Word Frequency Does Not Predict Grammatical Knowledge in Language Models

Charles Yu, Ryan Sie, Nico Tedeschi, and Leon Bergen
{cty002,rdsie,njtedeschi,lbergen}@ucsd.edu
University of California, San Diego

Abstract

Neural language models learn, to varying degrees of accuracy, the grammatical properties of natural languages. In this work, we investigate whether there are systematic sources of variation in the language models' accuracy. Focusing on subject-verb agreement and reflexive anaphora, we find that certain nouns are systematically understood better than others, an effect which is robust across grammatical tasks and different language models. Surprisingly, we find that across four orders of magnitude, corpus frequency is unrelated to a noun's performance on grammatical tasks. Finally, we find that a novel noun's grammatical properties can be few-shot learned from various types of training data. The results present a paradox: there should be less variation in grammatical performance than is actually observed.

1 Introduction

Neural language models (Howard and Ruder, 2018; Devlin et al., 2019; Dai et al., 2019; Yang et al., 2019; Radford et al., 2019) have achieved success in both text prediction and downstream tasks such as question-answering, text classification, and natural language inference. The strong performance of these models raises scientific questions about the knowledge they have acquired, in particular, about the abstractness and generality of their linguistic representations.

Previous work has investigated the linguistic representations of neural language models in several domains, and found varying evidence for how linguistically adequate these representations are (Lau et al., 2017; Marvin and Linzen, 2018; Goldberg, 2019; Futrell et al., 2019). This work has employed psycholinguistic methodology in order to elicit grammatical judgments from these models, inferring the models' underlying representations from the patterns of judgments.

In the current work, we focus on the variation in grammatical knowledge that potentially exists within a neural language model. Just as in human psycholinguistic tasks, previous work on neural LMs has observed variability in grammatical judgments between different sentences; not all violations of a grammatical constraint are judged to be equally bad. It is not clear, however, whether there are systematic sources of variation in these judgments, and if so, what the sources are.

We will focus on variation among lexical items, using English subject-verb agreement and reflexive anaphora as a case study. We first ask whether language models learn the grammatical properties of some nouns more accurately than for others. We do this by measuring the accuracy of language models when making grammatical judgments involving different nouns. We find systematic variation among nouns: nouns that perform well on one task or language model are more likely to perform well on other tasks or other language models. We then consider possible sources of the observed variation between nouns, finding that the grammatical properties of nouns are paradoxically easy to learn; our results suggest that there should be much less variation than is actually observed.[1]

Related work

A number of other studies have investigated the linguistic representations of neural models, both language models specifically and networks trained using other objectives. Linzen et al. (2016); Gulordava et al. (2018); Kuncoro et al. (2018) probe the ability of LSTMs to learn hierarchical structures. Warstadt et al. (2019b) introduces a large-scale corpus of grammatical acceptability judgements, trains RNNs to predict these judgments, and concludes

---

[1]All code and experimental materials are available at https://github.com/CharlesYu2000/lm-variation

that the models outperform unsupervised baselines, but fall far short of human performance. Lepori et al. (2020) finds that tree-based RNNs outperform sequential RNNs on number prediction tasks, but that fine-tuning on an artificially-generated augmentation set can bring the models closer to parity.

Other work has focused on probing whether neural language models have acquired adequate representations of specific linguistic phenomena. Marvin and Linzen (2018) and Goldberg (2019) use a minimal pair methodology to assess the grammatical knowledge of RNNs and BERT, looking at subject-verb number agreement, reflexive anaphora, and negative polarity items. Wilcox et al. (2018) examines whether RNN language models exhibit *wh*-licensing interactions on surprisal associated with gaps, concluding they can represent long-distance filler-gap dependencies and learn certain island constraints. Futrell et al. (2019) studies whether neural language models show evidence for incremental syntactic state representations using psycholinguistic methodology. Warstadt et al. (2019a) studies BERT's knowledge of NPI's, focusing on differences between tasks: boolean classification (e.g. Linzen et al. 2016 and Warstadt et al. 2019b), minimal pair comparisons (e.g. Marvin and Linzen 2018 and Wilcox et al. 2019), and probing tasks (e.g. Giulianelli et al. 2018).

2 Approach

We use the minimal pair methodology of Marvin and Linzen (2018) in order to investigate the grammatical judgments of neural language models. Given a minimal pair of sentences, i.e. a pair that differ from each other in their acceptability due to a difference in just one grammatical property. If the model understands the grammatical phenomenon being studied, it should assign higher probability to the grammatical sentence than to the ungrammatical sentence.

2.1 Grammatical tasks

Table 1 shows the 10 grammatical tasks (Marvin and Linzen, 2018) and the templates used for generating minimal pairs. The tasks fall into two general categories: subject-verb agreement (SVA) and reflexive anaphora (RA). The first SVA task, SVA Simple, probes whether the model understands that subject number must agree with the number of third-person present verbs:

(1) a.   The cat walks.

     b.   *The cat walk.

The other SVA tasks probe whether the models have more sophisticated representations of number agreement. For example, the SVA PP task measures whether the model is able to ignore distractors ("boys") which occur between the head of the subject and the verb:

(2) a.   The cat next to the boys jumps.

     b.   *The cat next to the boys jump.

The object relative clause tasks probe whether the model accurately maintains the head's number in the presence of an embedded clause. Marvin and Linzen (2018) provide extensive discussion of the linguistic motivation for these tasks.

The RA tasks measure whether the language model understands the structural conditions on the binding of reflexive pronouns. The tasks make use of the following property of English reflexives: a reflexive pronoun needs to agree in number with its antecedent. The RA Sent.Comp task evaluates whether the model understands that reflexives must be in the same clause as their antecedents:

(3) a.   The lawyers said the defendant incriminated himself.

     b.   *The lawyers said the defendant incriminated themselves.

The RA tasks involving object relative clauses evaluate whether the models understand that reflexive anaphora do not bind to the noun in an embedded clause but rather to the head noun.

2.2 Measuring the performance of a noun

We use these tasks in order to measure how well the model understands the grammatical properties of a particular *target noun*. Given a specific target noun, it is substituted as the TargetNoun in each of the task templates shown in Table 1. This gives a partially specified template. For example, substituting the target noun "zombie" in the SVA Simple template results in:

(4)   The zombie ⟨Verb⟩.

Given each of these partially specified templates, 500 minimal pairs are randomly sampled by filling in the remaining lexical items. Finally, the model's

Task	Template
SVA Simple	The ⟨TargetNoun⟩ ⟨Verb⟩.
SVA Subj.Rel.Clause	The ⟨TargetNoun⟩ that liked the ⟨Noun⟩ ⟨Verb⟩.
SVA Sent.Comp.	The ⟨Noun⟩ said the ⟨TargetNoun⟩ ⟨Verb⟩.
SVA PP	The ⟨TargetNoun⟩ next to the ⟨Noun⟩ ⟨Verb⟩.
SVA Obj.Rel.Clause.That	The ⟨TargetNoun⟩ that the ⟨Noun⟩ liked ⟨Verb⟩.
SVA Obj.Rel.Clause.NoThat	The ⟨TargetNoun⟩ the ⟨Noun⟩ liked ⟨Verb⟩.
RA Simple	The ⟨TargetNoun⟩ ⟨PastTransVerb⟩ ⟨himself/themselves⟩.
RA Sent.Comp.	The ⟨NonGenderedNoun⟩ said the ⟨TargetNoun⟩ ⟨PastTransVerb⟩ ⟨himself/themselves⟩.
RA Obj.Rel.Clause.That	The ⟨TargetNoun⟩ that the ⟨NonGenderedNoun⟩ liked ⟨PastTransVerb⟩ ⟨himself/themselves⟩.
RA Obj.Rel.Clause.NoThat	The ⟨TargetNoun⟩ the ⟨NonGenderedNoun⟩ liked ⟨PastTransVerb⟩ ⟨himself/themselves⟩.

Table 1: Templates used for sentence generation. TargetNoun indicates the position of the target noun whose performance score is being calculated.

grammatical judgments on the 500 minimal pairs are computed (by taking the difference in scores between the grammatical and ungrammatical variants) and averaged, resulting in a task performance score for the noun.

2.3 Limitations

These analyses are limited in several respects. First, only two grammatical tasks are used. By using a wider range of tasks, it will be possible to investigate a larger set of grammatical phenomena outside of number agreement.

Second, while the study focuses on the grammatical information carried by nouns, other lexical types such as verbs are likely to carry this information as well. Future work can determine whether the approach generalizes to verbs and other lexical types.

Finally, while the study uses acceptability judgments in order to determine the models' grammatical knowledge, other probing tasks exist and may produce different results (Warstadt et al., 2019a). We use acceptability judgments because, to the best of our knowledge, feature probing has not been extensively studied for GPT-2 or Transformer-XL. Different probing architectures may produce different results for these models. It would be desirable to understand the robustness of the current results to the choice of experimental readout.

3 Methods

In this section we describe the process of calculating a target noun's task performance score in more detail.

3.1 Sentence generation

Using WordNet (Fellbaum, 1998) and VerbNet (Schuler, 2005), we compiled a list of lexical items

as shown in Table 2. The target nouns were drawn from the Noun list, which consisted of animate nouns. Only nouns with distinct singular and plural forms were included. All verbs in the Verb set have an intransitive reading. For each pair of task template and target noun, 500 sentences were randomly sampled by choosing lexical items from the appropriate word lists.

For each sampled sentence, 2*2 or 2*2*2 versions were generated (depending on the template). These versions varied the grammaticality of the sentence and the plurality of the target noun and any distractor nouns. For example, for the SVA Simple task, 2*2 versions are generated for every sampled sentence:

(5) a. Singular-Grammatical: The horse walks.

   b. Singular-Ungrammatical: *The horse walk.

   c. Plural-Grammatical: The horses walk.

   d. Plural-Ungrammatical: *The horses walks.

3.2 Models

Our experiments use three models, Transformer-XL (Dai et al., 2019), GPT-2 (Radford et al., 2019), and BERT (Devlin et al., 2019). We use the Hugging Face implementations (Wolf et al., 2019) with the pre-trained models *transfo-xl-wt103*, which is trained on the WikiText-103 dataset, *gpt2-xl*, which is trained on the WebText dataset, and *bert-base-uncased*, which is trained on BookCorpus and English Wikipedia.

3.3 Sentence scoring

We now describe how a score was calculated for a particular sampled sentence. For each of the

Set Name	Transformer-XL	GPT-2	BERT
Noun	916	723	704
Verb	615	228	406
NonGenderedNoun	870	679	663
PastTransVerb	1298	1034	1298

Table 2: Size of word sets for each model.

sentence variants (e.g. Example 5), the model computes a score. In the case of Transformer-XL and GPT-2, this score is simply the the log probability of the string. For example, for Transformer-XL:

$$\text{Score}_{\text{string}}(s) = \log P_{\text{TXL}}(s) \quad (1)$$

where P_{TXL} is the Transformer-XL language model probability distribution.

For BERT, given its masked language model architecture, we follow the approach of Goldberg (2019). For the SVA tasks, we compute the log conditional probability of the verb whose number must agree with the target noun. For the RA tasks, we compute the log conditional probability of the reflexive pronoun. Both conditional probabilities are computed conditional on the left and right contexts.

Given the scores for a sentence's variants, we compute an overall score for the sentence, which captures how much the model prefers the grammatical variants to the ungrammatical variants. For each sampled sentence S, there are either 2 or 4 minimal pairs among its variants. In Example 5, a. and b. is a minimal pair, and c. and d. is a minimal pair. Letting $s_a, ..., s_d$ denote these variants, the overall score for the sentence is given by:

$$\text{Score}_{\text{sent}}(S) = \frac{1}{2}(\text{Score}_{\text{string}}(s_a) - \text{Score}_{\text{string}}(s_b)$$
$$+ \text{Score}_{\text{string}}(s_c) - \text{Score}_{\text{string}}(s_d))$$

The formula when there are four minimal pairs is similar.

3.4 Noun scoring

We next compute an overall score for the target noun. As described in Section 3.1, for a specific target noun n and task, we sample 500 sentences $S_1, ..., S_{500}$. The noun's score for this task is then given by:

$$\text{Score}_{\text{noun}}(n) = \frac{1}{500} \sum_{i=1}^{500} \text{Score}_{\text{sent}}(S_i) \quad (2)$$

3.5 Word filtering and tokenization

Words were removed from a particular model if either their singular or plural form was tokenized to *unk*, or if their singular and plural forms were assigned different numbers of tokens.[2] For BERT, words in the Verb set were removed if they were assigned more than one token, as BERT does not model the joint distribution over multiple masked tokens.

For Transformer-XL, we add a padding text[3] and a start-of-sentence-token ($\langle \text{SOS} \rangle$) to the beginning of the sentence and an end-of-sentence token ($\langle \text{EOS} \rangle$) to the end of the sentence. For GPT-2, we make no modifications to the generated sentence (although prefix spaces are added to the strings for tokenization purposes). For BERT, since it is a masked language model, we replace the Verb (for SVA) or reflexive pronoun (for RA) with a [MASK] token after tokenization. Thus, each sentence will have a single mask token corresponding to the word that should agree with the target noun.

4 Results

4.1 Noun performance is correlated across tasks

We first examine how each noun's performance varies across the grammatical tasks. For each noun-task pair, we measure the average performance of the noun on that task, as described above. This gives 10 features per noun, corresponding to the 10 grammatical tasks.

Figure 1 shows the pairwise comparisons between performance on the different tasks for Transformer-XL. Results for BERT and GPT-2 are similar and are shown in the appendix. The figure shows that performance is correlated across the tasks; for many pairs of tasks, nouns which have higher performance on one task are likely to have higher performance on the other.

Using principal component analysis, we found that a single principal component explains 47% of task variance for Transformer-XL, and two principal components explain 73%. Results are similar for BERT and GPT-2, and are shown in the appendix. The first PC primarily measures performance on the four reflexive anaphora tasks, while the second PC measures performance on the subject-verb agreement across relative clause tasks.

---

[2]The latter constraint was used in order to simplify batching.
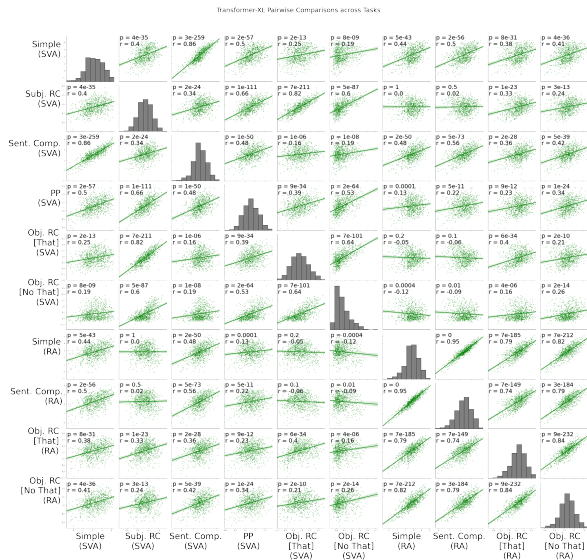
[3]https://tinyurl.com/y9kjuj5q

Figure 1: Pairwise comparisons between tasks with Transformer-XL. Rows and columns represent tasks, and one point represents a single noun's performance on a pair of tasks. The four tasks on the lower right, with strongest correlations, all involve reflexive anaphora.

This suggests that there is a dimension that characterizes whether the model understands how reflexive binding constraints operate for a noun, and a dimension for whether the model understands subject-verb agreement for the noun. Note that Figure 1 additionally demonstrates correlations between the reflexive tasks and the subject-verb agreement tasks.

These results provide evidence that language models' variation in performance on the grammatical tasks is, in part, explained by properties of the nouns which are stable across tasks. The models understand number agreement better for some nouns, and worse for others.

4.2 Noun performance is correlated across models

We next investigate whether nouns exhibit stable behavior across different neural language models. For each pair of the three language models, we measured how well a noun's task performance in one language model predicted its task performance in the other language model.

Figure 2 shows comparisons between pairs of language models on the 10 grammatical tasks. Of the 30 comparisons, 24 show significant positive correlations between the pairs of language models. 22 of the correlations remain significant after Bonferroni correction.

GPT-2 and Transformer-XL show the strongest correlation in performance. It is possible that this is due to methodological differences between the task setup for GPT-2 and Transformer-XL compared to BERT: GPT-2 and Transformer-XL are performing a language modeling task in which the probability of a full sentence is queried, while BERT performs masked language modeling on a single target word. The difference may also be due to corresponding training differences between BERT and the autoregressive language models.

The results provide evidence that nouns exhibit stable task performance across language models. The source of the correlation across language models must come from features of the training data. Properties of the natural text distribution of nouns lead some of these nouns to be better understood than others.

4.3 Effect of frequency on task performance

In Sections 4.1 and 4.2, we found evidence that nouns exhibit stable performance across different grammatical tasks and language models. One obvious explanation of these results is that nouns vary in their frequency in natural text, and language models learn more accurate grammatical representations for more frequent nouns.

In order to investigate this, we measured the frequency of each noun in two corpora: WikiText-103, a 103 million token subset of Wikipedia, which was used for training Transformer-XL; and Open WebText (Gokaslan and Cohen, 2019), an open-source implementation of the web corpus used to train GPT-2.[4] Word frequencies were measured separately for singular and plural noun forms. Figure 3 shows the relationship between frequency and task performance on each of the ten grammatical tasks. The appendix shows the results broken down by task type.

The results show no clear relationship between noun frequency and task performance. Frequency explains no more than 0.1% of the variation in performance. This holds true over more than four orders of magnitude in frequency. This provides evidence that 1) differences in corpus frequency do not explain the systematic differences observed between nouns, and 2) relatively few observations suffice for transformer language models to learn

---

[4]BERT was trained on a mix of Wikipedia text and Book-Corpus. Because, as of this writing, BookCorpus is no longer distributed, WikiText-103 was used as a proxy for BERT training frequencies.
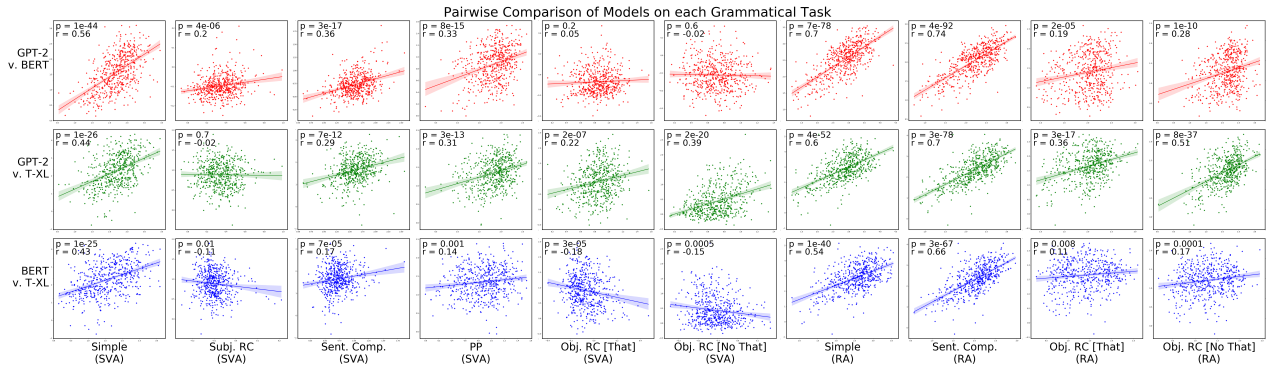
Figure 2: Pairwise comparisons between GPT-2, Transformer-XL, and BERT on the 10 grammatical tasks. Each row corresponds to a pair of language models, and each column is a single task. One point represents the performance of a noun on a single task.
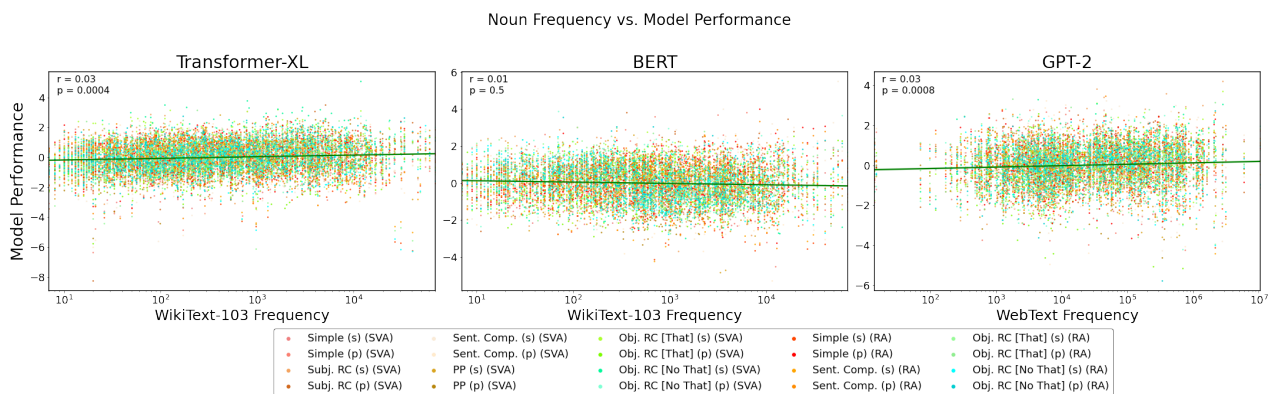

Figure 3: Relationship between corpus frequency and task performance for Transformer-XL, BERT, and GPT-2. Performance scores are z-normalized. Colors indicate the ten grammatical tasks and singular/plural form of the noun (*s* indicates singular, *p* indicates plural). Each point represents task performance for a single noun.

correct number agreement behavior for a noun. In the next section, we investigate this finding further.

5 Few-shot learning for novel lexical items

The results in the previous section provide evidence that nouns systematically vary in their performance on grammatical tasks; some nouns perform better than others across tasks and language models. However, this variation is not explained by frequency of occurrence in natural text. Nouns that occur on the order of 100 times in a corpus do not have systematically worse performance than nouns that occur 10^6 times.

The results raise a question: if frequency does not influence how well a noun is understood, what does? If low frequency nouns are understood as well as higher frequency nouns, then this suggests that language models few-shot learn the grammatical properties of nouns. We suggest that by study-

ing what makes a noun learnable in a few-shot setting, it may be psosible to better understand the sources of the observed variation.

We use a few-shot learning paradigm, introducing a new lexical item into the vocabulary of the language model, either "wug" (intended as a new singular noun), or "wuz" (intended as a plural). We then fine-tune the language model using several example sentences containing this word. Note that this paradigm is distinct from nearly all of the few-shot learning experiments performed in Radford et al. (2019); Brown et al. (2020), which operate on a known vocabulary.[5]

5.1 Learning agreement from syntactic data

We first look at whether training data containing explicit syntactic markers of number agreement is sufficient for few-shot learning. Table 3 describes

---

[5]Brown et al. (2020) perform several experiments on novel vocabulary items.

Training data type	Template
Simple	The wug/wuz ⟨PresentTenseVerb⟩.
Pred-adj	The wug/wuz is/are ⟨Adj⟩.
Reflexive	The wug/wuz ⟨Verb⟩ himself/themselves.

Table 3: The three types of training data used for syntactic fine-tuning.

the types of training data we examine. The three types of training data use different syntactic markers of plurality to indicate whether the new noun is singular or plural.

The language models are fine-tuned with 5 sentences drawn from a single training data type. GPT-2 was fine-tuned for 2 epochs, and BERT was fine-tuned for 4 epochs.[6] Transformer-XL was not used for the fine-tuning experiments, due to issues with introducing new vocabulary items given Transformer-XL's adaptive weight embedding.

After fine-tuning, each model was evaluated on the 10 grammatical tasks in Table 1. For each grammatical task, 500 sentences were sampled from the task template, and a performance score was calculated by averaging scores of the samples, as described in Section 3.4.

Figure 4 shows results for fine-tuning on the three types of syntactic data. Compared to model performance on real lexical items (shown in the leftmost column), both BERT and GPT-2 achieve qualitatively similar performance given the Pred-adj and Reflexive training data, but worse performance given the Simple training data. Performance is weakest on subject-verb agreement (SV-agreement) tasks involving relative clauses. When trained on data containing reflexive anaphora, both models achieve notably higher performance on the grammatical tasks involving reflexive anaphora.

The results provide evidence that small amounts of syntactic training data support learning the agreement properties of novel nouns. They also provide evidence of heterogeneity among different types of training data. Training from bare present tense verbs is least effective, and training from sentences containing reflexives leads to improved performance on tasks which require understanding of the conditions on reflexive binding.

5.2 Learning agreement from semantic data

We next examine whether purely semantic indicators of plurality are sufficient for learning a noun's

---

[6] Prior to more systematic experiments, we informally optimized the number of fine-tuning epochs.
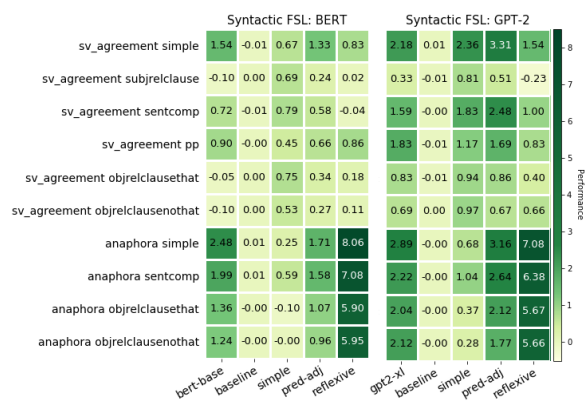

Figure 4: Few-shot learning from syntactic examples (averaging over plural and singular results). Columns show different types of training data, and rows show the 10 grammatical tasks. The bert-base and gpt2-xl columns indicate model performance on known lexical items, i.e. summarizing results from Section 4. The baseline columns indicate performance of non-fine-tuned models on the novel wug/wuz lexical items. Scores are differences of log-probabilities between grammatical and ungrammatical. The 95% confidence interval around each point estimate is always smaller than ±0.25.

number agreement properties. We look at several types of constructions which provide information about the plurality of a noun, but using predicates with past tense verbs that don't inflect for number so that there is no grammatical number agreement. In particular, we note the different possible readings with reference to the distributive and collective distinction described in the semantics literature (Lønning, 1997; Lasersohn, 2011; Champollion, 2015). For documentation of predicates that require a collective NP subject, see Levin (1993).

We use the fine-tuning method from Section 5.1.

Singular constructions

In order to induce singular noun interpretations, we use the singular-biased constructions shown at the top of Table 4. For example, if a wug worked all alone or came unaccompanied, it is likely that "wug" is both semantically and grammatically singular. However, these constructions do not gramatically require the head noun to be singular: they are compatible with distributive readings where the predicate individually applies to members of a group (e.g. "the lawyers worked all alone" means *each* lawyer worked alone).

BERT and GPT-2 were fine-tuned on 5 examples of each of the singular constructions. Figure 5 shows the results. None of the constructions con-

	Training data type	Example
Singular	all-alone	The wug worked all alone.
	unaccompanied	The wug came unaccompanied.
	separated-entire	The wug became separated from the entire group.
	personally	The wug personally thanked me.
Plural	unison	The wuz nodded in unison.
	together	The wuz ate together.
	simultaneously	The wuz jumped simultaneously.
	outnumbered	The wuz outnumbered the cats.
	constituted	The wuz constituted a majority of the team.
	gathered	The wuz gathered quietly.

Table 4: Types of training data used for semantic fine-tuning.

Singular Semantic FSL: BERT

	bert-base	baseline	all-alone	unaccompanied	separated-entire	personally
sv_agreement simple	0.58	-0.68	-0.53	0.65	1.80	0.09
sv_agreement subjrelclause	-1.41	-1.32	-1.40	-1.95	-1.64	-1.36
sv_agreement sentcomp	-0.75	-0.96	-1.07	-1.48	-0.96	-0.96
sv_agreement pp	0.40	0.30	-0.34	0.26	-0.01	0.02
sv_agreement objrelclausethat	-0.78	-0.64	-1.01	-1.49	-1.62	-0.91
sv_agreement objrelclausenothat	-0.51	-0.45	-0.47	-1.14	-1.08	-0.51
anaphora simple	3.07	1.45	3.17	3.55	1.26	3.47
anaphora sentcomp	1.98	0.61	1.54	2.19	0.38	1.86
anaphora objrelclausethat	1.61	0.54	1.51	1.98	-0.57	1.61
anaphora objrelclausenothat	1.52	0.48	1.54	2.25	-0.47	1.79

Singular Semantic FSL: GPT-2

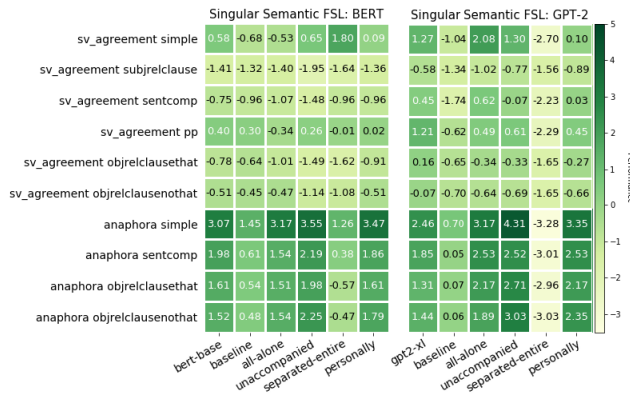|	gpt2-xl	baseline	all-alone	unaccompanied	separated-entire	personally
sv_agreement simple	1.27	-1.04	2.08	1.30	-2.70	0.10
sv_agreement subjrelclause	-0.58	-1.34	-1.02	-0.77	-1.56	-0.89
sv_agreement sentcomp	0.45	-1.74	0.62	-0.07	-2.23	0.03
sv_agreement pp	1.21	-0.62	0.49	0.61	-2.29	0.45
sv_agreement objrelclausethat	0.16	-0.65	-0.34	-0.33	-1.65	-0.27
sv_agreement objrelclausenothat	-0.07	-0.70	-0.64	-0.69	-1.65	-0.66
anaphora simple	2.46	0.70	3.17	4.31	-3.28	3.35
anaphora sentcomp	1.85	0.05	2.53	2.52	-3.01	2.53
anaphora objrelclausethat	1.31	0.07	2.17	2.71	-2.96	2.17
anaphora objrelclausenothat	1.44	0.06	1.89	3.03	-3.03	2.35

Figure 5: Few-shot learning from singular semantic examples. The bert-base and gpt2-xl columns indicate model performance on known lexical items, i.e. summarizing results from Section 4. The baseline columns indicate performance of non-fine-tuned models on the novel wug token.

sistently induced correct performance on the grammatical tasks across both models. Three of the constructions — *all-alone*, *unaccompanied*, and *personally* — led to strong performance on the reflexive anaphora tasks (stronger than the average performance calculated in Section 4). The *separated-entire* construction consistently decreased performance on the tasks relative to baseline.

Plural constructions

In order to provide the models with data indicating that a novel noun is plural, we use constructions which force either collective or distributive readings. For example, in Table 4, if the wuz constituted the majority of the team, then the word "wuz" must be semantically plural. The construction *constituted a majority* is collective because it must apply to the group as a whole:

(6) The doctors constituted a majority of the team.
    a. *Distributive reading: each of the doctors constituted a majority.

Plural Semantic FSL: BERT

	bert-base	baseline	unison	together	simultaneously	outnumbered	constituted	gathered
sv_agreement simple	2.51	0.65	-0.92	1.45	0.24	-0.63	-0.37	0.84
sv_agreement subjrelclause	1.21	1.32	1.25	1.33	1.21	0.99	1.39	1.04
sv_agreement sentcomp	2.19	0.94	0.83	1.02	1.01	0.67	0.97	0.81
sv_agreement pp	1.40	-0.30	-0.43	0.73	-0.10	-0.61	0.04	-0.02
sv_agreement objrelclausethat	0.68	0.65	0.77	1.05	0.81	0.43	0.88	0.73
sv_agreement objrelclausenothat	0.32	0.46	0.39	0.57	0.49	0.26	0.46	0.34
anaphora simple	1.89	-1.43	3.91	4.19	2.92	0.21	1.31	2.69
anaphora sentcomp	2.00	-0.59	3.32	4.16	2.72	0.97	1.72	1.88
anaphora objrelclausethat	1.11	-0.54	1.87	3.19	1.43	0.65	0.66	1.24
anaphora objrelclausenothat	0.97	-0.48	2.16	3.04	1.52	0.60	0.89	1.29

Plural Semantic FSL: GPT-2

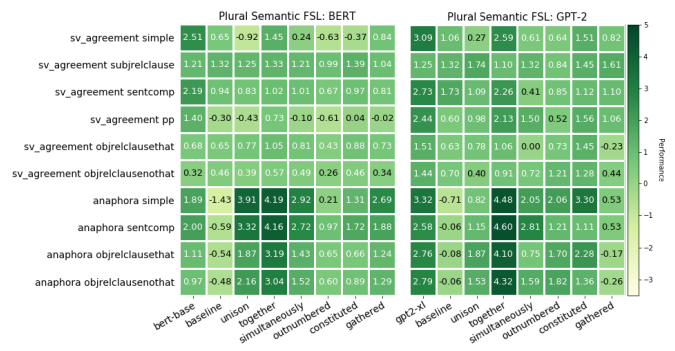|	gpt2-xl	baseline	unison	together	simultaneously	outnumbered	constituted	gathered
sv_agreement simple	3.09	1.06	0.27	2.59	0.61	0.64	1.51	0.82
sv_agreement subjrelclause	1.25	1.32	1.74	1.10	1.32	0.84	1.45	1.61
sv_agreement sentcomp	2.73	1.73	1.09	2.26	0.41	0.85	1.12	1.10
sv_agreement pp	2.44	0.60	0.98	2.13	1.50	0.52	1.56	1.06
sv_agreement objrelclausethat	1.51	0.63	0.78	1.06	0.00	0.73	1.45	-0.23
sv_agreement objrelclausenothat	1.44	0.70	0.40	0.91	0.72	1.21	1.28	0.44
anaphora simple	3.32	-0.71	0.82	4.48	2.05	2.06	3.30	0.53
anaphora sentcomp	2.58	-0.06	1.15	4.60	2.81	1.11	1.11	0.53
anaphora objrelclausethat	2.76	-0.08	1.87	4.10	0.75	1.70	2.28	-0.17
anaphora objrelclausenothat	2.79	-0.06	1.53	4.32	1.59	1.82	1.36	-0.26

Figure 6: Few-shot learning from plural semantic examples. The baseline columns indicate performance of non-fine-tuned models on the novel wuz token.

    b. Collective reading: the doctors as a group constituted a majority.

While the argument of a collective predicate must be semantically plural, it is not necessarily grammatically plural. For example, the singular "the group" could constitute the majority of the team.

Three of the constructions in Table 4 are collective: *outnumbered*, *constituted*, and *gathered*. The other three are distributive phrasal predicates, which force distributive readings:

(7) The architects nodded in unison.
    a. Distributive reading: each of the architects nodded.
    b. *Collective reading: the group of architects itself nodded.

Figure 6 shows the plural learning results. The 6 types of training data perform comparably on the subject-verb agreement tasks (and similar to the baseline model, which represents performance prior to fine-tuning). The three distributive phrasal constructions perform better on the reflexive anaphora tasks than the three collective constructions, though all constructions improve relative to the baseline.

6 Discussion

We have investigated the sources of variation in neural language models' grammatical judgments. We found that there are systematic differences between nouns: when a language model exhibits knowledge of a noun's grammatical properties in one task, it is more likely to do so in other tasks. Moreover, when one language model exhibits this knowledge, other language models are more likely to as well.

The study found two latent dimensions of variation between nouns: one corresponding to how well the models understood its behavior with reflexive pronouns, and the other corresponding to subject-verb agreement.

Subsequent analyses demonstrate a pair of empirical phenomena:

1. It is relatively easy to learn the number agreement properties of a noun. The models learn the agreement properties of a novel noun from just a few samples, and the data supporting few-shot learning appears to be densely distributed; nearly all types of syntactic and semantic data examined lead to improvements on the reflexive pronoun or subject-verb agreement tasks.

2. Nouns that occur more frequently during training are not learned more accurately. Many nouns that occur with high frequency are not learned accurately.

These results suggest that nouns should vary less in their grammatical performance than is actually observed; the study finds *excess variation* in grammatical performance. If number agreement can be correctly learned from a few samples (FSL samples), then one would expect model performance to either a) improve with more data, as more FSL samples are observed, or b) improve with more data up to some threshold, and then asymptote after learning has saturated. In either case, for high frequency nouns, a sufficient number of FSL samples should be observed for these nouns to be learned very accurately.

A potential explanation of the results is that they are caused by catastrophic forgetting (Ratcliff, 1990; French, 1999): although a sufficient number of FSL samples are observed for a noun, these samples are forgotten during training, causing the performance of the noun to degrade. This explanation is implausible. If catastrophic forgetting is occurring, then the problem should be more severe for infrequent nouns than for frequent nouns, as the interval between training samples will be longer for infrequent nouns. This would predict better performance for frequent nouns.

Acknowledgements

References

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Lucas Champollion. 2015. Distributivity, collectivity and cumulativity. In L. Matthewson, C. Meier, H. Rullmann, and T. E. Zimmermann, editors, *Companion to Semantics*. Wiley-Blackwell.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

C. Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

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. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus.

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

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, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. 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, Melbourne, Australia. Association for Computational Linguistics.

Peter Nathan Lasersohn. 2011. Mass nouns and plurals. In *Semantics: An International Handbook of Natural Language Meaning*, pages 1131–1153. De Gruyter.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Michael A Lepori, Tal Linzen, and R Thomas McCoy. 2020. Representations of syntax [MASK] useful: Effects of constituency and dependency structure in recursive LSTMs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, Washington. Association for Computational Linguistics.

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

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.

Jan Tore Lønning. 1997. Plurals and collectivity. In *Handbook of logic and language*, pages 1009–1053. Elsevier.

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*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Roger Ratcliff. 1990. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285.

Karin Kipper Schuler. 2005. Verbnet: A broad-coverage, comprehensive verb lexicon.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019a. Investigating BERT's knowledge of language: Five analysis methods with NPIs. 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 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019b. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

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*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 3302–3312, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

A Further analyses

This section contains several additional analyses: principal components for task performance among

the three language models (Tables 5-8); pairwise comparison between task performance for BERT and GPT-2 (Figures 7 and 8); and more fine-grained comparisons between word frequency and model performance (Figures 9 and 10).

PC Number	Transformer-XL	BERT	GPT-2
1	0.4663	0.3865	0.4146
2	0.7299	0.5619	0.6873
3	0.8511	0.7073	0.8059
4	0.9083	0.8034	0.8720
5	0.9499	0.8919	0.9175

Table 5: Cumulative proportion of variance explained by the top (of 10) PCs for each model as detailed in Section 4.1.

Contributor by Rank	PC 1	PC 2	PC 3
1	RA ObjRelClauseNoThat - 0.386980	SV SubjRelClause - 0.449504	SV SentComp - 0.534710
2	RA ObjRelClauseThat - 0.376354	SV ObjRelClauseNoThat - 0.442445	SV Simple - 0.516031
3	RA SentComp - 0.359096	SV ObjRelClauseThat - 0.439015	RA ObjRelClauseThat - 0.402452
4	RA Simple - 0.347978	RA Simple - 0.376192	SV ObjRelClauseThat - 0.312466

Table 6: Top contributors (tasks) to top few (of 10) PCs for Transformer-XL's noun performance as detailed in Section 4.1. Cells contain the task name followed by their (absolute) component value in the eigenvector.

Contributor	PC 1	PC 2	PC 3
1	RA ObjRelClauseNoThat - 0.456096	SV ObjRelClauseThat - 0.577452	SV SentComp - 0.686402
2	RA ObjRelClauseThat - 0.444272	SV ObjRelClauseNoThat - 0.576572	SV Simple - 0.499513
3	RA Simple - 0.383953	SV PP - 0.353213	SV SubjRelClause - 0.346437
4	RA SentComp - 0.383866	RA Simple - 0.288091	SV PP - 0.248977

Table 7: Top contributors (tasks) to top few (of 10) PCs for BERT's noun performance as detailed in Section 4.1. Cells contain the task name followed by their (absolute) component value in the eigenvector.

Contributor	PC 1	PC 2	PC 3
1	RA ObjRelClauseNoThat - 0.454492	SV ObjRelClauseNoThat - 0.477923	SV SentComp - 0.549969
2	RA Simple - 0.447148	SV SubjRelClause - 0.444648	SV Simple - 0.525072
3	RA SentComp - 0.441366	SV ObjRelClauseThat - 0.426385	SV ObjRelClauseThat - 0.478782
4	RA ObjRelClauseThat - 0.425359	SV SentComp - 0.385486	SV ObjRelClauseNoThat - 0.362165

Table 8: Top contributors (tasks) to top few (of 10) PCs for GPT-2's noun performance as detailed in Section 4.1. Cells contain the task name followed by their (absolute) component value in the eigenvector.
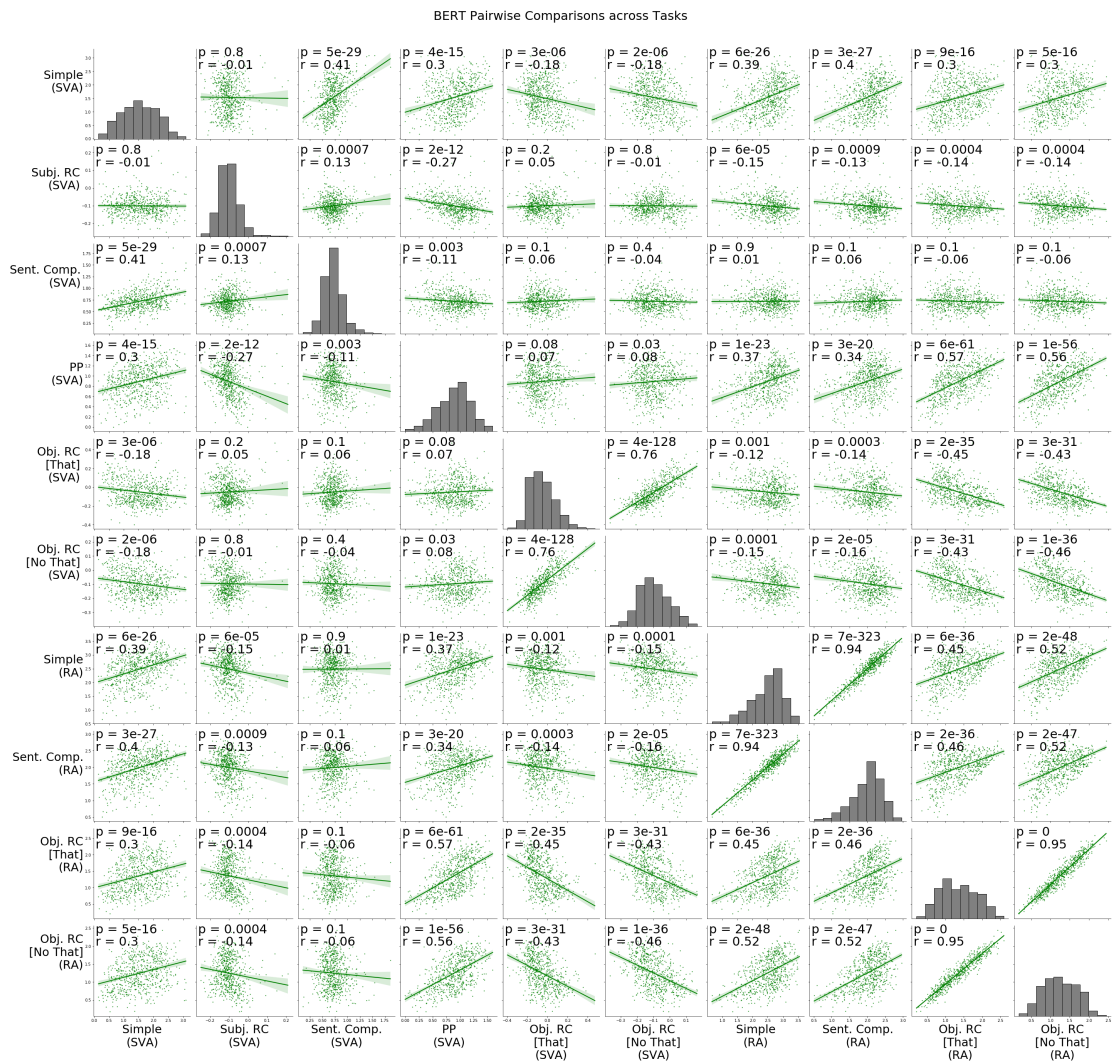
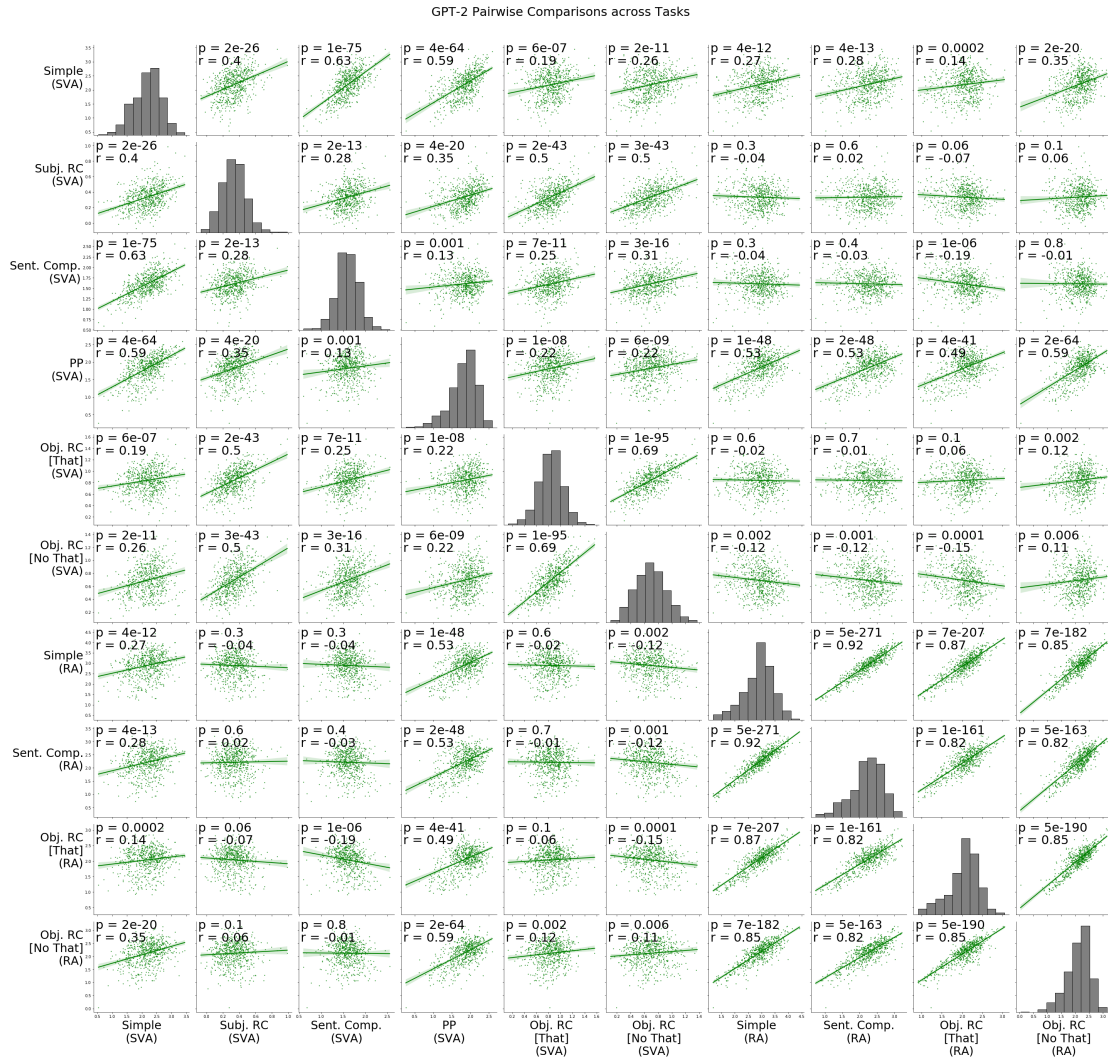Figure 7: BERT: Pairwise comparisons between tasks.

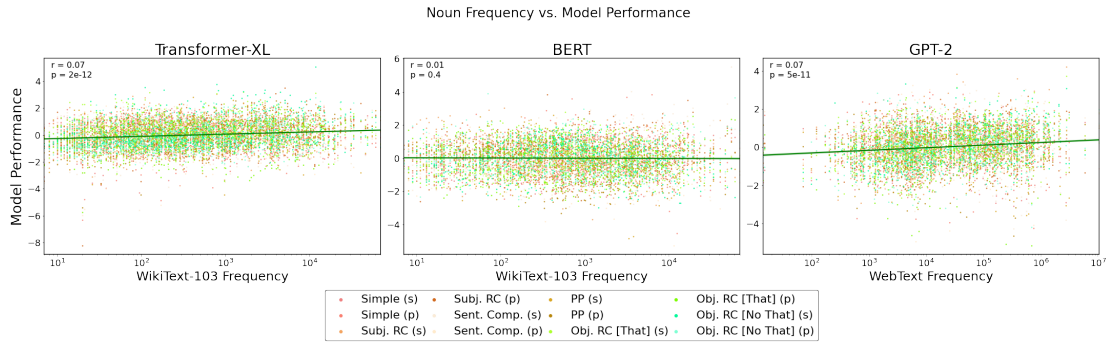Figure 8: GPT-2: Pairwise comparisons between tasks.

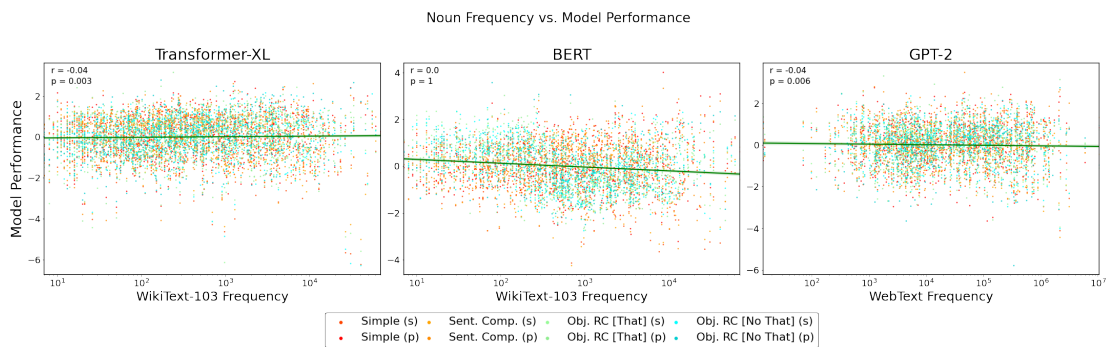Figure 9: Noun Frequency vs. Model Performance on Subject-Verb Agreement Tasks

Figure 10: Noun Frequency vs. Model Performance on Reflexive Anaphora Tasks