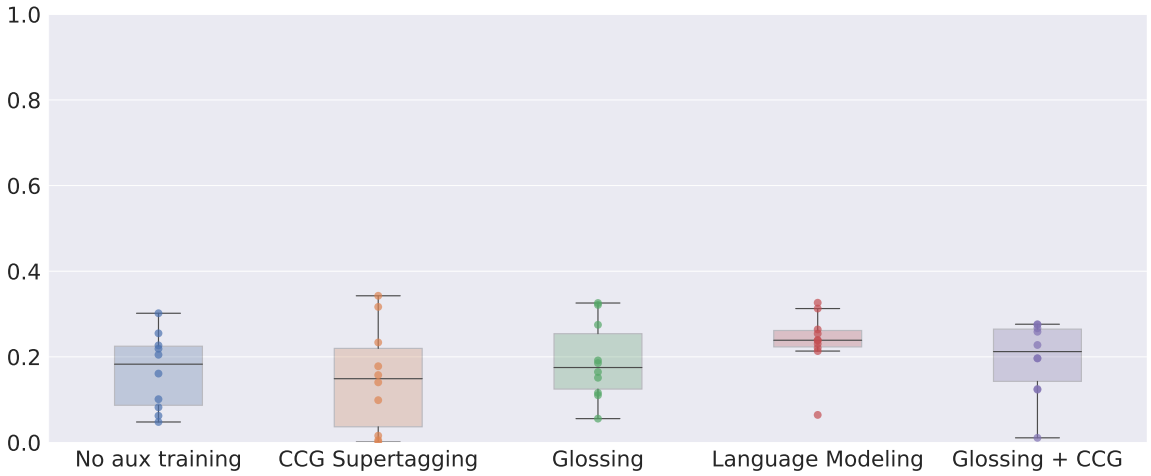


Figure 5-3. Comparison of auxiliary training objectives for the **bidirectional LSTM (BiLSTM)** model. Only standard deviations greater than 0.01 are shown.

5.2.6 Error Analysis: Effect of the Glossing Objective

We saw that the glossing task added as an auxiliary objective to the Transformer models yielded substantial gains in performance. Our hypothesis in Section 5.2.1.2 when proposing the glossing task was that it would help reduce single lexical errors. We compared the error patterns of the baseline models and the models auxiliary-trained on the glossing objective to verify whether this prediction was borne out. The average single lexical error rate in baseline Transformer models across the 10 random runs was 25.1% (i.e., 25.1% of all generalization test examples were single lexical errors). This error rate was substantially reduced in Transformer models with glossing as auxiliary training, with an average single lexical error rate of 5.7%. As an illustration, in one of the random runs, the model's single lexical error (11-b) was successfully corrected to (11-c):

(11) a. INPUT: A [w_1] burned Sophia .

b. OUTPUT (BASELINE TRANSFORMER): director(x_1) AND burn.agent (x_2, x_1) AND burn.theme (x_2, Sophia)

c. OUTPUT (TRANSFORMER + GLOSSING): [w_1](x_1) AND burn.agent (x_2, x_1) AND burn.theme (x_2, Sophia)

It remains to be answered why the glossing objective fails to be effective for models with other architectures. We did find that Transformer models were more effective at learning to perform the glossing task itself, with the lowest average development set cross-entropy loss on the glossing task (Transformer: 0.17; LSTM: 0.37; BiLSTM: 0.26). Then, perhaps, changing the auxiliary training setup so that the LSTM/BiLSTM models could achieve better performance on the glossing may lead to better generalization on the target task. Why learning the glossing task was easier for the Transformer models under our setup is unclear, since both LSTM/BiLSTM models are equipped with an attention mechanism that allows the models to attend to the input sequence during the generation of the output sequence. Note that the attention mechanism is directly relevant to the correspondence relation in faithfulness constraints discussed in Section 5.2.1.2. The attention mechanism, conceptually, should make the task of glossing quite easy—therefore, the discrepancy we observe in the difficulty of the glossing task between Transformer and LSTM models must be rooted in the difference in how the encoder-decoder attention is actually implemented[17]—the Transformer models use multi-headed scaled dot-product attention whereas the LSTM models use global dot product attention.

[17]The self-attention mechanism in Transformer models does not seem very relevant to the glossing task, because the glossing task primarily concerns deducing the correct input-output correspondence (rather than within-input or within-output dependency that self-attention would be helpful for).

5.2.7 Existing Ideas with a Similar Effect to the Glossing Objective

Adding a glossing objective is not the only way to provide bias towards faithful translations. One existing idea in sequence-to-sequence learning (including semantic parsing and machine translation) that may yield a similar effect is copy mechanism. Models with a copy mechanism are augmented with the capacity to directly copy tokens from the input when producing the output, in addition to the usual mechanism to produce tokens in the output vocabulary space (Jia and Liang 2016; Gu et al. 2016; Gulcehre et al. 2016; Zhang et al. 2018, *i.a.*). The copy mechanism is relevant to our discussion of faithfulness constraints because the output tokens produced via the copy mechanism would never violate MAX and DEP constraints. However, the copy mechanism has an important limitation. That is, it can only reflect a fixed view of what the mapping function from the object language to representation language can be. Namely, the mapping function (let's call this function *rep*) assumed by the copy mechanism is the identity function ($rep(i) = i$). While this may be a reasonable assumption under some scenarios, it is easy to see where this would fall short—all cases where the mapping function should not be identity. For instance, in our dataset, the token that corresponds to the lexical translation[18] of a verb in the representation language is always the root form, although the verbs in the input sentence may be tensed (e.g., in (10), the lexical translation of 'appreciated' in the input is 'appreciate', not 'appreciated').[19] There is no way to handle non-identity mappings like this through the copy mechanism.

[18] I use this term to refer to the output token 'dance' in John danced $\rightsquigarrow$ dance.agent(John); output 'dance' is the lexical translation of 'danced' in the input, whereas the lexical part plus the saturated function would be the full translation of 'danced'.

[19] Technically, no input-output mapping under the current experimental setup is the identity function, because the input and output vocabulary spaces are disjoint. Only when the vocabulary space for the object and representation language are shared, the identity assumption can have any effect.

Akyürek and Andreas (2021) offers a more promising generalization of the copy mechanism in this regard, called the lexical translation mechanism. Briefly, this mechanism enables the models to learn a different mapping function, moving away from the strong assumption of the copy mechanism that the mapping function is the identity function. The mapping function allowed under their setup is a token-to-token lexical mapping function: $rep(i) = j$ $(i \in \mathbf{V}_{in}, j \in \mathbf{V}_{out})$, where $\mathbf{V}_{in}$ is the input vocabulary space and $\mathbf{V}_{out}$ is the output vocabulary space ($\mathbf{V}_{in}$ and $\mathbf{V}_{out}$ may be shared). This method still maintains a strict restriction on what the mapping function could be, although is not necessarily a limitation. However, in a semantic representation language like ours, the translation rule is rarely at a single token-level (with the exception of proper nouns: John $\rightsquigarrow$ John′). Conceptually, this approach differs from the glossing auxiliary task approach because the mapping function implied by the glossing task can be revised in the training phase of the target task, and there are no hard constraints on this revision. In effect, both approaches seem to contribute similar degrees of improvement in lexical generalization, for similar reasons related to strengthening the models' faithfulness bias.

5.3 Experiment 2: Effect of Dataset Size in Language Modeling Pretraining

In Experiment 1, we saw that language modeling as an auxiliary objective did not provide clear benefits for compositional generalization. However, the models that are commonly used in NLP are trained on a much greater amount of data than the amount we used in Experiment 1 (and it has been observed that downstream task performance in general improves as the amount of auxiliary training data increases; e.g., Kaplan et al. 2020). In this follow-up experiment, we test a Transformer encoder-decoder model that is trained on a much larger amount of data to investigate whether the generalization capacity improves when the auxiliary training on language modeling is

conducted at scale. Furthermore, we train several versions of this model with different amounts of pretraining data to gauge the effect of the amount of data that models are exposed to.

5.3.1 Model and Training

We use the T5-small model of Raffel et al. (2020), which is a Transformer-based sequence-to-sequence model with 60 million parameters.[20] We make a departure from the specific sequence-to-sequence Transformer model that we have been using in Chapter 4 and Experiment 1 in this chapter, since this lets us evaluate the effect of pretraining that is outside the scale of computing resources available to us. Specifically, the T5-small model is pretrained on 34 billion tokens of English text from the Colossal Clean Crawled Corpus (C4[21]; see Dodge et al. 2021 for more details about the corpus). We take this model and further train it (i.e., finetune it) on our compositional generalization dataset. We also take a randomly initialized T5 and pretrain it on varying amounts of data: 0 (i.e., randomly initialized and not pretrained), 1M, 5M, 25M, 50M, 100M, and 1B tokens. We used 10% of the datasets, respectively, as development sets to determine early stopping points with a patience of 5. Then, we finetune each model on the compositional generalization task and compare the results. We used English Wikipedia data[22] instead of C4 for the additional pretraining variations because using C4 (even for subsampling purposes) required running an extremely compute-intensive preprocessing pipeline.[23] The finetuning was run 5 times for each model using different random seeds.

[20]The choice of T5-small over other variations like T5-base and T5-large is due to computing resource constraints. We initially conducted several pilot experiments using the T5-base model, and found that the generalization performance was not substantially different from T5-small (but we have not tested variations of the training data).

[21]https://www.tensorflow.org/datasets/catalog/c4

[22]Downloaded from https://www.tensorflow.org/datasets/catalog/wikipedia

[23]https://github.com/google-research/text-to-text-transfer-transformer states that ~7TB of raw data must be downloaded first and should be put through a preprocessing step that takes ~335 CPU days on a machine without support for distributed computing.

Because we used the modified version of the generalization dataset (Section 5.2.3) replacing some of the lexical items with special tokens $[w_n]$, we added these tokens to the vocabulary of the T5-small model before the finetuning step. The input and output were tokenized with the tokenizer of T5[24], which is based on SentencePiece (Kudo and Richardson 2018). This tokenizer uses subword units and treats whitespaces as a character, rather than treating whitespace boundaries as tokenization boundaries.

Number of tokens in pretraining	Generalization Accuracy	Data
0	0.52 (± 0.04)	-
1M	0.43 (± 0.08)	Wikipedia
5M	0.44 (± 0.10)	Wikipedia
25M	0.24 (± 0.05)	Wikipedia
50M	0.22 (± 0.06)	Wikipedia
100M	0.37 (± 0.04)	Wikipedia
1B	0.27 (± 0.06)	Wikipedia
34B	0.05 (± 0.003)	C4

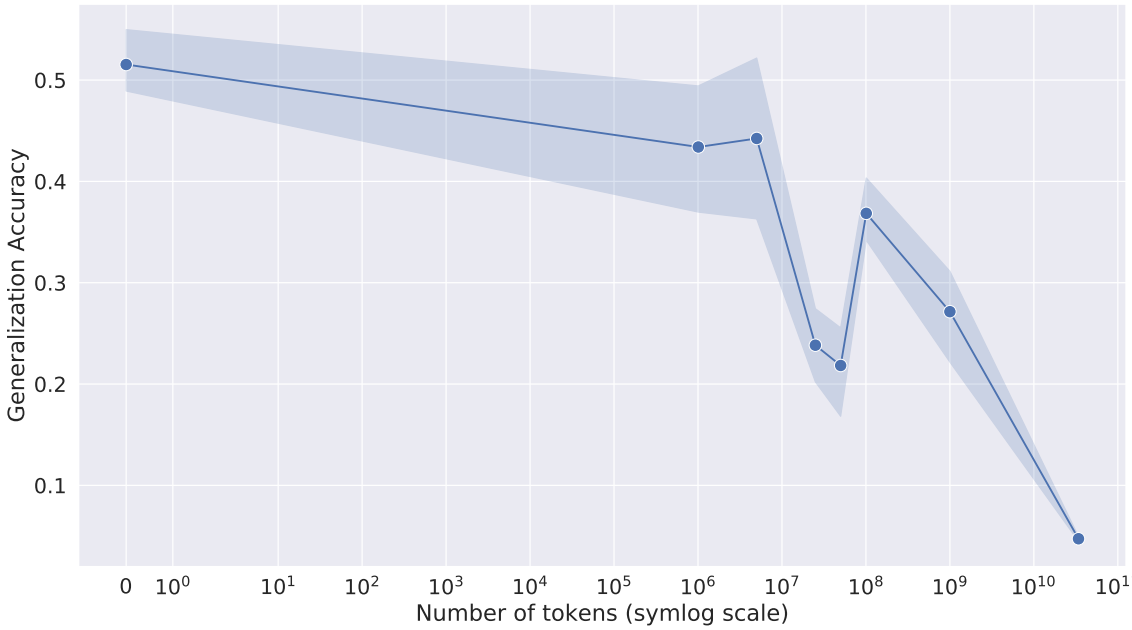


Figure 5-4. Generalization accuracy of T5-small models pretrained on different amounts of data. The x-axis shows the number of tokens in symmetrical log scale which allows us to include the value 0 in the plot.

[24] https://huggingface.co/transformers/model_doc/t5.html#t5tokenizer

5.3.2 Results

Figure 5-4 shows the generalization accuracy of models pretrained with the language modeling objective on varying amounts of data. Surprisingly, the result demonstrates a clear negative effect of pretraining—more data provided to the models during pretraining led to lower generalization accuracy.

The effect of real words in the generalization dataset: Our finding that more pretraining leads to worse generalization performance is contrary to the findings of Furrer et al. (2020) and Tay et al. (2021). We believe this is due to the issue we raised in Section 5.2.3. When we used the compositional generalization dataset without the word substitution modification described in Section 5.2.3 (e.g., substituting the real word *hedgehog* with $[w_0]$ to maintain the training-generalization systematic gap of the word not appearing in the object position) to finetune the T5-small model, the results were drastically different, with the fully pretrained T5-small model at 79% generalization accuracy.[25] This suggests that breaking the distributional mismatch at the pretraining stage does substantially impact the results, and we must make the suggested modification if data other than the compositional generalization dataset itself is additionally introduced.

Failure to learn exposure examples in the 34B-token model: An analysis of the training errors revealed that the models pretrained on 34B tokens did not properly learn the exposure examples (Section 4.4.3), learning the correct input-output mapping for only around 8% of the exposure examples (the model with the maximum accuracy among the 5 random seeds had 13.3% accuracy, equivalent to correctly learning 2/15 of the exposure examples).[26] Since learning the meanings of

[25] This aligns with the reported performance in Tay et al. (2021).

[26] There are only 15 exposure examples in the training set that contains 24,000 examples in total. So, if the model succeeds in learning everything else except for the exposure examples, the models could still display high training set accuracy. The development set, which was used to determine the early stopping point, did not contain exposure examples. Therefore, it is possible that the models do

the exposure examples is critical, we re-trained the 34B-token models on only the exposure examples after they were trained on the compositional generalization task, until the accuracy on the exposure examples was 15/15 (100%). However, even after the models successfully learned the exposure examples, the generalization accuracy did not improve substantially (~7%), suggesting that the low accuracy shown in Table 5-4 was not just due to this failure to learn the exposure examples during training.

5.3.3 The Effect of Pretraining on Structural Accuracy of Outputs

We conducted a manual error analysis of the first 50 errors of the best performing model among the models without pretraining (59% generalization accuracy) and the best performing model among the models that were pretrained on 34B tokens (5% generalization accuracy). We found that 50/50 errors in the 34B-token model was a single lexical error, where the model correctly predicted everything else in the output except for the special token $[w_n]$. On the other hand, the error patterns of the model with no pretraining were more diverse, including single lexical errors (9/50), missing conjuncts (4/50), producing x_n instead of lexical denotations (2/50), and a combination of these errors (23/50). We cannot conclude too much from this analysis, since the absolute numbers of errors between these two models were very different (8666 and 19,293, respectively), and it could be the case that the former is a subset of the latter (i.e., the error patterns seem more diverse in the former just because all of the single lexical errors have been fixed). To verify whether this is the case, we also examined the 50 errors in the no pretraining model and tracked how the predictions changed with pretraining for these particular examples in the 34B-token model. The result showed that the pattern of the errors *shifted* to single lexical errors: 41/50 errors that were not single lexical errors in the no pretraining model changed to single lexical

not learn the exposure examples properly. However, this failure to learn the exposure examples was only observed in the 34B-token models.

errors. This suggests that the pattern of errors are becoming more uniform in the 34B-token model (12)-(13). For example, the errors of missing a full conjunct (12-c) or missing a full conjunct *plus* producing a lexical denotation in place of a skolem constant (13-c) (missing recipient conjunct and produced *Emma* in place of x_2) in the model without pretraining both shifted to a single lexical error of producing the wrong $[w_n]$ token inside correct output structures.

(12) a. INPUT: $[w_5]$ drew Natalie.

b. OUTPUT (GOLD): draw.agent(x_1, $[w_5]$) AND draw.theme(x_1, Natalie)

c. OUTPUT (NO PRETRAINING): draw.agent(x_1, $[w_5]$)

d. OUTPUT (34B-TOKEN): draw.agent(x_1, $[w_3]$) AND draw.theme(x_1, Natalie)

(13) a. INPUT: The cat $[w_{19}]$ Emma the donut.

b. OUTPUT (GOLD): *cat(x_1) ; *donut(x_5) ; $[w_{20}]$.agent(x_2, x_1) AND $[w_{20}]$.recipient(x_2, Emma) AND $[w_{20}]$.theme(x_2, x_5)

c. OUTPUT (NO PRETRAINING): *cat(x_1) ; *donut(x_5) ; $[w_{20}]$.agent(x_2, x_1) AND $[w_{20}]$.theme(x_2, Emma)

d. OUTPUT (34B-TOKEN): *cat(x_1) ; *donut(x_5) ; $[w_{18}]$.agent(x_2, x_1) AND $[w_{18}]$.recipient(x_2, Emma) AND $[w_{18}]$.theme(x_2, x_5)

In light of this observation of the shift in the pattern of errors, we introduce *output structure match rate*—the ratio of structurally correct predictions—as an additional metric to gauge whether the rate of structurally correct outputs increased following the increase in the size of the training data. As shown in Figure 5-5, it was indeed the case that models with greater amount of language modeling pretraining produced outputs that structurally matched the target outputs, contrary to the overall decrease in full accuracy (Figure 5-4). If we believe that similar patterns of error share a similar underlying cause, it might be easier to identify solutions to the errors in the models

Number of tokens in pretraining	Output Structure Match Rate	Data
0	0.70 (± 0.03)	-
1M	0.68 (± 0.005)	Wikipedia
5M	0.69 (± 0.04)	Wikipedia
25M	0.69 (± 0.03)	Wikipedia
50M	0.68 (± 0.01)	Wikipedia
100M	0.75 (± 0.03)	Wikipedia
1B	0.79 (± 0.02)	Wikipedia
34B	0.80 (± 0.03)	C4

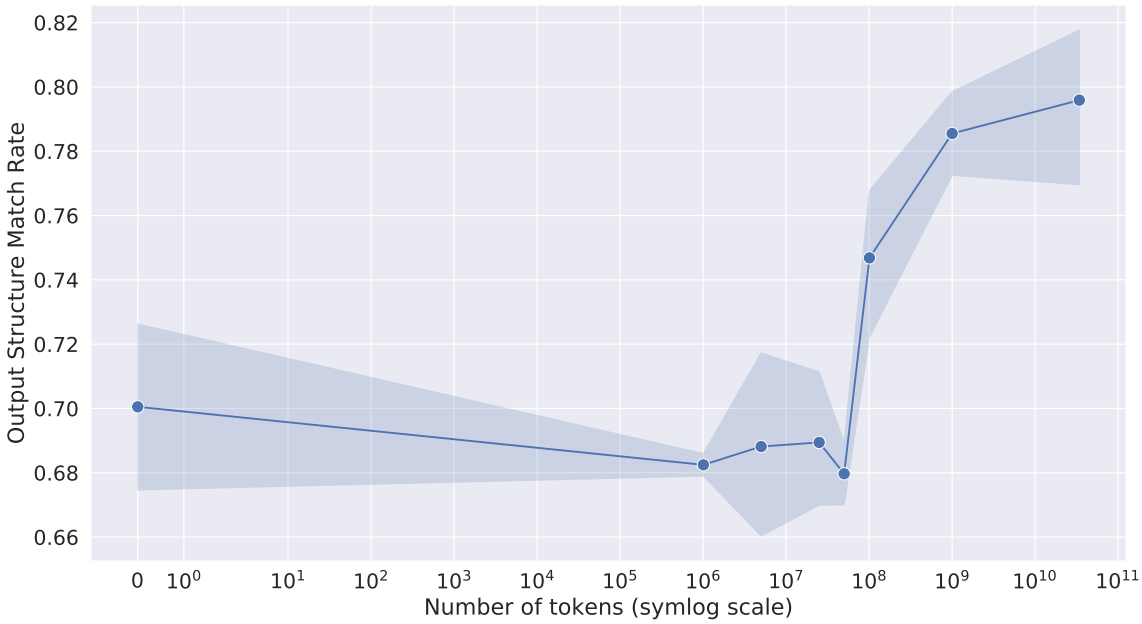


Figure 5-5. Output structure match rate of T5-small models pretrained on different amounts of data. The x-axis shows the number of tokens in symmetrical log scale which allows us to include the value 0 in the plot.

pretrained with more data, compared to the models with less pretraining data that showed more heterogeneous patterns of error. If we were to identify a solution that improves single lexical errors, we would be able to achieve around 80% generalization accuracy.

Language modeling objective promotes probable translations: While the negative impact of language modeling pretraining on generalization performance may be surprising, given an across-the-board beneficial effect of language modeling on

many language tasks as discussed in Section 5.2.1.3, a constraint-based description of the output patterns that we used to describe the effect of the glossing objective is also relevant here. Recall that we described the glossing objective as having an effect of promoting faithful translations that are likely in competition with translations that are more probable given the training data. The language modeling objective seems to have exactly the opposite effect—promoting probable translations over faithful translations, exacerbating the issue of single lexical errors as a result. We leave the correction of these single lexical errors to future work, but one way to move forward could be combining the glossing objective with the language modeling objective. This might be a more challenging problem than simply adding the glossing objective, because the two auxiliary objectives seem to be directly at competition in terms of the generalizations they promote.

5.4 Limitations and Future Work

The goal of the array of experiments conducted in this section was to identify an auxiliary objective (or a set of objectives) that aids compositional generalization. We proposed and tested three auxiliary objectives (added to the models used in Chapter 4): CCG supertagging, glossing, and language modeling. The results showed that the glossing task greatly benefited Transformer models, correcting the output patterns that violated faithfulness constraints that we predicted the objective would be helpful in improving. We also observed a minor contribution of language modeling to BiLSTM models. We furthermore tested the effect of scaling up the amount of language modeling pretraining, finding that generalization performance in fact was hurt by more data seen during pretraining. While this is a surprising finding, an error analysis revealed that the errors in models pretrained on greater amounts of data became more structurally accurate. The homogeneity of the error patterns suggests that these models may be easier to improve (compared to models with no pretraining

that show more heterogenous error patterns). The expected ceiling performance if we are able to debug the single lexical errors in the pretrained models is around 80%, which amounts to almost all of the lexical generalization cases. Furthermore, most errors were single lexical errors, implying that the language modeling objective has the opposite effect to the glossing objective; it promotes probable translations over faithful translations.

While these are encouraging findings, none of the models we tested achieved any degree of success on structural generalization (around 0% generalization accuracy for all models tested). To this end, we plan to continue investigating other factors that could contribute helpful learning biases. Furthermore, representational analysis of the models we trained could be used to investigate how (or whether) the linguistic structures crucial for solving the structural generalization cases are currently represented in these models. We expect analysis methods such as ROLE (Soulos et al. 2020) to be useful in this process, providing us with new insights to make progress on structural generalization.

Chapter 6

Conclusion and Future Work

Summary of findings: In this dissertation, I have investigated the patterns of compositional linguistic generalization in several artificial neural networks (ANNs). First, I showed that a Transformer model pretrained on the language modeling objective achieves a nontrivial degree of success on category abstraction—an important precondition for compositional generalization—using a method adapted from a developmental study. Second, I showed that Transformer and LSTM models only achieve a limited degree of success in matching the expected compositional generalization patterns of human learners (with the caveat that the space of Transformer and LSTM model configurations is much larger than the set of models tested). In particular, the models I tested completely failed to assign meaning representations to sentences with novel structure. Three auxiliary training objectives (CCG supertagging, glossing, and language modeling) were added to the Transformer and LSTM models to gauge whether these objectives can change the inductive bias of the models, so that they display stronger compositional generalization. The result showed that the glossing objective contributed a helpful learning bias to the Transformer model, increasing the mean generalization accuracy from 37% to 62%, and furthermore substantially reducing the variance in performance across different random seeds (standard deviation of 0.19 to 0.04). This suggests that the ANN models without auxiliary training

lack sufficient inductive bias towards translations to meaning representations that satisfy faithfulness constraints MAX and DEP, and this inductive bias in Transformer models could be injected by adding an auxiliary objective of glossing, a task that maximally satisfies these faithfulness constraints. Additionally, I explored the effect of scaling up the amount of auxiliary training with the language modeling objective (amounts ranging from 1M to 34B tokens), which led to a surprising finding that the generalization accuracy decreased as the amount of auxiliary training data increased. An error analysis revealed that the pattern of errors became more homogeneous as the training data increased, with most of the errors in the largest auxiliary-trained model consisting of single lexical retrieval errors (with all structural aspects of the output correct). This suggests that the language modeling objective has an opposite effect to the glossing objective—it strengthens the inductive bias for probable translations, promoting probable translations over faithful ones.

Motivation for a follow-up human subject study: While auxiliary training with glossing and language modeling objectives led to interesting modifications to the generalization patterns to some models tested, no notable changes were observed in their structural generalization—all models almost completely failed to translate novel structures regardless of auxiliary training. The proposed structural generalization tests also highlight an important gap in the experimental literature: they have not been experimentally attested in human learners, although the shared assumption is that human learners are able to assign meaning to novel structures of the kind that were tested. For instance, a widely adopted view that language is recursive predicts that arbitrary degrees of embedded structures, even outside of the depth directly observed by the learners, can be assigned correct meaning. McCoy et al. (2021) tested this assumption on unseen degrees of embedding and found promising results, although the recursive structure they examined was center-embedding (rather than the right-branching structures in our tests). They also tested for acceptability of the

structures rather than production or interpretation. Generalization of modification to different syntactic positions, to the best of our knowledge, has not been empirically attested, although the assumption is that human learners will be able to assign appropriate meaning to modifiers appearing in novel syntactic positions (as long as the novel syntactic positions occupied by the phrase that contains the modification is allowed by the grammar). For instance, we would expect a speaker of English who can interpret nominal prepositional phrase (PP) modification to be able to interpret PP modification in all syntactic positions that allows for noun phrases (NPs), even if they have not been exposed to a PP modification in all possible syntactic positions that NPs can occupy—it is unlikely that exposure to a PP-modified grammatical subject of a triply nested complementizer phrase (CP) be necessary before being able to interpret PP modification in that position. However, this is a theoretical prediction (e.g., a context-free rewrite rule like NP → NP PP and a corresponding translation function would give rise to this prediction) rather than an empirical observation. This leads to a proposal for a follow-up study with human subjects, which I discuss in detail below.

6.1 Proposal: Testing Structural Generalization in Human Learners

Much experimental evidence supports lexical generalization that requires novel composition of a known lexical item and a known structure in human learners (e.g., Tomasello and Olguin 1993; Brooks et al. 1999; Rowland and Noble 2010; Messenger and Fisher 2018, *i.a.,* see Section 2.2). However, our empirical understanding of structural generalization (production and comprehension of novel structures) in human learners is comparatively limited. In fact, most empirical evidence in the literature supporting structural generalization is indirect: for instance, the distributional discrepancy between PP modification in different syntactic positions. I analyzed the Epochs corpus of Perfors et al. (2011) and found that PP modification on the grammatical object occurs

much more frequently (≥ 100 times) compared to PP modification on the grammatical subject (12 times). Furthermore, all subject PP modification are subjects of depth 0-1 CPs, showing that the distribution of modification is skewed. Furthermore, learners are unlikely (if not impossible) to be exposed to nominal PP modification (i.e., [NP PP]) in *all* possible syntactic positions that an unmodified NP may occupy (e.g., subject, object, part of an oblique argument PP, subject of an degree 2 embedded CP, etc.). Nevertheless, the assumption about the adult English speakers' knowledge of PP modification is that it is general ('PP modifiers modify NPs') rather than specific to a certain syntactic position ('PP modifiers modify NPs in subject positions, NPs in object positions, NPs in object positions of CPs of depths 0–1...').

While a distributional mismatch between modification in different syntactic positions is attested, it remains untested whether this kind of generalization is possible without having seen *any* examples of modification in other grammatical roles, since data from corpora do not account for the full input available to the learners. I propose to test this generalization in human learners using artificial language learning, in order to gain more insights about structural generalization in human learners. This will also help establish a credible human estimate of structural generalization cases tested in Chapter 4. Artificial language learning has long been used to experimentally investigate human learning biases, by testing how humans generalize to novel examples (Braine 1963; Morgan et al. 1987; Wonnacott et al. 2008; Reeder et al. 2017, *i.a.*). This experimental paradigm has been used across various linguistic domains; for example, Finley and Badecker (2009) tested generalization of phonological features to novel segments, Culbertson et al. (2012) have tested generalization in word order, and Saratsli et al. (2020) have tested the learnability of evidentiality. Specifically, I propose to test whether adult English speakers are able to generalize to subject NPs a modification strategy that only appears with object NPs. This generalization corresponds to a

specific subset of structural generalization tested in Chapter 4 (Section 4.7.2),[1] on which all ANN models, auxiliary-trained or not, performed close to 0% accuracy.

6.1.1 Pilot Study

Methodology: I designed a pilot artificial language learning experiment based on Culbertson et al. (2012) and Martin et al. (2019) to test modifier generalization with human subjects. In this experiment, participants are asked to view a series of short animations denoting an action (Verb) performed by an agent shape (Subject) on a patient shape (Object), and simultaneously are presented with text descriptions that consist of three nonce words corresponding to the Verb, Subject, and Object in the scene (in the left figure of Figure 6-1a, Verb is 'zog', Subject is 'slov', and Object is 'VAB'). Shapes appear in two visually different versions: the default and the modified; the default is an unfilled shape and the modified is a blue shape (Figure 6-1). The default and the modified versions are distinguished in the description by a modification marking strategy (capitalization: 'vab' denotes 'star' and 'VAB' denotes 'blue star').

In the pilot, VSO word order was used for the description—an order different from English, the native language of the participants. Capitalization, which is a string manipulation method that does not involve appending a word to a noun, was chosen as the modification strategy. This choice was to block potential transfer from English, namely that blueness is expressed by a prenominal adjective, and that this is possible for both subjects and objects.[2] 12 animation-description pairs were used for training and 12 for testing. One shape (circle: 'fim') was never shown in blue and was only shown in the object position during training. All other shapes appeared as

[1] The human subject version of the experiment is production rather than comprehension (which would be the closer analog of the ANN experiment). However, I believe the production version of the experiment to be more challenging than comprehension.

[2] One may argue that this modification strategy is too far removed from anything that would be found in natural language—I address this concern in Section 6.1.2 in the discussion of planned future experiments.

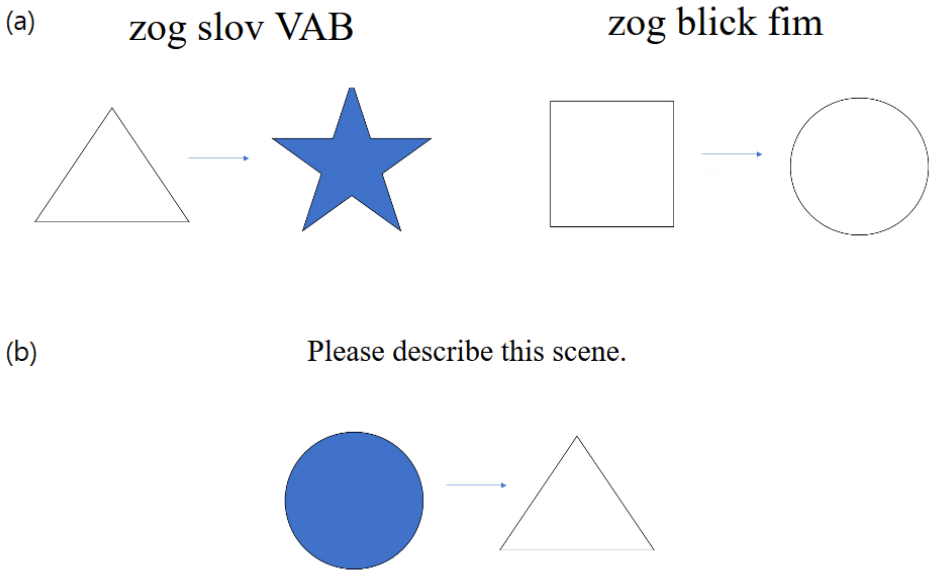


Figure 6-1. (a) Example stimuli from the training set: color modification optionally appears on the grammatical object (the shape being hit). (b) Example target stimuli from the test set: color modification appears on the grammatical subject (the shape hitting another shape). Circles do appear in the training set but only as a grammatical object, and never co-occurs with modification (i.e., blue circle is unseen).

both subject and object. However, not all combinations of shape, grammatical role and color were shown during training; 12 examples were randomly sampled from all possible training examples ($n = 36$). The test set included 3 types of examples: seen, unseen but in-distribution (unseen examples but sampled from the same distribution as the training examples), and target (examples with unseen shape—blue circle—in subject position). To reduce the burden on memory, a pictorial dictionary was always shown to the participants (Figure 6-2). The expected total duration of the pilot was around 15 minutes.

Training phase: After viewing each animation for 7 seconds, the description disappeared, and the participants were asked to recall and type out what the description corresponding to the animation was. If their answer was incorrect, they were given corrective feedback and were shown the correct description until their typed answer

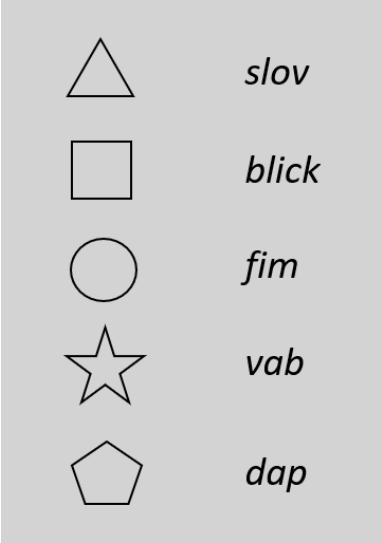


Figure 6-2. The pictorial dictionary shown to the participants in the pilot study.

exactly matched the description. During this training phase, participants were only exposed to animations with modification on the Object. Moreover, we restricted the distribution of a specific entity (e.g., the circle in Figure 6-1), by not letting this entity appear as a modified version or in the subject position.

Testing phase: Participants were asked to describe the animations based on the descriptions that they have learned during the training phase. The testing phase included animations in which the modified version of the restricted entity appeared as the subject (e.g., blue circle in Figure 6-1b). The participants had to employ generalization in order to describe such animations, because they would not have seen the modified version of the entity before, and would have only seen this entity appear as grammatical object.

Interpreting participant responses: There exist several plausible generalization strategies. For example, we could hypothesize: (1) frequency-based, (2) similarity-based (or shape-based), and (3) compositional generalization strategies.

1. Frequency-based: Since a blue circle is an entity that the participants never saw before, learners could opt for the most frequent lexical item of the same

grammatical category (i.e., the most frequent noun) that they observed during training. This was 'vab' in the pilot, so 'zog vab slov' for Figure 6-1b would correspond to this generalization.

2. Similarity-based: Since an unfilled circle is the most similar in appearance to a blue circle, learners could opt for the lexical item that denotes an unfilled circle ('zog fim slov').

3. Compositional: The learners might realize that blueness and modification co-occur, and compose the modification strategy with the word that denotes the unfilled circle. Furthermore, in order for the learners to choose the compositional strategy, they must be generalizing modification on the object to the subject, since they have never seen any modified subjects during training. This is the expected generalization, and 'zog FIM slov' would be the corresponding answer.

Results: 23 participants were recruited through Amazon Mechanical Turk. All participants were native speakers of English. An overwhelming majority of the participants (22/23, 96%) opted for the compositional generalization strategy (using 'FIM', a word form that they have never encountered during training), transferring a novel modification strategy only seen in object position to subject position. 1 participant used similarity-based generalization, consistently using the uncapitalized word 'fim' to describe examples like Figure 6-1b.[3]

6.1.2 Planned Experiments

As follow-up experiments, I plan to conduct variations of the pilot experiment in terms of modification strategy and the semantics of modification. Additionally, I will conduct a control experiment in which generalization is not expected, thus showing

[3]It was clear that this was not due to a technical issue (e.g., not being able to use capitalization while typing) or general unwillingness to capitalize anything, because this participant correctly used capitalization during the training phase.

that generalization of nominal markers is constrained by *consistent form-meaning mapping*.

Figure 6-3. Illustration of different modification strategies.

Different modification strategies: While capitalization was chosen as the modification strategy in the pilot to prevent transfer from English as much as possible, this kind of purely orthographic manipulation is impossible to be a nominal modification strategy that appears in natural language (unless the orthographic manipulation is a way of representing strategies such as stress). To address this concern, different modifier form and complexity will be tested to ensure that similar generalization patterns are observed for more plausible modification strategies. Specifically, I will test: (1) affixation (*vab taka* for 'blue star' (*taka*=blue); Figure 6-3a) and (2) full postpositional phrase modification (*puku ko vab* for 'star on the shelf' (*puku*=shelf, *ko*=on); Figure 6-3b).

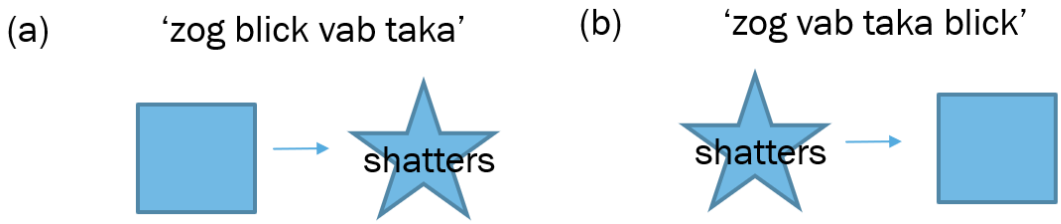


Figure 6-4. Illustration of nominal modification with resultative semantics.

Semantics of modification not available in English: The two modification strategies that I proposed to test both make use of semantics of nominal modification that is equally available in English (color and location of the object modified). I plan

to add a test case where the modification contributes a meaning that is not available in English through modifying a noun phrase. This test case will be such that the modifier describes the modified shape's state as a result of the action, similar to the semantics of a resultative construction in English such as *Lisa painted the wall green*. However, this meaning will be marked on the nominal unlike the English resultative. Using the same capitalization strategy as the pilot, and English words instead of nonce words for exposition, (1-a) would mean 'The square hit the star, and as a result of the hitting event, the star (the hittee) changed state', and (1-b) would mean 'The star hit the square, and as a result of the hitting event, the star (the hitter) changed state'. A nonce version of these sentences, with lexical modification is shown in (2) and Figure 6-4. The resulting state will be the modified object shattering into pieces, which is a plausible consequence of a hitting event that can happen to both the hitter and the hittee, and also a consequence that can easily be shown visually.

(1) a. hit square STAR 'The square hit the star, and the star shattered'
b. hit STAR square 'The star hit the square, and the star shattered'

(2) a. zog blick vab taka 'The square hit the star, and the star shattered'
b. zog vab taka blick 'The star hit the square, and the star shattered'

The pilot study used 'blue' as the content of modification, which introduces a potential transfer from English, since blueness is also expressed as a nominal modifier (adjective) in English and appears with both subject and object NPs. Using the resultative semantics has the effect of removing this potential confound, because similar semantics in English cannot be expressed as a nominal modifier. In fact, this not only blocks transfer but is actively adversarial in this respect. Therefore, if participants do generalize object modification to subject in this resultative semantics scenario, it will be stronger evidence towards compositional generalization.

Control experiment (no generalization case): A control experiment will also be conducted in order to verify that transfer of marking strategy across nominals is not licensed unrestrictedly. In this experiment, the same marker as in the main experiment will be used to mark a grammatical relation instead of a property inherent to the noun's referent (like color). For example, modification strategy (capitalization: (3) or affixation: (4)) will occur uniformly with the last word of the sentence, similarly to an object marker. In this case, we do not expect participants to transfer the marker to the subject position, even though this generalization does not contradict any observations. The conclusion we expect is that human learners would transfer object markers to subject markers only in the presence of semantic evidence that relates the marker with an semantic property of the referent, and restrict transfer when the marker denotes a grammatical property.

(3) a. hit square STAR 'The square hit the star'
b. hit star SQUARE 'The star hit the square'

(4) a. hit square star taka 'The square hit the star'
b. hit star square taka 'The star hit the square'

Grammaticality judgment version of the task: If there are heuristic strategies that can lead to the target response without compositional generalization, using grammaticality judgment as a task format rather than free-form answers will help tease apart heuristics from genuine structural generalization. While I could not identify an immediate heuristic for the capitalization experiment, if affixation is used as the modification strategy, a bag-of-words heuristic can (probabilistically) lead to the target response. This heuristic is: 'whenever there is a blue shape in the scene, add the affix to a random word in the description'. Under this heuristic, there is a $1/n$ chance of getting the target response that is equivalent to the compositional response (n=number of

words in the sentence containing no modification). A grammaticality judgment version of the task can distinguish a true case of compositional generalization and this heuristic—if participants are employing a bag-of-words heuristic, any sentence containing the affix will be judged grammatical regardless of the position. If participants are generalizing compositionally, only the sentence with the affix attached to the subject nominal will be judged grammatical.

6.2 Future Work: Promising Directions for Achieving Structural Generalization in ANNs

Going back to the experiments in the dissertation, the results revealed a pressing challenge for current ANN models—their almost complete incapacity to assign correct meaning representations to novel structures.[4] What could be done to improve structural generalization in ANN models? One possibility that I have already discussed in Chapter 4 is leveraging the approach of Chen et al. (2020b) that adds to an ANN an explicit mechanism of symbolic manipulation that supports recursion, as discussed in Section 4.6. Below, I discuss several additional ideas (some repeated from Section 2.3.1) that may contribute helpful learning biases towards structural generalization.

Models with invariance property: Gogate and Hollich (2010) have proposed a hypothesis that a general perceptual mechanism of detecting invariance may also underlie sensitivity to linguistic regularities. In this regard, models with invariance bias such as a convolutional neural networks may help models to recognize the structural regularities that are crucial for structural generalization (e.g., phrasal constitutency). Prior studies reported that the convolutional encoder-decoder model of Gehring et al. (2017) showed better performance on the SCAN dataset (Dessì and Baroni 2019)

[4]While the exact capability of human structural generalization is to be quantified through the experiments proposed in Section 6.1, we believe our pilot results, results of McCoy et al. (2021) on unseen center embedding structures and results of Morgan and Ferreira (2021) on unseen types of relative clauses all point towards the possibility of structural generalization in human learners.

and several synthetic generalization datasets generated from PCFGs (Hupkes et al. 2020) compared to recurrent neural networks, suggesting that invariance bias is indeed helpful for compositional generalization.

Imposing structural constraints: Looking at model architectures with explicit structural constraints, such as tree-structured recurrent neural networks (Tai et al. 2015) or tree-structured Transformers (Nguyen et al. 2020; Harer et al. 2019; Wang et al. 2019c; Shiv and Quirk 2019) may be a promising direction, considering the results of Bowman et al. (2015) and McCoy et al. (2019) that tree-structured architectures improve compositional generalization. In particular, Bowman et al. (2015) have shown that tree-structured LSTMs generalize to sequences longer than what was observed in the training data, suggesting that explicit structure in the model architecture can help generate novel structures. We could furthermore consider models that use internal representations (rather than architectures) with an explicitly compositional structure, such as the Tensor Product Transformer of Schlag et al. (2019) (which uses the Hadamard product of filler and role representations as a simplification of Tensor Product Representations Smolensky 1990). These representations have been shown to facilitate compositional generalization (and interpretability of internal representations) in mathematics problem solving (Chen et al. 2020a). Approaches that augment pretrained models like BERT with explicit syntactic parses is another promising direction. In particular, Sachan et al. (2020) and Zanzotto et al. (2020) provide ways to inject representations of tree structures to the BERT model.

Bibliography

Kenneth Aizawa. Exhibiting versus explaining systematicity: a reply to Hadley and Hayward. *Minds and Machines*, 7(1):39–55, 1997.

Ekin Akyürek and Jacob Andreas. Lexicon learning for few-shot neural sequence modeling. *arXiv:2106.03993*, 2021.

Afra Alishahi, Grzegorz Chrupała, and Tal Linzen. Analyzing and interpreting neural networks for NLP: A report on the first BlackboxNLP workshop. *Natural Language Engineering*, 25(4):543–557, 2019.

Richard N. Aslin and Elissa L. Newport. Distributional language learning: Mechanisms and models of category formation. *Language Learning*, 64(s2):86–105, 2014. doi: https://doi.org/10.1111/lang.12074. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/lang.12074.

Carl L. Baker. Syntactic theory and the projection problem. *Linguistic Inquiry*, 10(4): 533–581, 1979.

Mark C. Baker. *Lexical categories: Verbs, nouns and adjectives*. Cambridge University Press, Cambridge, 2003.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2322.

Srinivas Bangalore and Aravind Joshi. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265, 1999.

Marco Baroni. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791):20190307, 2020.

Maija S. Blaubergs and Martin D. Braine. Short-term memory limitations on decoding self-embedded sentences. *Journal of Experimental Psychology*, 102(4):745–748, 1974.

Paul A. Bloom and Ira Fischler. Completion norms for 329 sentence contexts. *Memory & Cognition*, 8(6):631–642, 1980.

Mikael Bodén. Generalization by symbolic abstraction in cascaded recurrent networks. *Neurocomputing*, 57:87–104, 2004.

Bernd Bohnet, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez. Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2642–2652, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1246. URL https://www.aclweb.org/anthology/P18-1246.

Arielle Borovsky, Marta Kutas, and Jeffrey L. Elman. Learning to use words: Event-related potentials index single-shot contextual word learning. *Cognition*, 116(2):289–296, 2010.

Arielle Borovsky, Jeffrey L. Elman, and Marta Kutas. Once is enough: N400 indexes semantic integration of novel word meanings from a single exposure in context. *Language Learning and Development*, 8(3):278–302, 2012.

Johan Bos, Valerio Basile, Kilian Evang, Noortje J. Venhuizen, and Johannes Bjerva. The groningen meaning bank. In *Handbook of Linguistic Annotation*, pages 463–496. Springer, 2017.

Melissa Bowerman. Evaluating competing linguistic models with language acquisition data: Implications of developmental errors with causative verbs. *Quaderni di semantica*, 3:5–66, 1982.

Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. Recursive neural networks can learn logical semantics. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 12–21, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4002. URL https://www.aclweb.org/anthology/W15-4002.

Martin D. S. Braine. On learning the grammatical order of words. *Psychological Review*, 70 (4):323–348, 1963.

Joan Bresnan and Marilyn Ford. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language*, 86(1):168–213, 2010.

Patricia J. Brooks and Michael Tomasello. Young children learn to produce passives with nonce verbs. *Developmental Psychology*, 35(1):29–44, 1999.

Patricia J. Brooks, Michael Tomasello, Kelly Dodson, and Lawrence B. Lewis. Young children's overgeneralizations with fixed transitivity verbs. *Child Development*, 70(6): 1325–1337, 1999.

Timothy A. Cartwright and Michael R. Brent. Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition*, 63(2):121–170, 1997.

David J.s Chalmers. Connectionism and compositionality: Why fodor and pylyshyn were wrong. *Philosophical Psychology*, 6(3):305–319, 1993. doi: 10.1080/09515089308573094. URL https://doi.org/10.1080/09515089308573094.

Kezhen Chen, Qiuyuan Huang, Hamid Palangi, Paul Smolensky, Ken Forbus, and Jianfeng Gao. Mapping natural-language problems to formal-language solutions using structured neural representations. In *International Conference on Machine Learning*, pages 1566–1575. PMLR, 2020a.

Xinyun Chen, Chen Liang, Adams Wei Yu, Dawn Song, and Denny Zhou. Compositional generalization via neural-symbolic stack machines. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1690–1701. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper/2020/file/12b1e42dc0746f22cf361267de07073f-Paper.pdf.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://www.aclweb.org/anthology/D14-1179.

Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA, 1965.

Morten H. Christiansen. The (non) necessity of recursion in natural language processing. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, pages 665–670, 1992.

Morten H. Christiansen and Nick Chater. Generalization and connectionist language learning. *Mind & Language*, 9(3):273–287, 1994. doi: https://doi.org/10.1111/j.1468-0017.1994.tb00226.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0017.1994.tb00226.x.

Morten H. Christiansen and Maryellen C. MacDonald. A usage-based approach to recursion in sentence processing. *Language Learning*, 59(s1):126–161, 2009. doi: https://doi.org/10.1111/j.1467-9922.2009.00538.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9922.2009.00538.x.

Guglielmo Cinque. *Adverbs and functional heads: A cross-linguistic perspective*. Oxford University Press, Oxford, 1999.

Erin Conwell and Katherine Demuth. Early syntactic productivity: Evidence from dative shift. *Cognition*, 103(2):163–179, 2007.

Leyang Cui and Yue Zhang. Hierarchically-refined label attention network for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4115–4128, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1422. URL https://aclanthology.org/D19-1422.

Jennifer Culbertson, Paul Smolensky, and Géraldine Legendre. Learning biases predict a word order universal. *Cognition*, 122(3):306–329, 2012.

Deborah A. Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. Expanding the scope of the ATIS task: The ATIS-3 corpus. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. URL https://aclanthology.org/H94-1010.

Donald Davidson. The logical form of action sentences. In Nicholas Rescher, editor, *The Logic of Decision and Action*, pages 81–95. University of Pittsburgh Press, Pittsburgh, 1967.

Roberto Dessì and Marco Baroni. CNNs found to jump around more skillfully than RNNs: Compositional generalization in seq2seq convolutional networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3919–3923, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1381. URL https://www.aclweb.org/anthology/P19-1381.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, and Matt Gardner. Documenting the English Colossal Clean Crawled Corpus. *arXiv:2104.08758*, 2021.

David Dowty. Thematic proto-roles and argument selection. *Language*, 67(3):547–619, 1991.

David Dowty. Compositionality as an empirical problem. In Chris Barker and Pauline Jacobson, editors, *Direct Compositionality*, volume 14, pages 23–101. Oxford Unviersity Press, Oxford, 2007.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

Jeffrey L. Elman. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7(2):195–225, 1991.

Allyson Ettinger. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 01 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00298. URL https://doi.org/10.1162/tacl_a_00298.

Larry Fenson, Virginia A. Marchman, Donna J. Thal, Phillip S. Dale, J. Steven Reznick, and Elizabeth Bates. *MacArthur-Bates communicative development inventories*. Paul H. Brookes Publishing Company, Baltimore, MD, 2007.

Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. Improving text-to-SQL evaluation methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 351–360, Melbourne, Australia,

July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1033. URL https://aclanthology.org/P18-1033.

Sara Finley and William Badecker. Artificial language learning and feature-based generalization. *Journal of Memory and Language*, 61(3):423–437, 2009.

Dan Flickinger. On building a more effcient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28, 2000. doi: 10.1017/S1351324900002370.

Jerry A. Fodor and Brian P. McLaughlin. Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, 35(2):183–204, 1990.

Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

Stefan L. Frank. Learn more by training less: systematicity in sentence processing by recurrent networks. *Connection Science*, 18(3):287–302, 2006.

Stefan L. Frank, Willem Haselager, and Iris van Rooij. Connectionist semantic systematicity. *Cognition*, 110(3):358–379, 2009.

Gottlob Frege. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50, 1892.

Gottlob Frege. Compound thoughts. *Mind*, 72(285):1–17, 1963. ISSN 00264423, 14602113. URL http://www.jstor.org/stable/2251920.

Daniel Furrer, Marc van Zee, Nathan Scales, and Nathanael Schärli. Compositional generalization in semantic parsing: Pre-training vs. specialized architectures. *arXiv:2007.08970*, 2020.

Kanishk Gandhi and Brenden M. Lake. Mutual exclusivity as a challenge for deep neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14182–14192. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/a378383b89e6719e15cd1aa45478627c-Paper.pdf.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR, 2017.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Susan A. Gelman and Marjorie Taylor. How two-year-old children interpret proper and common names for unfamiliar objects. *Child Development*, 55(4):1535–1540, 1984. ISSN 00093920, 14678624. URL http://www.jstor.org/stable/1130023.

Louann Gerken, Rachel Wilson, and William Lewis. Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, 32(2):249–268, 2005. doi: 10.1017/S0305000904006786.

Yael Gertner, Cynthia Fisher, and Julie Eisengart. Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science*, 17(8): 684–691, 2006.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL http://proceedings.mlr.press/v9/glorot10a.html.

Lakshmi J. Gogate and George Hollich. Invariance detection within an interactive system: A perceptual gateway to language development. *Psychological Review*, 117(2):496–516, 2010. URL https://doi.org/10.1037/a0019049.

Yoav Goldberg. Assessing BERT's syntactic abilities. *arXiv:1901.05287*, 2019.

Rebecca L. Gómez and LouAnn Gerken. Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, 4(5):178–186, 2000. URL https://www.sciencedirect.com/science/article/abs/pii/S1364661300014674.

Rebecca L. Gómez and Laura Lakusta. A first step in form-based category abstraction by 12-month-old infants. *Developmental Science*, 7(5):567–580, 2004.

Jonathan Gordon, David Lopez-Paz, Marco Baroni, and Diane Bouchacourt. Permutation equivariant models for compositional generalization in language. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylVNerFvr.

Jane Grimshaw. Form, function, and the language acquisition device. In Carl L. Baker and John J. McCarthy, editors, *The Logical Problem of Language Acquisition*, pages 165–182. MIT Press, Cambridge, MA, 1981.

Jeroen Groenendijk and Martin Stokhof. Dynamic predicate logic. *Linguistics and Philosophy*, 14(1):39–100, 1991. URL http://www.jstor.org/stable/25001418.

Jeroen Groenendijk and Martin Stokhof. Why compositionality? In Gregory N. Carlson and Francis J. Pelletier, editors, *Reference and Quantification: The Partee Effect*, pages 83–106. CSLI Press, 2004. URL https://pure.uva.nl/ws/files/2213352/27378_wc.pdf.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1154. URL https://aclanthology.org/P16-1154.

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. Pointing the unknown words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 140–149, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1014. URL https://aclanthology.org/P16-1014.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 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)*, pages 1195–1205, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1108. URL https://www.aclweb.org/anthology/N18-1108.

Jiaqi Guo, Qian Liu, Jian-Guang Lou, Zhenwen Li, Xueqing Liu, Tao Xie, and Ting Liu. Benchmarking meaning representations in neural semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1520–1540, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.118. URL https://aclanthology.org/2020.emnlp-main.118.

Robert F. Hadley. Systematicity in connectionist language learning. *Mind & Language*, 9 (3):247–272, 1994.

Robert F. Hadley. Cognition, systematicity and nomic necessity. *Mind & Language*, 12(2): 137–153, 1997.

Robert F. Hadley and Vlad C. Cardei. Language acquisition from sparse input without error feedback. *Neural Networks*, 12(2):217–235, 1999.

Robert F. Hadley, Adam Rotaru-Varga, Dirk V. Arnold, and Vlad C. Cardei. Syntactic systematicity arising from semantic predictions in a hebbian-competitive network. *Connection Science*, 13(1):73–94, 2001.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3153–3160, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/510_Paper.pdf.

Jacob Harer, Chris Reale, and Peter Chin. Tree-transformer: A Transformer-based method for correction of tree-structured data. *arXiv:1908.00449*, 2019.

Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579, 2002.

Han He and Jinho D. Choi. Establishing strong baselines for the new decade: Sequence tagging, syntactic and semantic parsing with BERT. *arXiv:1908.04943*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24–27*, 1990. URL https://aclanthology.org/H90-1021.

Jonathan Herzig, Peter Shaw, Ming-Wei Chang, Kelvin Guu, Panupong Pasupat, and Yuan Zhang. Unlocking compositional generalization in pre-trained models using intermediate representations. *arXiv:2104.07478*, 2021.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://www.aclweb.org/anthology/D19-1275.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Julia Hockenmaier and Mark Steedman. Ccgbank: a corpus of ccg derivations and dependency structures extracted from the penn treebank. *Computational Linguistics*, 33(3):355–396, 2007.

Barbara Höhle, Jürgen Weissenborn, Dorothea Kiefer, Antje Schulz, and Michaela Schmitz. Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy*, 5(3):341–353, 2004. URL https://onlinelibrary.wiley.com/doi/abs/10.1207/s15327078in0503_5.

Juan Hu, Nancy Budwig, Kaya Ono, and Hang Zhang. Individual differences in preschoolers' ability to generalize unaccusative intransitive constructions in novel verb experiments: Evidence from their familiar verb usage in naturalistic play contexts. In H. Caunt-Nulton, S. Kulatilake, and I. Woo, editors, *A Supplement to the Proceedings of the 31st Boston University Conference on Language*, 2007.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: how do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2017.

Stephen James, Andrew J. Davison, and Edward Johns. Transferring end-to-end visuomotor control from simulation to real world for a multi-stage task. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg, editors, *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 334–343. PMLR, 13–15 Nov 2017. URL https://proceedings.mlr.press/v78/james17a.html.

Theo M. V. Janssen. *Foundations and Applications of Montague Grammar, Part 1: Philosophy, Framework, Computer Science*. CWI Tract 19. Center for Mathematics and Computer Science, Amsterdam, 1986.

Theo M. V. Janssen. Compositionality. In J. van Benthem and A. ter Meulen, editors, *Handbook of Logic and Linguistics*, pages 417–473. Elsevier Science Publishers, Amsterdam, 1997.

Robin Jia and Percy Liang. Data recombination for neural semantic parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12–22, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1002. URL https://aclanthology.org/P16-1002.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Kent Johnson. On the systematicity of language and thought. *The Journal of Philosophy*, 101(3):111–139, 2004.

Peter W. Jusczyk and Richard N. Aslin. Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29(1):1–23, 1995.

Hans Kamp and Uwe Reyle. *From Discourse to Logic: Introduction to Model Theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers, Dordrecht, 1993.

David Kaplan. Demonstratives. In Joseph Almog, John Perry, and Howard Wettstein, editors, *Themes from Kaplan*, pages 481–563. Oxford University Press, 1989.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.

Fred Karlsson. Syntactic recursion and iteration. In Harry van der Hulst, editor, *Recursion and Human Language*, pages 43–67. De Gruyter Mouton, 2010.

Deborah G. Kemler Nelson, Peter W. Jusczyk, Denise R. Mandel, James Myers, Alice Turk, and LouAnn Gerken. The head-turn preference procedure for testing auditory perception. *Infant Behavior and Development*, 18(1):111–116, 1995.

Nenagh Kemp, Elena Lieven, and Michael Tomasello. Young children's knowledge of the "determiner" and "adjective" categories. *Journal of Speech, Language, and Hearing Research*, 48(3):592–609, 2005. doi: 10.1044/1092-4388(2005/041).

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SygcCnNKwr.

Eugene Kharitonov and Rahma Chaabouni. What they do when in doubt: a study of inductive biases in seq2seq learners. *arXiv:2006.14953*, 2020.

Najoung Kim and Tal Linzen. Compositionality as directional consistency in sequential neural networks. *NeurIPS Workshop on Context and Compositionality in Biological and Artificial Neural Systems*, 2019.

Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.731. URL https://www.aclweb.org/anthology/2020.emnlp-main.731.

Najoung Kim and Paul Smolensky. Testing for grammatical category abstraction in neural language models. *Proceedings of the Society for Computation in Linguistics*, 4(1):467–470, 2021.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/S19-1026. URL https://www.aclweb.org/anthology/S19-1026.

Najoung Kim, Kyle Rawlins, Benjamin Van Durme, and Paul Smolensky. Predicting the argumenthood of English prepositional phrases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6578–6585, 2019b.

Karin Kipper-Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, 2005. URL http://verbs.colorado.edu/~kipper/Papers/dissertation.pdf.

Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan, July 2003. Association for Computational Linguistics. doi: 10.3115/1075096.1075150. URL https://www.aclweb.org/anthology/P03-1054.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P17-4012.

Melissa Kline and Katherine Demuth. Syntactic generalization with novel intransitive verbs. *Journal of Child Language*, 41(3):543–574, 2014.

Angelika Kratzer. Stage-level and individual-level predicates. In Greg Carlson and Jeff Pelletier, editors, *The Generic Book*, pages 125–175. Chicago University Press, Chicago, 1995.

Angelika Kratzer. Severing the external argument from its verb. In Johan Rooryck and Laurie Zaring, editors, *Phrase Structure and the Lexicon*, pages 109–137. Springer Netherlands, Dordrecht, 1996. ISBN 978-94-015-8617-7. doi: 10.1007/978-94-015-8617-7_5. URL https://doi.org/10.1007/978-94-015-8617-7_5.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://www.aclweb.org/anthology/D18-2012.

Brenden M. Lake. Compositional generalization through meta sequence-to-sequence learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/f4d0e2e7fc057a58f7ca4a391f01940a-Paper.pdf.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2873–2882. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/lake18a.html.

Brenden M. Lake, Tal Linzen, and Marco Baroni. Human few-shot learning of compositional instructions. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 611–617, 2019.

Geoffrey Leech, Paul Rayson, and Andrew Wilson. *Word frequencies in written and spoken English: based on the British National Corpus*. Longman, 2001. URL http://ucrel.lancs.ac.uk/bncfreq/flists.html.

Géraldine Legendre, Yoshiro Miyata, and Paul Smolensky. Harmonic Grammar—a formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 884–891, 1990.

Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press, 1993.

Yuanpeng Li, Liang Zhao, Jianyu Wang, and Joel Hestness. Compositional generalization for primitive substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4293–4302, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1438. URL https://www.aclweb.org/anthology/D19-1438.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

João Loula, Marco Baroni, and Brenden Lake. Rearranging the familiar: Testing compositional generalization in recurrent networks. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 108–114, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5413. URL https://www.aclweb.org/anthology/W18-5413.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1166. URL https://www.aclweb.org/anthology/D15-1166.

Brian MacWhinney. *The CHILDES project: Computational tools for analyzing talk*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1995a.

Brian MacWhinney. *The CHILDES project: Computational tools for analyzing talk. The database, Vol. 2*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1995b.

Brian MacWhinney and Catherine Snow. The child language data exchange system. *Journal of Child Language*, 12(2):271–295, 1985. doi: 10.1017/S0305000900006449.

Ofra Magidor. Category mistakes are meaningful. *Linguistics and Philosophy*, 32(6):553–581, 2009.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2): 313–330, 1993. URL https://www.aclweb.org/anthology/J93-2004.

David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman, San Francisco, 1982. ISBN 0716715678.

Alexander Martin, Klaus Abels, David Adger, and Jennifer Culbertson. Do learners' word order preferences reflect hierarchical language structure? In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 2303–2309, 2019.

Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1151. URL https://www.aclweb.org/anthology/D18-1151.

John J. McCarthy and Alan Prince. Faithfulness and reduplicative identity. *Papers in Optimality Theory*, 10, 1995. URL https://scholarworks.umass.edu/linguist_faculty_pubs/10.

Michael McCloskey. Networks and theories: The place of connectionism in cognitive science. *Psychological Science*, 2(6):387–395, 1991.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://www.aclweb.org/anthology/P19-1334.

Tom McCoy, Jennifer Culbertson, Paul Smolensky, and Géraldine Legendre. Infinite use of finite means? evaluating the generalization of center embedding learned from an artificial grammar, May 2021. URL psyarxiv.com/r8ct2.

Brian P. McLaughlin. The connectionism/classicism battle to win souls. *Philosophical Studies*, 71(2):163–190, 1993.

Katherine Messenger and Cynthia Fisher. Mistakes weren't made: Three-year-olds' comprehension of novel-verb passives provides evidence for early abstract syntax. *Cognition*, 178: 118–132, 2018.

Elliot Meyerson and Risto Miikkulainen. Pseudo-task augmentation: From deep multitask learning to intratask sharing—and back. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3511–3520. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/meyerson18a.html.

Stephan C. Meylan, Michael C. Frank, Brandon C. Roy, and Roger Levy. The emergence of an abstract grammatical category in children's early speech. *Psychological Science*, 28 (2):181–192, 2017. doi: 10.1177/0956797616677753. URL https://doi.org/10.1177/0956797616677753. PMID: 28074675.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1090.

Toben H. Mintz. Category induction from distributional cues in an artificial language. *Memory & Cognition*, 30(5):678–686, 2002.

Toben H. Mintz. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91–117, 2003.

Toben H. Mintz, Elissa L. Newport, and Thomas G. Bever. The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, 26(4):393–424, 2002.

Tom M. Mitchell. The need for biases in learning generalizations. In Jude W. Shavlik and Thomas G. Dietterich, editors, *Readings in Machine Learning*, pages 184–191. Morgan Kauffman, 1980. Book published in 1990.

Yusuke. Miyao. *From linguistic theory to syntactic analysis: Corpus-oriented grammar development and feature forest model*. PhD thesis, University of Tokyo, 2006. URL https://ci.nii.ac.jp/naid/10030678297/en/.

Padraic Monaghan, Nick Chater, and Morten H. Christiansen. The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, 96(2): 143–182, 2005.

Richard Montague. Universal grammar. *Theoria*, 36(3):373–398, 1970. doi: 10.1111/j.1755-2567.1970.tb00434.x.

Adam M. Morgan and Victor S. Ferreira. Beyond input: Language learners produce novel relative clause types without exposure. *Journal of Cognitive Psychology*, 0(0):1–35, 2021. doi: 10.1080/20445911.2021.1928678. URL https://doi.org/10.1080/20445911.2021.1928678.

James L. Morgan, Richard P. Meier, and Elissa L. Newport. Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive psychology*, 19(4):498–550, 1987.

Xuan-Phi Nguyen, Shafiq Joty, Steven Hoi, and Richard Socher. Tree-structured attention with hierarchical accumulation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJxK5pEYvr.

Lars F. Niklasson and Tim Van Gelder. On being systematically connectionist. *Mind & Language*, 9(3):288–302, 1994.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajič, and Zdeňka Urešová. SemEval 2015 task 18: Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 915–926, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-2153. URL https://aclanthology.org/S15-2153.

Raquel Olguin and Michael Tomasello. Twenty-five-month-old children do not have a grammatical category of verb. *Cognitive Development*, 8(3):245–272, 1993.

Kaya Ono and Nancy Budwig. Young children's use of unaccusative intransitives in novel verb experiments. In *A Supplement to the Proceedings of the 30th Boston University Conference on Language*, 2006.

Peter Pagin. Communication and the complexity of semantics. In Markus Werning, Wolfram Hinzen, and Edouard Machery, editors, *The Oxford Handbook of Compositionality*, pages 510–529. Oxford University Press, 2012.

Peter Pagin. Compositionality, computability, and complexity. *The Review of Symbolic Logic*, page 1–52, 2020. doi: 10.1017/S1755020320000027.

Peter Pagin and Dag Westerståhl. Compositionality II: Arguments and problems. *Philosophy Compass*, 5(3):265–282, 2010.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, 03 2005. ISSN 0891-2017. doi: 10.1162/0891201053630264. URL https://doi.org/10.1162/0891201053630264.

Terence Parsons. *Events in the Semantics of English*. MIT Press, Cambridge, MA, 1990.

Barbara H. Partee. Compositionality. In F. Landman and F. Veltman, editors, *Varieties of Formal Semantics*, pages 281–311. Dordrecht: Foris, 1984.

Barbara H. Partee. Lexical semantics and compositionality. In Lila Gleitman and Mark Lieberman, editors, *An Invitation to Cognitive Science, Part I: Language*, volume 1, pages 311–360. MIT Press, 1995.

John Payne, Rodney Huddleston, and Geoffrey K. Pullum. The distribution and category status of adjectives and adverbs. *Word Structure*, 3(1):31–81, 2010.

Francis Jeffry Pelletier. The principle of semantic compositionality. *Topoi*, 13(1):11–24, 1994.

Andrew Perfors, Joshua B. Tenenbaum, and Terry Regier. The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338, 2011.

Steven T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130, 2014.

Steven Pinker. *Language Learnability and Language Development (1984/1996)*. Harvard University Press, Cambridge, MA, 1984.

Alan S. Prince and Paul Smolensky. Optimality Theory: Constraint interaction in generative grammar. *Rutgers Optimality Archive*, 2002. URL http://roa.rutgers.edu/article/view/547.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.467. URL https://www.aclweb.org/anthology/2020.acl-main.467.

Geoffrey K. Pullum and Barbara C. Scholz. Systematicity and natural language syntax. *Croatian Journal of Philosophy*, 7(21):375–402, 2007.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

Siva Reddy, Oscar Täckström, Michael Collins, Tom Kwiatkowski, Dipanjan Das, Mark Steedman, and Mirella Lapata. Transforming Dependency Structures to Logical Forms for Semantic Parsing. *Transactions of the Association for Computational Linguistics*, 4: 127–140, 04 2016. ISSN 2307-387X. doi: 10.1162/tacl_a_00088. URL https://doi.org/10.1162/tacl_a_00088.

Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. Universal semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 89–101, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1009. URL https://www.aclweb.org/anthology/D17-1009.

Martin Redington, Nick Chater, and Steven Finch. Distributional information and the acquisition of linguistic categories: A statistical approach. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, pages 848–853, 1993.

Martin Redington, Nick Chater, and Steven Finch. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4):425–469, 1998.

Patricia A. Reeder, Elissa L. Newport, and Richard N. Aslin. From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, 66(1):30–54, 2013.

Patricia A. Reeder, Elissa L. Newport, and Richard N. Aslin. Distributional learning of subcategories in an artificial grammar: Category generalization and subcategory restrictions. *Journal of Memory and Language*, 97:17–29, 2017.

Peter A. Reich. The finiteness of natural language. *Language*, 45(4):831–843, 1969. ISSN 00978507, 15350665. URL http://www.jstor.org/stable/412337.

Douglas Roland, Frederic Dick, and Jeffrey L. Elman. Frequency of basic english grammatical structures: A corpus analysis. *Journal of Memory and Language*, 57(3):348–379, 2007.

Caroline F. Rowland and Claire L. Noble. The role of syntactic structure in children's sentence comprehension: Evidence from the dative. *Language Learning and Development*, 7(1):55–75, 2010.

Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M. Lake. A benchmark for systematic generalization in grounded language understanding. *Advances in Neural Information Processing Systems*, 33, 2020.

Devendra Singh Sachan, Yuhao Zhang, Peng Qi, and William Hamilton. Do syntax trees help pre-trained transformers extract information? *arXiv:2008.09084*, 2020.

Dionysia Saratsli, Stefan Bartell, and Anna Papafragou. Cross-linguistic frequency and the learnability of semantics: Artificial language learning studies of evidentiality. *Cognition*, 197:104194, 2020. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2020.104194. URL https://www.sciencedirect.com/science/article/pii/S0010027720300135.

Imanol Schlag, Paul Smolensky, Roland Fernandez, Nebojsa Jojic, Jürgen Schmidhuber, and Jianfeng Gao. Enhancing the transformer with explicit relational encoding for math problem solving. *NeurIPS Workshop on Context and Compositionality in Biological and Artificial Neural Systems*, 2019.

Rose M. Scott and Cynthia Fisher. Two-year-olds use distributional cues to interpret transitivity-alternating verbs. *Language and cognitive processes*, 24(6):777–803, 2009.

Vighnesh Shiv and Chris Quirk. Novel positional encodings to enable tree-based transformers. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/6e0917469214d8fbd8c517dcdc6b8dcf-Paper.pdf.

Amy Skipp, Kirsten L. Windfuhr, and Gina Conti-Ramsden. Children's grammatical categories of verb and noun: a comparative look at children with specific language impairment (sli) and normal language (nl). *International Journal of Language & Communication Disorders*, 37(3):253–271, 2002. doi: 10.1080/13682820110119214.

URL https://www.tandfonline.com/doi/abs/10.1080/13682820110119214. PMID: 12201977.

Paul Smolensky. On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11(1):1–23, 1988.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1):159–216, 1990. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(90)90007-M. URL https://www.sciencedirect.com/science/article/pii/000437029090007M.

Paul Smolensky. The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. In *Connectionism and the Philosophy of Mind*, pages 281–308. Springer, 1991.

Paul Smolensky. Constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture. In C. Macdonald and G. Macdonald, editors, *Connectionism: Debates on Psychological Explanation*, volume 2, pages 221–290. Blackwell, 1995.

Paul Smolensky. Computational levels and integrated connectionist/symbolic explanation. In Paul Smolensky and Géraldine Legendre, editors, *The Harmonic mind: From Neural Computation to Optimality-Theoretic Grammar, Volume II: Linguistic and Philosophical Implications*, pages 503–592. MIT Press, 2006.

Paul Soulos, R. Thomas McCoy, Tal Linzen, and Paul Smolensky. Discovering the compositional structure of vector representations with role learning networks. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 238–254, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.23. URL https://www.aclweb.org/anthology/2020.blackboxnlp-1.23.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv:1904.01172*, 2019.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1548. URL https://www.aclweb.org/anthology/D18-1548.

Dhanasekar Sundararaman, Vivek Subramanian, Guoyin Wang, Shijing Si, Dinghan Shen, Dong Wang, and Lawrence Carin. Syntax-infused Transformer and BERT models for machine translation and natural language understanding. *arXiv:1911.06156*, 2019.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*,

pages 1556–1566, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1150. URL https://www.aclweb.org/anthology/P15-1150.

Yi Tay, Mostafa Dehghani, Jai Gupta, Dara Bahri, Vamsi Aribandi, Zhen Qin, and Donald Metzler. Are pre-trained convolutions better than pre-trained transformers? *arXiv:2105.03322*, 2021.

Wilson L. Taylor. "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433, 1953.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJzSgnRcKX.

Michael Tomasello and Raquel Olguin. Twenty-three-month-old children have a grammatical category of noun. *Cognitive Development*, 8(4):451–464, 1993.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 14954–14964. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/749a8e6c231831ef7756db230b4359c8-Paper.pdf.

Johan Van Benthem. The logic of semantics. In *Essays in Logical Semantics*, pages 198–214. Springer Netherlands, Dordrecht, 1986. ISBN 978-94-009-4540-1. doi: 10.1007/978-94-009-4540-1_10. URL https://doi.org/10.1007/978-94-009-4540-1_10.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5831–5837, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1592. URL https://www.aclweb.org/anthology/D19-1592.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

David Vilares, Michalina Strzyz, Anders Søgaard, and Carlos Gómez-Rodríguez. Parsing as pretraining. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 9114–9121, 2020.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.14. URL https://www.aclweb.org/anthology/2020.emnlp-main.14.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1439. URL https://www.aclweb.org/anthology/P19-1439.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

Yaushian Wang, Hung-Yi Lee, and Yun-Nung Chen. Tree transformer: Integrating tree structures into self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1061–1070, Hong Kong, China, November 2019c. Association for Computational Linguistics. doi: 10.18653/v1/D19-1098. URL https://www.aclweb.org/anthology/D19-1098.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 09 2019. ISSN 2307-387X. doi: 10.1162/tacl_a_00290. URL https://doi.org/10.1162/tacl_a_00290.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. *OntoNotes Release 5.0, LDC2013T19*. Linguistic Data Consortium, Philadelphia, PA, 2013. URL https://catalog.ldc.upenn.edu/LDC2013T19.

Markus Werning. Right and wrong reasons for compositionality. In Markus Werning, Edouard Machery, and Gerhard Schurz, editors, *The Compositionality of Meaning and Content. Volume I: Foundational Issues*, pages 285–310. De Gruyter, 2005. doi: doi: 10.1515/9783110323627.285. URL https://doi.org/10.1515/9783110323627.285.

Dag Westerståhl. On mathematical proofs of the vacuity of compositionality. *Linguistics and Philosophy*, 21(6):635–643, 1998. URL http://www.jstor.org/stable/25001726.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://www.aclweb.org/anthology/N18-1101.

Colin Wilson. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, 30(5):945–982, 2006. URL https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog0000_89.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Lynsey Wolter. Situation variables and licensing by modification in opaque demonstratives. *Proceedings of Sinn und Bedeutung*, 11:612–626, Aug. 2019. URL https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/668.

Elizabeth Wonnacott, Elissa L. Newport, and Michael K. Tanenhaus. Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, 56(3):165–209, 2008.

Zenan Xu, Daya Guo, Duyu Tang, Qinliang Su, Linjun Shou, Ming Gong, Wanjun Zhong, Xiaojun Quan, Nan Duan, and Daxin Jiang. Syntax-enhanced pre-trained model. *arXiv:2012.14116*, 2020.

Hitomi Yanaka, Koji Mineshima, and Kentaro Inui. SyGNS: A systematic generalization testbed based on natural language semantics. *arXiv:2106.01077*, 2021.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1425. URL https://aclanthology.org/D18-1425.

Wlodek Zadrozny. From compositional to systematic semantics. *Linguistics and philosophy*, 17(4):329–342, 1994.

Fabio Massimo Zanzotto, Andrea Santilli, Leonardo Ranaldi, Dario Onorati, Pierfrancesco Tommasino, and Francesca Fallucchi. KERMIT: Complementing transformer architectures with encoders of explicit syntactic interpretations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 256–267, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.18. URL https://www.aclweb.org/anthology/2020.emnlp-main.18.

John M. Zelle and Raymond J. Mooney. Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, volume 2, page 1050–1055. AAAI Press, 1996. ISBN 026251091X.

Luke S. Zettlemoyer and Michael Collins. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, page 658–666, Arlington, Virginia, USA, 2005. AUAI Press. ISBN 0974903914.

Kelly Zhang and Samuel Bowman. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5448. URL https://aclanthology.org/W18-5448.

Linfeng Zhang, Muzhou Yu, Tong Chen, Zuoqiang Shi, Chenglong Bao, and Kaisheng Ma. Auxiliary training: Towards accurate and robust models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 372–381, June 2020.

Qilin Zhang, Gang Hua, Wei Liu, Zicheng Liu, and Zhengyou Zhang. Can visual recognition benefit from auxiliary information in training? In *Asian Conference on Computer Vision*, pages 65–80. Springer, 2014.

Sheng Zhang, Kevin Duh, and Benjamin Van Durme. Cross-lingual semantic parsing. *arXiv:1804.08037*, 2018.

Victor Zhong, Caiming Xiong, and Richard Socher. Seq2SQL: Generating structured queries from natural language using reinforcement learning. *arXiv:1709.00103*, 2017.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 19–27, December 2015.