Structural Supervision Improves Few-Shot Learning and Syntactic Generalization in Neural Language Models

Ethan Wilcox[1], Peng Qian[2], Richard Futrell[3], Ryosuke Kohita[4], Roger Levy[2] and Miguel Ballesteros[5]
[1]Department of Linguistics, Harvard University
[2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[3]Department of Language Science, UC Irvine [4]IBM Research [5]Amazon AI
`wilcoxeg@g.harvard.edu`, {`pqian,rplevy`}`@mit.edu`
`rfutrell@uci.edu, kohi@ibm.com, ballemig@amazon.com`

Abstract

Humans can learn structural properties about a word from minimal experience, and deploy their learned syntactic representations uniformly in different grammatical contexts. We assess the ability of modern neural language models to reproduce this behavior in English and evaluate the effect of structural supervision on learning outcomes. First, we assess few-shot learning capabilities by developing controlled experiments that probe models' syntactic nominal number and verbal argument structure generalizations for tokens seen as few as two times during training. Second, we assess invariance properties of learned representation: the ability of a model to transfer syntactic generalizations from a base context (*e.g.*, a simple declarative active-voice sentence) to a transformed context (*e.g.*, an interrogative sentence). We test four models trained on the same dataset: an *n*-gram baseline, an LSTM, and two LSTM-variants trained with explicit structural supervision (Dyer et al., 2016; Charniak et al., 2016). We find that in most cases, the neural models are able to induce the proper syntactic generalizations after minimal exposure, often from just two examples during training, and that the two structurally supervised models generalize more accurately than the LSTM model. All neural models are able to leverage information learned in base contexts to drive expectations in transformed contexts, indicating that they have learned some invariance properties of syntax.

1 Introduction

Recurrent Neural Network language models (Elman, 1990; Hochreiter and Schmidhuber, 1997) have been shown to learn many aspects of natural language syntax including a number of long-distance dependencies and representations of incremental syntactic state (Marvin and Linzen, 2018; Wilcox et al., 2018; Futrell et al., 2018). However, previous studies have not investigated the relationship between a token's frequency in the training corpus and syntactic properties models learn about it. In this work, we assess neural models' ability to make robust syntactic generalizations about a token's nominal number or verbal argument structure based on minimal exposure with the token during training. Because of the Zipfian distribution of words in a corpus, the vast majority of word types will be seen only a handful of times during training (Zipf, 1949). Therefore, the few-shot learning capabilities of neural LMs are critical to their robustness as an NLP system and as a cognitive model.

However, human learning goes beyond simply learning syntactic properties in particular constructions. People apply the same properties across different constructions, meaning that their representations of the syntactic features of a word are in some sense *invariant* to the grammatical context of that word. For example, speakers and listeners are sensitive to a verb's argument structure relationships and can easily recognize that a verb which cannot take a direct object in active, declarative sentences cannot be passivized (as in the ungrammatical sentence *"The ship was arrived."*) The relationship between an active sentence and a passive sentence has been termed a *transformation* in the linguistic literature (Chomsky, 1957). Many semantic-syntactic rules that govern word co-occurrence in one form, such as a verb's argument structure relationships, hold uniformly across transformations. It remains an open question whether models learn grammatical rules invariant to their surface realization, a property we call *syntactic invariance*.

We combine assessment of few-shot learning and syntactic invariance for two grammatical features of English: whether a noun is singular or

---

plural (nominal number) and whether a verb is transitive or intransitive (verbal argument structure). We assess whether a model is able to make different predictions based on number or argument structure in a simple active voice *base context*. We then assess whether models are able to make similar distinctions in a *transformed context*—passive voice for verbs and polar questions for nouns. In the *transformed contexts*, we test models with tokens that occur only in the *base context* during training. (As a control, we also test verbal passive voice with tokens that occur in the passive voice in the training data to establish that models have learned the proper syntactic rules for this context.) For models to succeed in the transformed contexts they must represent syntactic features in a way that is invariant to the specific realization of those features in terms of word co-occurrences in different constructions. For each grammatical feature, we introduce a suite of novel targeted test sentences, similar to those presented in Marvin and Linzen (2018).

We find that all neural models tested are able to induce the proper syntactic generalizations in the *base* and *transformed* contexts after just two or three exposures, whereas a baseline *n*-gram model fails to learn the relevant generalizations. For all constructions tested our two neural models enhanced with explicit structural supervision outperform the purely sequence model. Assessing invariance properties, we find that neural models demonstrate proper behavior in transformed contexts, even for tokens seen only in base contexts during training. This behavior indicates that models are able to deploy generalizations learned in one syntactic context into different syntactic environments, a key component of human linguistic capabilities that has been so far untested in the neural setting.

1.1 Related Work

Bayesian models of word learning have shown successes in acquiring proper syntactic generalizations from minimal exposure (Tenenbaum and Xu, 2000; Wang et al., 2017), however it is not clear how well neural network models would exhibit these rapid generalizations. Comparing between neural network architectures, recent work has shown that models enhanced with explicit structural supervision during training produce more human-like syntactic generalizations (Kuncoro et al., 2017,

2018; Wilcox et al., 2019), but it remains untested whether such supervision helps learn properties of tokens that occur rarely during training.

Previous studies have found that Artificial Neural Networks (ANNs) are capable of learning some argument structure paradigms and make correct predictions across multiple frames (Kann et al., 2018), however these capabilities remain untested for incremental language models. Much has been written about the ability of ANNs to learn number agreement (Linzen et al., 2016; Gulordava et al., 2018; Giulianelli et al., 2018), including their ability to maintain the dependency across different types of intervening material (Marvin and Linzen, 2018) and with coordinated noun phrases (An et al., 2019). Hu et al. (2020) find that model architecture, rather than training data size, may contribute most to performance on number agreement and related tasks. Focusing on RNN models, Lakretz et al. (2019) find evidence that number agreement is tracked by specific "number" units that work in concert with units that carry more general syntactic information like tree depth. Jumelet et al. (2019) argue that when learning dependencies RNNs acquire a default form (which they postulate to be singular and masculine), and predicting a non-default form requires explicit contrary evidence. Our results support their hypothesis. Models are more accurate with singular nouns and transitive verbs seen only a few times in training, behavior that indicates these forms are expected when evidence is sparse.

2 General Methods

2.1 Psycholinguistic Assessment Paradigm

In order to assess the learning outcomes of neural LMs, we adopt the Psycholinguistic Assessment Paradigm (Linzen et al., 2016; Futrell et al., 2018). In this paradigm models are exposed to sentences that reveal the syntactic generalizations learned during training. For example, Linzen et al. (2016) used as input sentences with the prefix *The keys to the cabinet . . .* and inspected the relative probabilities of the continuations *is* and *are*. If the model has learned the grammatical rule that the head of a subject noun phrase governs plural agreement, then P(*are*) should be greater than P(*is*).

In order to assess the few-shot learning capabilities of the models tested, we sample words from eight "exposure buckets" based on the number of times the word appears in the training corpus, with the majority of buckets for words seen less than 10
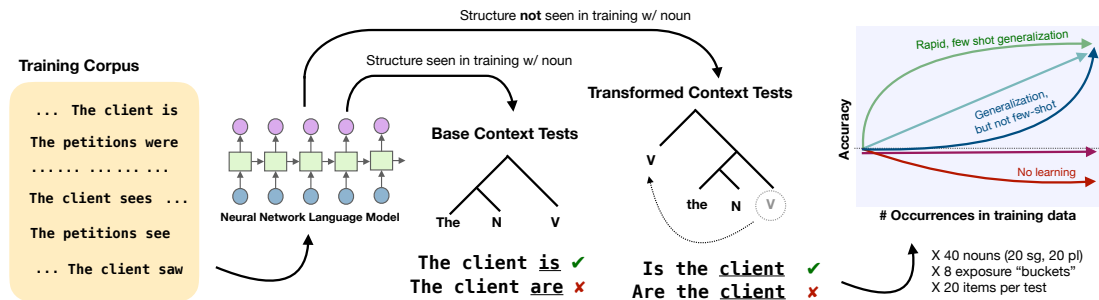
Figure 1: Experimental pipeline, given for Nominal Number tests without any modification.

times during training.[1] For each test, we sample 40 words balanced across syntactic categories. For each word, we generate 20 test sentences—each with a grammatical and ungrammatical condition—and measure the proportion of sentences in which the model shows higher probability in the grammatical condition in a particular critical region. In total, each test contains 12,800 sentences.

For each experiment, we report two metrics: First, we report the number of exposure buckets for which the models' accuracy is significantly above chance, which is 50% in all cases. Second, in order to assess the impact of structural supervision we report the results of a logistic regression model trained to predict accuracy with fixed effects of model class and exposure and random by-sentence intercepts.[2] A significant positive effect of model class means that the architecture contributes to more robust learning across all exposure buckets tested.

2.2 Models Tested

All models are trained on sections 2-21 of the Wall Street Journal portion of the Penn Treebank (PTB) (Marcus et al., 1993), which contains ∼1,000,000 tokens of newswire text. While this dataset is relatively small compared to the ones on which neural models were trained in Linzen et al. (2016) and Gulordava et al. (2018), it was imperative that we collect accurate metadata for all tokens in the training data and PTB is one of the largest datasets expertly hand annotated with both syntactic structure and part-of-speech information. Argument structure

statistics were obtained from a Universal Dependency representation of the dataset, converted from its original phrase structure parse via the Stanford Parser (Schuster and Manning, 2016). Additional argument structure information was collected from the Celex2 dataset (Baayen et al., 1995).

n-**Gram Baseline** We used a 5-gram baseline with modified Kneser-Ney smoothing trained using SRILM (Stolcke, 2002).

Recurrent Neural Network LMs model a sentence in a purely sequential basis, without explicitly representing the latent syntactic structure. We use the LSTM architecture (Hochreiter and Schmidhuber, 1997) and, following Futrell et al. (2018) derive the word surprisal from the LSTM language model by directly computing the negative log value of the predicted conditional probability from the softmax layer. This and subsequent neural models were trained with embedding size 256, dropout 0.3 following the hyper-parameters in Van Schijndel and Linzen (2018).

Recurrent Neural Network Grammars (RN-NGs) (Dyer et al., 2016) jointly model a sentence as well as its syntactic parse. The model explicitly represents parse trees and composes partially built phrase structures, an approach that may result in better performance on tree-structurally local but linearly distal relationships (see (Dyer et al., 2016)). Models are supervised with Penn-Treebank style parses during training; we assess whether this explicit syntactic supervision translates into better few-shot learning and syntactic invariance outcomes. We use the same hyperparameters used by Dyer et al. (2016).

ActionLSTM Model: We ablate the composition function of an RNNG, producing a model that predicts the **action sequence** of a parse tree as well as the upcoming word. In this sense, it is an incrementalized version of the Parsing-as-Language-

---

[1]Our exposure buckets were: 2, 3, 4, 5, 6-10, 11-20, 21-30, 50-100, and we use the end of the range to label the buckets in figures. Thus, if a token is in exposure bucket "4" it occurred four times in the training corpus; if it is in exposure bucket "20," it occurred between 11 and 20 times in the training corpus.

[2]The R code used to run the model was `glm(accuracy ∼ model + exposure-bucket + (1|sentence-id))`

Modeling configuration presented in Charniak et al. (2016). We refer to this model as the "ActionLSTM" model in the following sections.[3]

3 Nominal Number

In English, the matrix verb of a tensed clause must agree in number with the head of the subject Noun Phrase. Neural LMs are capable of learning this relationship based on a pure language modeling objective in multiple languages and across a variety of intervening material (Gulordava et al., 2018; Marvin and Linzen, 2018). Previous work assessing neural learning of subject-verb number agreement has not directly compared the learning outcome to the amount of experience models receive during training.

3.1 Base Contexts: Active Voice

In order to assess the few-shot learning capabilities of neural models in base contexts, we randomly select 20 plural and 20 singular nouns in each exposure bucket and generated test items for each following (1). (In (1) and following examples, ungrammatical sentence variants are marked with a *, which is for presentational purposes only and not included in test items. Underlined portions of the sentences indicate critical regions, whose joint probability was used to calculate accuracy scores.) To test whether models' representations are impervious to modification, we also generate a set of test items with prepositional phrases (PPs) and object relative clauses (ORCs) modifying the head verb, following (2) and (3). In this and all following experiments, sentences are generated using templates and—other than the target token—contain tokens that occur at least 50 times in the training data. As described in 2.1, strength of nominal number feature learning was evaluated by calculating model accuracy, or the proportion of times the models preferred the grammatical variant.

(1) **Base No Modifier** (singular example)
    a. The president <u>is</u>...
    b.*The president <u>are</u>...

(2) **Base w/ PP Modifier** (plural example)
    a.*The petitions near the old investment <u>is</u>...

---

[3]As RNNG and ActionLSTM jointly model terminal words and syntactic parses, we use word-synchronous beam search (Stern et al., 2017) to compute surprisal values incrementally. ActionLSTM was able to achieve a parsing F1 score of 92.81 on the PTB, which is in the same range as the original architecture on the same test set, as reported in Kuncoro et al. (2017).

    b. The petitions near the old investment <u>are</u>...

(3) **Base w/ ORC Modifier** (singular example)
    a. The client that the lawyers like <u>is</u>...
    b.*The client that the lawyers like <u>are</u>...

The results for this experiment can be seen in Figure 2, in the left panels. This is the same presentational paradigm we will use for all results in this paper: the y-axis is the model's accuracy, pooled across performance on singular and plural nouns, and the x-axis is the model's exposure for each token—the number of times it occurs during training. The scale of the x-axis is log-transformed. Points represent mean accuracy for each exposure bucket and error bars are 95% binomial confidence intervals. Lines show logistic regression fits from the raw data, with standard errors.

In the *Base Simple* experiment, the n-gram shows moderate few-shot learning, above chance in 5/8 exposure buckets. All neural models show robust few-shot learning and are above chance in all exposure buckets. We find a significant effect of structural supervision, with both the ActionLSTM and RNNG outperforming the LSTM model ($p < 0.05$ and $p < 0.001$ respectively). The experiments with modifiers prove more difficult: In the *Base PP* experiment, we find no few-shot generalization for the n-gram model, weak few-shot generalization for the LSTM (1/8 buckets), moderate generalization for the RNNG (4/8 buckets) and strong generalization for the ActionLSTM (7/8 buckets). In the *Base RC* experiments we find a similar pattern: No generalization for the n-gram (0/8 buckets), weak generalization for the LSTM (2/8 buckets), but stronger generalization for the two structurally supervised models (6/8 and 7/8 buckets for the RNNG and ActionLSTM respectively). For these two experiments, we find an effect of structural supervision on accuracy, with both the ActionLSTM and RNNG out-performing the LSTM ($p < 0.001$ except for the RNNG RC-Modifier where $p < 0.05$). Our results are generally in line with those presented in Marvin and Linzen (2018), who find performance in the 50-60% accuracy range for number agreement across PP and RC modifiers. Some studies, such as Lakretz et al. (2019) find higher performance performance on a similar task; we attribute these differences to the relative size of the training data.

Overall, these results indicate that all models are capable of making grammatical generalizations based on minimal exposure with a token, and capa-
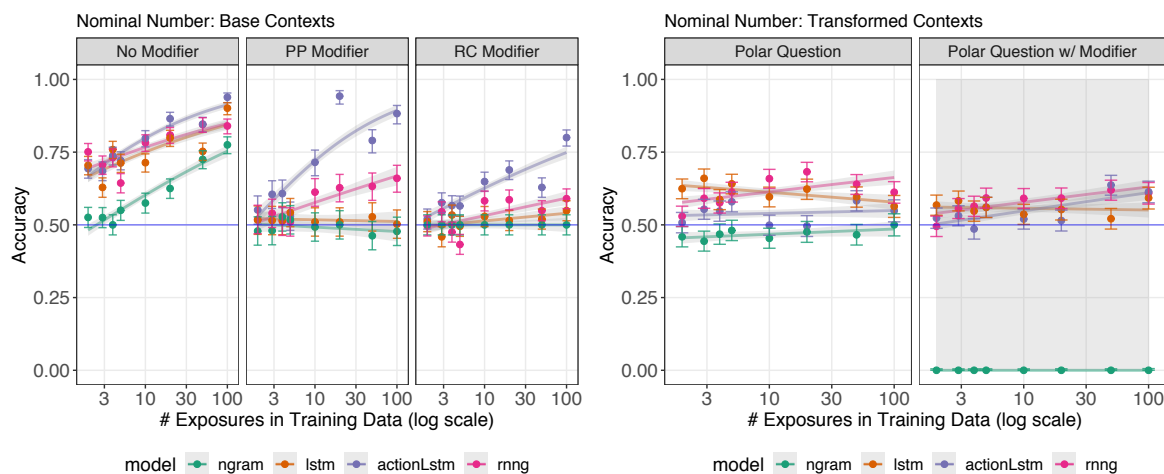
Figure 2: Results from the novel word learning experiments for Base Contexts (left) and Transformed Contexts (right). Results are averaged across singular and plural nouns. Points represent by-exposure bucket means, with 95% binomial confidence intervals. Smooth lines are results of logistic regression model fits on the raw data.

ble of leveraging increased exposure to make more accurate number agreement predictions. (For statistical analysis of the effects of exposure on accuracy across all of our experiments, see Appendix A.) Although the graphs in Figure 2 are pooled across noun category, models demonstrate much higher accuracy for *singular* nouns, especially if they occur only a few times in training. Improvement from increased exposure comes as models get better at accurately predicting number agreement for plural nouns. These findings are in line with the singular bias reported in Marvin and Linzen (2018) and support the hypothesis from Jumelet et al. (2019) that models have learned a "default" prediction, in this case *singular*. For results and analysis from all of our experiments broken down by grammatical category, see Appendix B.

3.2 Transformed Contexts & Syntactic Invariance

Because subjects precede verbs in English, most evidence available to a neural model for a noun's number follows the noun linearly. However, in English *polar question* formation, the matrix verb is moved to the front of the sentence inverting the base noun-verb order, as in (4). If models have learned nominal number feature that is invariant to linear order, verbs that set off polar questions should set up expectations for nominal subjects that match in number. In order to assess whether models were robust to such transformations, we created test items following the template in (4) and (5), which includes an additional four word modifi-

cation. Half of the sentences were in present tense, half in past tense. We measure the model's accuracy at the noun directly and pool accuracy scores across singular and plural nouns.

(4) **Polar Question** (singular example)
   a. Is the president...
   b. *Are the president...

(5) **Polar Question w/ Modifier** (plural example)
   a. *Is the very big and important hearings...
   b. Are the very big and important hearings...

Although polar questions are relatively rare in the WSJ section of the Penn Treebank—the ratio of active to inverted polar sentences is ∼1000:1—some nouns do occur in both base for and inverted form. Because our aim here is to assess models' generalization to novel syntactic frames, we filtered every noun from our previous set that occurred in both frames, a total of 15 nouns. The results presented here therefore address whether the models have learned a representation of number that is invariant to linear order. Successful learning, in this case, means that models have learned that nouns which set up expectations for singular verbal inflections should also be more likely in contexts where singular nouns are expected, and likewise for plurals.

The results for this experiment can be seen in Figure 2 in the right-hand panel. In the *Transformed Simple* experiment we find no generalization for the *n*-gram model, which was not above chance in any exposure bucket, moderate generalization for the ActionLSTM (above chance in 5/8 buckets) and strong generalization for the LSTM and RNNG model (8/8 buckets and 7/8 buckets

respectively). Results are similar for the *Transformed Modifier* experiment. In this case the *n*-gram model is at 0% accuracy for all buckets; this is because it assigns equal probability to the critical region in each condition, which we count as a "failure." The ActionLSTM displays moderate generalization (4/8 buckets) and the RNNG and LSTM stronger generalization (7/8 and 8/8 buckets respectively). In these experiments, the ActionLSTM under-performs compared to the LSTM in the *No Modifier* experiment ($p < 0.001$), but the RNNG outperforms it in the *Transformed Modifier* experiment ($p < 0.05$). These results indicate that all the neural models are able to leverage information gained in the base contexts to drive expectations in the inverted context, however note that the accuracy scores are lower here than in the base contexts, with no model breaking the 70% accuracy threshold. This may be due to the relatively few number of polar questions in the corpus.

4 Verbal Argument Structure

In this section, we assess the ability of neural models to represent verbal argument structure, which we simplify to whether a verb is transitive or intransitive. If it is transitive, then it requires a theme, which must be realized as a direct object in the active voice. If it is intransitive then it requires an empty theme position and cannot have an object in the active voice. Verbal argument structure is a hard task, insofar as both intransitivity and transitivity can only be inferred through *indirect negative evidence*. We assess neural models' sensitivity for indirect negative evidence by investigating how much experience models need with a particular verb before they make robust predictions about whether an object should follow that token at test time.

4.1 Base Contexts: Active Voice

In order to assess the generalizations models have learned about verbal argument structure, we selected 20 transitive verbs and 20 intransitive verbs. Verb transitivity was assessed using the hand-coded Celex2 Corpus, and double-checked using a Universal Dependencies representation of the original PTB Phrase Structure trees; verbs marked "transitive" in Celex were dropped if they occurred without arguments in 3rd person past tense active voice less than 90% of the time and verbs marked "intransitive" in Celex were dropped if they took arguments in 3rd person past tense more than 10% of the time.

We generated sentences for each verb in an *infinitival* construction, following (6) (which gives a transitive example) and in *past tense* following (7) (which gives an intransitive example). Infinitival tense sentences were generated because most verbs are ambiguous between their past tense and past participle forms, which can occur without a direct object, even for transitive verbs (e.g. *The patient was cured*). Model accuracy was assessed by comparing the surprisal of the `adverb + period` region at the end of each sentence (e.g. "today ." in the examples below) and accuracy scores are pooled across verb type. For transitive verbs, this region should be more surprising if an object is absent; for intransitive verbs this region should be more surprising if it is present.

(6) **Active, Infinitival Tense** (transitive example)
   a. The doctor can cure the patient <u>today.</u>
   b. *The doctor can cure <u>today.</u>

(7) **Active, Past Tense** (intransitive example)
   a. *The doctor slept the patient <u>today.</u>
   b. The doctor slept <u>today.</u>

The results for this experiment can be seen in Figure 3 on the left, with the *infinitival* experiment on the far left and the *past tense* experiment in the middle. For the *Active Infinitival* results we find no generalization for the *n*-gram model, but strong generalization for the three neural models: The LSTM is significantly above chance in 6/8 exposure buckets, the ActionLSTM in 8/8 and the RNNG in 7/8. For the *Active Past* experiment we find no few-shot learning for the *n*-gram model (it is above chance in 0/8 buckets), moderate few-shot learning for the LSTM model (5/8 buckets) but strong few-shot learning for the two supervised models (8/8 buckets for both). We find a significant effect of structural supervision, whereby the ActionLSTM and RNNG outperform the LSTM in both experiments ($p < 0.001$). These results indicate that not only that all neural models have learned the basic facts of argument structure, but that they are willing to generalize about the likelihood of an upcoming object after just two exposures during training.

Interestingly, the LSTM model shows a decrease in accuracy in our Past Tense test as words grow more frequent in the training data. We hypothesized that this is because high-frequency tokens ending in "-ed" are more likely to be used in pas-
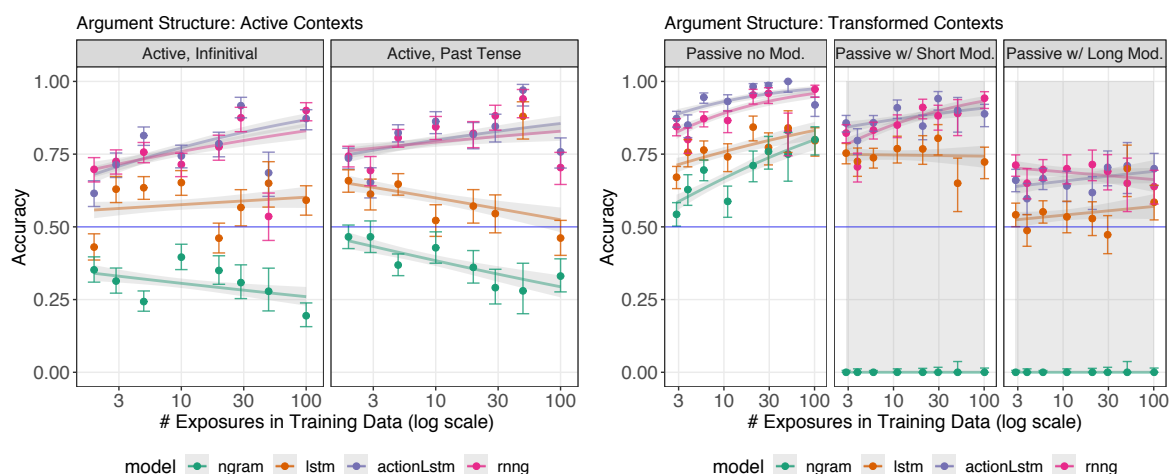
Figure 3: Results from the few shot learning experiments for active contexts (left) and transformed, passive, contexts (right). Points represent mean accuracy for each exposure and error bars are 95% binomial confidence intervals. Lines show logistic regression fits from the raw data, with standard errors.

sive voice, where they occur without a direct object. Models that lack POS disambiguation, would erroneously come to expect these tokens to occur without an object, even when they are being used in active voice, which explains why we do not see the trend for infinitival tests, nor for the supervised models which predict POS tags. We confirmed our hypothesis with two follow-up statistical tests: We found a positive correlation between a token's frequency and the percentage of time it is used in passive voice in the training data ($cor = 0.39; p < 0.001$). And we ran a statistical model looking at the effect of passive usage on accuracy, finding a positive effect for the two supervised models and a negative effect for the LSTM ($p < 0.001$ in all cases).[4]

4.2 Transformed Context: Passive Voice

When verbs are realized in different syntactic frames, or *syntactic transformations*, their argument structure properties are preserved. For example, because passive voice promotes the semantic *theme*—which plays the syntactic role of object in active voice—to the subject position, it is impossible for a truly intransitive verb to be realized in passive voice, as in the ungrammatical (8-c).

In order to assess whether models have learned the rules governing the passive transformation, we designed tests with the same verbs as in Section 4.1). Verbs were dropped if they had different

forms for 3rd person past-tense and passive participle, such as the verb *give* (*gave*, *given*). We generated items following three tests sketched in (8), (9) and (10). Model accuracy was assessed by comparing the surprisal of the `verb + adverb + period` following a prefix that contains a passive "was" versus a prefix that does not, and accuracy scores are pooled across transitive and intransitive verbs. If models are learning the proper grammatical generalizations, then intransitive verbs should be unexpected in passive voice, and the verb should be more surprising when it follows the passivizing "was". Conversely, transitive verbs should be more likely in passive voice than in active voice without a direct object, which is ruled out by the `adverb + period` portion of our continuation, and therefore should be more likely when the passivizing "was" is absent. Because we use verbs that do appear in passive voice during training this section does not test models' invariance properties, but rather their few-shot learning capabilities for this grammatical context. After these control experiments, we turn to invariance tests in Section 4.3.

(8) **Passive Voice: No Modifier**
Example with Transitive verb
a. The doctor was <u>cured yesterday</u> .
b. *The doctor <u>cured yesterday</u> .
Example with Intransitive Verb
c. *The doctor was <u>arrived yesterday</u> .
d. The doctor <u>arrived yesterday.</u>

(9) **Passive Voice: Short Modifier**
a. The dog was quickly and fully <u>cured today.</u>
b. *The dog quickly and fully <u>cured today.</u>

---

[4]The statistical model used was `glm(accuracy ∼ percent_VBN + (1|sentence_id)`, where the predictor `percent_VBN` indicates the proportion of total occurrences the token is tagged as a passive participle.
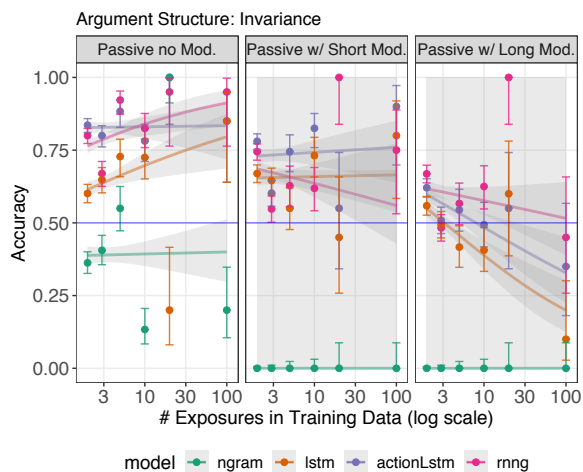
Figure 4: Invariance to transformation for argument structure. Error bars are 95% confidence intervals.

(10) **Passive Voice: Long Modifier**

　　a. The dog was quickly, suddenly, and entirely cured yesterday.

　　b. *The dog quickly, suddenly, and entirely cured yesterday.

The results for this experiment can be seen in Figure 3, in the right panel. For the *Passive No-Modifier* test, all models are significantly above chance for all exposure buckets tested. For the *Short Modifier* test, we find no few-shot learning for the *n*-gram model (above chance in 0/8 buckets), but strong few-shot learning for all neural models tested (8/8 buckets). For the *Long Modifier* experiment the *n*-gram shows no few-shot learning (0/8 buckets) and the LSTM shows moderate few-shot generalization (4/8 buckets), but the RNNG and ActionLSTM are still robust (8/8 buckets for both). Across all three experiments, we find that the structurally supervised neural models perform better than the LSTM ($p < 0.001$).

For these experiments, models were more accurate with transitive verbs, especially for ones that occur infrequently during training. This transitive bias was also present in the *base context* tests for the ActionLSTM and RNNG, indicating that transitivity may be the default assumption, and models expect verbs to be able to occur in passive frames unless they have a large amount of indirect negative evidence to the contrary.

4.3 Syntactic Invariance

In this section, we run the same *passive* experiments described in Section 4.2, however we use verbs that occur only in the *active* voice during training. In order for models to achieve higher than 50% accuracy, they must learn the co-variation between direct objects in the active voice and passive nominal subjects in the passive voice, even for verbs which they have only seen in the active voice. That is, they must learn a grammatical rule that is invariant to syntactic transformation and verb type.

We sampled all the verbs that occurred with the VBD but not the VBN (past participle) part-of-speech in our training data and generated 20 sentences for each verb. This came to 56 verbs in total, with no transitive verbs in the 50 or 100 exposure buckets. The results from this experiment can be seen in Figure 4. For the *No-Modifier* experiment, the *n*-gram model shows little few-shot learning, above chance in only 1/8 exposure buckets. However, the neural models show moderate few-shot learning, with accuracy scores above chance in 5/8 buckets (LSTM), 6/8 (ActionLSTM) and 6/8 (RNNG). For the *Short Modifier* experiment we find no few-shot learning for the *n*-gram models, but moderate few-shot learning for all neural models tested (LSTM: 4/8 bucekts; ActionLSTM: 5/8 and RNNG: 6/8). Models fare worse in the *Long Modifier* experiment with week few-shot learning for the LSTM and ActionLSTM (1/8 buckets each) and moderate few-shot learning for RNNG model (4/8 buckets each). The *n*-gram is below chance in all buckets. Turning to the effects of structural supervision: We find that the RNNG and the ActionLSTM generally outperform the LSTM ($p < 0.001$ for all three, except RNNG/*Short Modifier* which is not significant). Because there were so few verbs in our training data that occurred only in active verbal frames, error estimates are larger for this experiment and the results are somewhat less consistent. Despite this, it is clear that in the *No Modifier* and *Short Modifier* tests, all the neural models show moderate accuracy outcomes, indicating that their learned representations are at least somewhat invariant to syntactic information.

5 Discussion

In this paper, we have tested the few-shot learning capabilities of neural language models, as well as whether these models can learn grammatical representations that are invariant to syntactic transformation. First, we addressed neural models' ability to learn **nominal number**, introducing a novel testing paradigm that leveraged *polar questions* to assess subject/verb number agreement learning in syntactically transformed settings. Second, we

Test	Few-Shot Learning				Structural Supervision	
	n-gram	LSTM	Action LSTM	RNNG	Action LSTM	RNNG
Number	5/8	8/8	8/8	8/8	* * *	*
Number w/ PP Modifier	0/8	1/8	7/8	4/8	* * *	* * *
Number w/ RC Modifier	0/6	2/8	7/8	5/8	* * *	*
Verbal Arg. Struct. Infinite	0/6	6/8	8/8	7/8	* * *	* * *
Verbal Arg. Struct. Past	0/8	5/8	8/8	8/8	* * *	* * *
Arg. Struct. Transformation	8/8	8/8	8/8	8/8	* * *	**
Arg. Struct. w/ Modifier	0/8	8/8	8/8	8/8	*	* * *
Arg. Struct w/ Long Modifier	0/8	4/8	8/8	8/8	* * *	* * *
Number Transformation Simple	0/8	8/8	5/8	7/8	! * * *	n.s.
Number Transformation w/ Modifier	0/8	7/8	4/8	7/8	n.s.	*
Arg. Struct. Transf.	1/8	5/8	6/8	6/8	* * *	* * *
Arg. Struct. Transf. w/ Modifier	0/8	4/8	5/8	6/8	* * *	n.s.
Arg. Struct Transf. w/ Long Modifier	0/8	1/8	1/8	4/8	* * *	* * *

Table 1: Left columns: Few shot learning outcomes with the results from our tests of syntactic invariance in the bottom quadrant. Colors correspond to the proportion of exposure buckets for which each model achieved accuracy scores significantly above chance, colored by tertiles. Right columns indicate whether the two structurally supervised models outperform the LSTM for each test, where *s indicate the significance level from our statistical tests and !s indicate significantly worse performance than the LSTM.

turned to neural models' ability to represent verbal **argument structure**, developing two novel suites of tests that assessed preference for *themes*—either realized as direct objects or passive subjects—in both active contexts and passive contexts. In each experiment we assessed the effect of syntactic supervision on learning outcomes by comparing two supervised models to one purely sequence model.

A summary of our results can be seen in Table 1, with few-shot learning outcomes in colored cells on the left, and the effect of structural supervision on the right. The results from experiments that assess syntactic invariance are on the bottom, below the line break. This table makes it clear that all neural models are capable of making syntactic generalizations about a token from minimal exposure during training. Although model accuracy is reduced for tests that assess syntactic invariance, all neural models show at least a moderate ability to generalize across syntactic transformations. Furthermore, Table 1 shows that syntactic invariance is enhanced in structurally supervised models. Interestingly, both ActionLSTM and RNNG have access to syntactic information, but the comparison in Table 1 indicates that RNNG can leverage that information more effectively to produce syntactic invariance. Therefore we suggest that RNNG's improved performance does not come from the mere *presence* of syntactic information in the training and test data, but rather from the fact that it uses syntactic information to structure its computation in a non-sequential way.

Because these experiments require careful and robust syntactic analysis of the training data, we evaluated models trained on a relatively small, human-annotated corpus. While the small training data poses some limitations when interpreting the results, it makes them more relevant to low-resource NLP applications and suggests that using structurally supervised models can lead to better generalization in a sparse data environment. While sub-word tokenization schemes such as Byte-Pair Encoding (Sennrich et al., 2015) have helped reduce the number of individual lexical items that need to learned, they do not completely eliminate the long tail of sub-word units. Thus, robust few-shot generalization is still an important problem in these environments. It may be that larger amounts of training data support even better few-shot learning and syntactic invariance outcomes. Scaling these carefully-controlled methods to the larger data setting will be an important next step. However, even with the relatively small models tested here, the results support a growing body of evidence that incremental statistical models of language are able to induce many key features of human linguistic competence.

Acknowledgements

References

Aixiu An, Peng Qian, Ethan Wilcox, and Roger Levy. 2019. Representation of constituents in neural language models: Coordination phrase as a case study. *arXiv preprint arXiv:1909.04625*.

R Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. The celex lexical database (release 2). *Distributed by the Linguistic Data Consortium, University of Pennsylvania*.

Eugene Charniak et al. 2016. Parsing as language modeling. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Noam Chomsky. 1957. *Syntactic structures*. Walter de Gruyter.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. *arXiv preprint arXiv:1808.08079*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P Levy. 2020. A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692*.

Jaap Jumelet, Willem Zuidema, and Dieuwke Hupkes. 2019. Analysing neural language models: Contextual decomposition reveals default reasoning in number and gender assignment. *arXiv preprint arXiv:1909.08975*.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R Bowman. 2018. Verb argument structure alternations in word and sentence embeddings. *arXiv preprint arXiv:1811.10773*.

Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig, and Noah A. Smith. 2017. What do recurrent neural network grammars learn about syntax? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. 2018. Lstms can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in lstm language models. *arXiv preprint arXiv:1903.07435*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Sebastian Schuster and Christopher D Manning. 2016. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Mitchell Stern, Daniel Fried, and Dan Klein. 2017. Effective inference for generative neural parsing. *arXiv preprint arXiv:1707.08976*.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing*.

Joshua B Tenenbaum and Fei Xu. 2000. Word learning as bayesian inference. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Marten Van Schijndel and Tal Linzen. 2018. A neural model of adaptation in reading. *arXiv preprint arXiv:1808.09930*.

Su Wang, Stephen Roller, and Katrin Erk. 2017. Distributional modeling on a diet: One-shot word learning from text only. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler-gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. 2019. Structural supervision improves learning of non-local grammatical dependencies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

George Kingsley Zipf. 1949. *Human behavior and the principle of least effort.* addison-wesley press.

A Effect of Exposure on Model Accuracy

In this section we report the result for statistical tests assessing the effect of a token's frequency in training on model accuracy for that token. We derive significance from a general linear model with # of exposures as a sole predictor, with random by-item intercepts (`glm(accuracy ~ #_occurrences + (1|item_number))`))

Nominal Number For the base context, in the no modifier condition we find a positive effect of increased exposure for all models ($p < 0.001$). For the *PP modifier* test we find an effect of exposure for the ActionLSTM and the RNNG ($p < 0.001$), and a negative, but insignificant effect for the *n*-gram and the LSTM. For the *RC Modifier* experiment we find an effect of increased exposure for all three neural models ($p < 0.001$ for the RNNG and ActionLSTM; $p < 0.05$ for the LSTM), but no effect for the *n*-gram. For the inverted contexts: in the *no modifier* tests we find no effect of increased exposure, except for the LSTM, where the effect is negative ($p < 0.01$). For the *modifier* tests, we find a significant effect for the ActionLSTM and the RNNG ($p < 0.001$).

Argument Structure For the base context (active voice): In the infinitival tests, we find a significant effect of exposure on accuracy for the ActionLSTM and the RNNG ($p < 0.001$) and a negative effect for the *n*-gram model ($p < 0.001$). In the past-tense, we find no significant effect for the RNNG
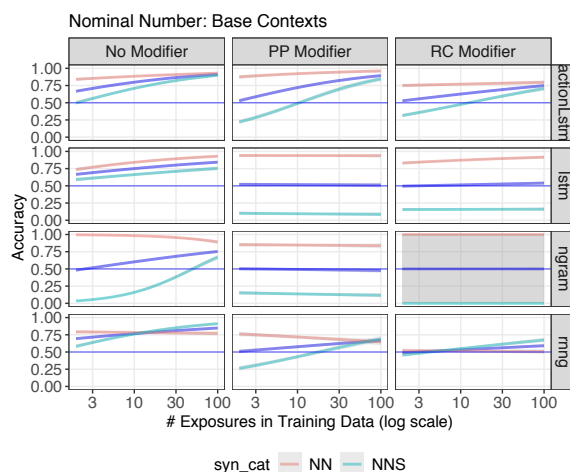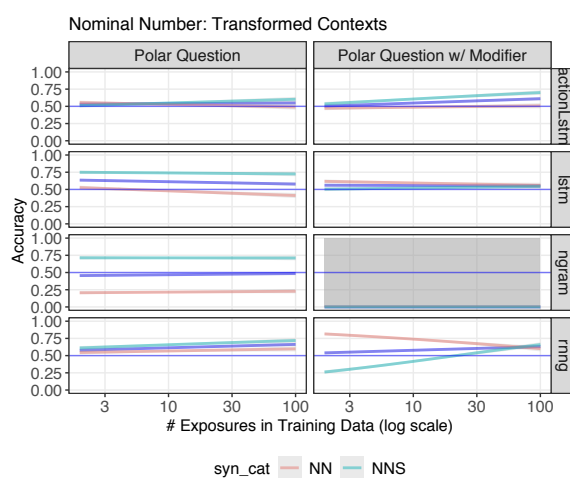

Figure 5: Number: Base Contexts

Figure 6: Number: Transformed Contexts

or the ActionLSTM, and a negative effect for the *n*-gram and LSTM models ($p < 0.001$). In the transformed contexts (passive voice), for the *no-modifier* tests we find a significant effect of exposure for all models ($p < 0.001$ for all except ActionLSTM where $p < 0.05$). For the *short-modifier* tests we find an effect for the ActionLSTM ($p < 0.05$) and the RNNG ($p < 0.001$). And in the *long-modifier* test we find a marginally significant effect for the three neural models ($p \sim 0.05$ for all).

B Learning Outcomes by Grammatical Condition

In this section, for each test reported in the paper, we break down model performance by grammatical category, either singular vs plural nouns (for nominal number tests) or transitive vs. intransitive verbs (for our argument structure tests). Charts follow the same presentational paradigm: *y*-axis shows

accuracy and *x*-axis the number of times each word appears during training, on a log-10 scale. Smooth lines are results of logistic regression model fits on the raw data, with shaded regions indicating standard error. Dark blue lines show model performance averaged between the two conditions (these are the same fits that appear in charts in the main body of the paper).

The data presented here are consistent with the hypothesis from (Jumelet et al., 2019). When models receive scant evidence of a token's syntactic properties in training, they assume that it belongs to a "base" category, which is *singular* for nouns and *transitive* for verbs. Thus, models are more accurate for singular nouns and transitive verbs seen rarely in training. As the model receives more evidence that a token is not in the base category, its predictions flip. Hence, gains in overall-accuracy tend to come from models learning the proper agreement for non-base tokens (plural nouns and intransitive verbs). Generally, these effects are stronger for nominal number learning, and stronger for structurally supervised models than for the LSTM, which is consistent with the findings presented in the main body of the text.

B.1 Number: Base Contexts

The nominal number breakdown for base contexts can be seen in Figure 5, with accuracy scores for singular nouns (NN) in red and plural nouns NNS) in teal. Over all, models tended to show higher accuracy scores for singular nouns, which indicates the presence of a singular bias. Interestingly, the ActionLSTM and the RNNG are capable of overcoming the singular bias when presented with sufficient data, however the LSTM remains equally biased for tokens seen 2 and 100 times in training.

B.2 Number: Transformed Contexts

The nominal number breakdown for transformed can be seen in Figure 6. The empirical picture is more complicated here, however if anything models show higher performance for plural nouns. This behavior suggests that *is* sets up weaker expectations for singular nouns than *are* does for plural nouns. Such a pattern is consistent with the hypothesis that models learn the singular as a base form, in which case it would set up weaker expectations for singular nouns. These results compliment those from An et al. (2019) (section 6), who also test in inverted settings and find that models tend not to be surprised at coordinated NPs following a singular
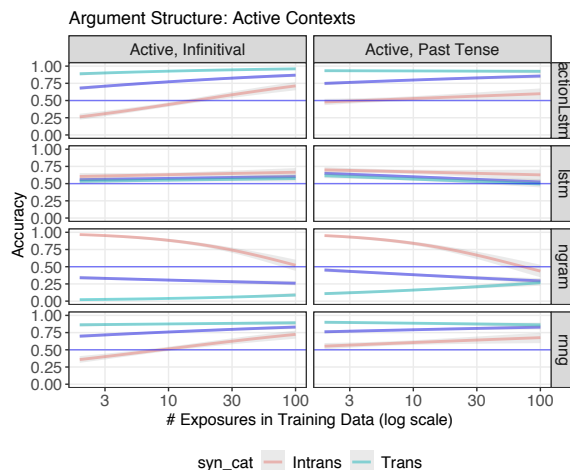

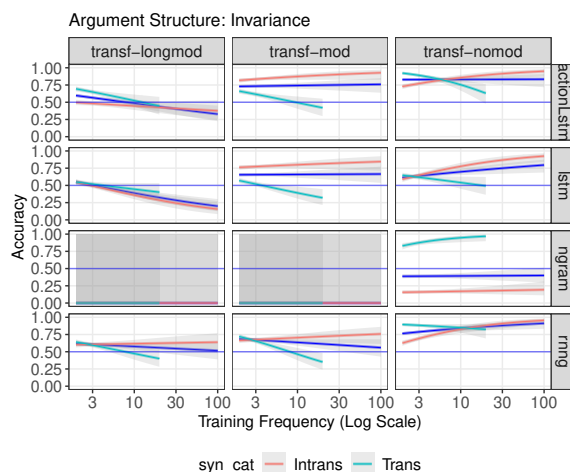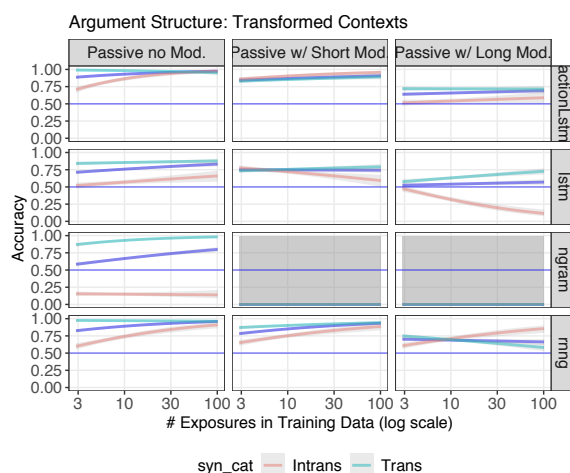Figure 7: Number: Arg Struct: Base Contexts

Figure 8: Argument Structure: Transformed Contexts (top) and Invariance (bottom)

verb, as in the ungrammatical sentence *What is the pig and the cat eating?*

B.3   Argument Structure: Base Contexts

The breakdown for argument structure learning base contexts can be seen in Figure 7, with accuracy scores for *intransitive* verbs in red and *transitive* verbs in teal. Here, we see a strong transitive bias for the two structurally supervised models, with no obvious bias for the LSTM and an intransitive bias for the *n*-gram.

B.4   Argument Structure: Transformed Contexts and Invariance

The breakdown for argument structure learning in the transformed contexts can be seen in Figure 8 with transformation tests on the top and invariance tests on the bottom. In this case, where performance is different between the two conditions models display higher accuracy scores for transitive verbs.