Frequency Effects on Syntactic Rule Learning in Transformers

Jason Wei[1] **Dan Garrette**[1] **Tal Linzen**[2*] **Ellie Pavlick**[1,3]
[1]Google Research [2]New York University [3]Brown University
{jasonwei,dhgarrette,linzen,epavlick}@google.com

Abstract

Pre-trained language models perform well on a variety of linguistic tasks that require symbolic reasoning, raising the question of whether such models implicitly represent abstract symbols and rules. We investigate this question using the case study of BERT's performance on English subject–verb agreement. Unlike prior work, we train multiple instances of BERT from scratch, allowing us to perform a series of controlled interventions at pre-training time. We show that BERT often generalizes well to subject–verb pairs that never occurred in training, suggesting a degree of rule-governed behavior. We also find, however, that performance is heavily influenced by word frequency, with experiments showing that both the absolute frequency of a verb form, as well as the frequency relative to the alternate inflection, are causally implicated in the predictions BERT makes at inference time. Closer analysis of these frequency effects reveals that BERT's behavior is consistent with a system that correctly applies the SVA rule in general but struggles to overcome strong training priors and to estimate agreement features (singular vs. plural) on infrequent lexical items.

1 Introduction

Many natural language phenomena are best described as the product of applying rules to abstract symbols, without access to the content of these symbols (Smolensky, 1988; Fodor and Pylyshyn, 1988). Most speakers of English will agree, for example, that if "*gorp*" is a singular noun, then, regardless of the meaning of "*gorp*", the utterance "*the gorp* ***adds*** *nothing*" is grammatical, but "*the gorp* ***add*** *nothing*" is not.

The success of contemporary neural language models such as BERT (Devlin et al., 2019) on language understanding tasks, as well as in more targeted linguistic evaluations (Marvin and Linzen, 2018; Goldberg, 2019), raises the question of whether these systems acquire such symbolic rules. While previous studies have attempted to address such questions, particularly in relation to BERT (Rogers et al., 2020), prior work has generally not analyzed the relationship between the model's pre-training data and its behavior. As a result, it has been difficult to tease apart the many factors that may influence a model's test time performance.

In this paper, we investigate whether pre-trained transformer-based language models learn and apply symbolic rules, focusing on BERT's ability to follow the English subject–verb number agreement rule (§3) as a case study. On our evaluation stimuli (§4), we find that BERT achieves high performance, even on subject–verb pairs that never occurred together in the training set (§5.1–§5.2). In exploratory data analysis, however, we find that this performance is also influenced by effects from both absolute and relative frequency of verb forms in the training data (§5.3–§5.4). To confirm these phenomena causally, we perform a series of training interventions where we pre-train BERT models on training data for which we have carefully manipulated the frequencies of verb forms (§6). We further use probing classifiers to attribute observed mistakes either to errors in rule-following or to errors in lexical categorization (§7).

These experiments reveal several insights about BERT in the context of rule-governed tasks. First, the high performance of BERT on subject–verb combinations that never occurred in the training set is consistent with a model that learns abstract representations of lexical items and patterns, i.e., abstract features and rules. Second, BERT's performance is influenced by absolute frequency effects, but probing classifiers show that this influence can be explained by the model's inability to learn the features of a verb form (singular vs. plural) for infrequent lexical items, rather than a failure to apply the rule when the verb form has been classified. Fi-

[*] Work done while visiting Google.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 932–948
November 7–11, 2021. ©2021 Association for Computational Linguistics

nally, although BERT generally applies rules with high accuracy, it fails to overcome strong priors during training—when one verb form is much more frequent than another, BERT tends to produce the more common form, even when it is not consistent with the rule.

2 Experimental Logic

2.1 Hypotheses

We aim to investigate BERT's ability to reason over abstract symbols. As a case study, we focus on subject–verb agreement (SVA) in English, for which the grammaticality rule of interest is:

$$\textsc{number}(\textit{subject}) = \textsc{number}(\textit{verb})$$

We consider three alternative hypotheses about the process underlying BERT's behavior on SVA.

H1: Idealized Symbolic Learner. In theory, symbolic reasoners operate over abstract categories, such as the *agreement feature* NUMBER, and rules, such as *"if* NUMBER(*subject*) = SINGULAR, *then* NUMBER(*verb*) = SINGULAR." Early work (Fodor and Pylyshyn, 1988) which discusses the behavior of such symbolic systems often presents an idealized version, for the sake of theoretical argument. Thus, under H1, this system would not make errors such as misclassifying inputs or erroneously parsing the sentence, and is not affected by word-specific properties (e.g., frequency).

H2: Item-Specific Learner. The antithesis of the idealized symbolic learner is a model that reasons entirely using word co-occurrences. This system does not represent any abstractions over the immediate inputs it receives, and thus cannot reason over features such as singular/plural. Conceptually, it is analogous to early phrase-based MT systems (Brown et al., 1990) that build a literal string lookup table in order to predict the most likely output given an input. By definition, it performs poorly on noun–verb pairs that never co-occurred in training, as the lookup table will not have the relevant entry.[1]

H3: Symbolic Learner with Noisy Observations. Both H1 and H2 represent extreme, largely theoretical models of system behavior. In practice, we expect systems like BERT to display some hybrid of the two. However, to our knowledge, there has been no work to date which proposes a specific hypothesis of what type of hybridization best explains BERT's behavior. In this work, we consider one such hybrid: a system that is symbolic at its core but has noisy observations.[2] That is, under H3, the system represents symbols (e.g. singular/plural word categories) and rules (e.g., SVA) correctly, but can make errors in mapping from inputs to symbols. Conceptually, it is analogous to a BayesNet (Pearl, 1988) that correctly represents nodes and causal connections internally but may nonetheless incorrectly process an input, activating the wrong nodes and thus producing the wrong output. Thus, unlike in H1, systems consistent with H3 make errors when they cannot identify whether a subject or verb is singular or plural, potentially due to frequency effects (present at all levels of processing; Marantz, 2013).

2.2 Predictions and Summary of Findings

We use three diagnostics to differentiate the above hypotheses: (1) generalization to unseen noun-verb pairs, (2) the presence of frequency effects when making predictions for seen noun-verb pairs, (3) and correlation between specific types of errors.

Generalization to unseen noun-verb pairs allows us to differentiate H2 from H1 and H3. For instance, since whether the sentence *"the section adds nothing"* obeys the SVA rule depends only on NUMBER(*"section"*) and NUMBER(*"adds"*), a symbolic reasoner's ability to assess grammaticality should not depend on how frequently the words *"section"* and *"adds"* have been seen together in the data. Instead, we would expect such a system to learn the correct agreement features of the two words independently and apply a general SVA rule to them. In contrast, an item-specific learner, which does not represent abstract agreement features, would rely on probabilities defined over specific lexical items, and thus may fail to reason correctly about rare or unseen situations, for which such probabilities are poorly calibrated.

The **presence of frequency effects** in BERT's performance allows us to differentiate H1 from H2

[1] We do not specify whether such a model has access to abstract features other than agreement because such features (e.g., the notion of subject) do not affect the specific hypotheses we consider. For example, a model that does not represent agreement feature and only learns word co-occurrences will perform poorly on unseen items, regardless of whether it has access to correct parses.

[2] Here, "observations" involves both the parser as well as the lexicon. I.e., H3 allows for errors to arise due to incorrect lexical entries and/or incorrect parses. However, since our experiments (§7) don't differentiate lexicon errors from parse errors, we do not differentiate them within this hypothesis. Future work that differentiates these two errors sources could be worthwhile.

and H3. That is, under both H2 and H3, the model may perform worse on less frequent words (albeit for different reasons). In contrast, a system consistent with H1 should not exhibit any differences in performance due to differences in inputs below the abstraction of singular/plural.

Our experiments show that BERT generalizes well (though not perfectly) to unseen noun-verb pairs (§5.1–§5.2) and exhibits clear frequency effects (§6). Together, these results are most consistent with a hybrid system like H3. To confirm this, we use probing classifiers to investigate H3's specific prediction about **correlations between types of errors**, i.e., that errors on SVA should be explained by errors in classifying singular vs. plural (§7). We find that the expected error patterns explain some frequency effects (those due to absolute frequency) but not others (those due to relative frequency). Thus, we ultimately conclude that, of the hypotheses considered, H3 is the best model of BERT's behavior, though BERT exhibits additional sensitivity to frequency imbalances between competing word forms that H3 leaves underspecified.

3 Related Work

Targeted Syntactic Evaluation. We use the targeted syntactic evaluation framework of Linzen et al. (2016) and Marvin and Linzen (2018) to measure the model's ability to learn and apply the SVA rule. Following the setup from Goldberg (2019), each test instance consists of a sentence in which a verb has been masked out, and BERT's masked language modeling (MLM) parameters are used to score whether the singular or plural form of the verb is a better fit for the masked position. For example, given the sentence *"The section `[MASK]` nothing to the info."* and set of verb inflections {*"add"*, *"adds"*}, the model would be considered correct if the MLM prediction assigns a higher score to the singular form *"adds"* than the plural form *"add"* since the subject of the masked verb position is *"section,"* which is singular.

Due to the particulars of BERT's MLM task setup, the model is only able to score words that are represented by a single wordpiece. While Goldberg (2019) dealt with this limitation by restricting evaluation to just those verbs that appear in the original BERT model's vocabulary as a single wordpiece, we are able to avoid such compromises because pre-training the models ourselves means that we can add any entries we want to the vocabulary.

Syntactic Reasoning in LMs. There has been substantial prior work on the ability of language models to perform abstract syntactic processing tasks (Hu et al., 2020) (see Linzen and Baroni (2020) for a review). On SVA specifically, Goldberg (2019) found that BERT achieves high accuracy on both natural sentences (97%) and nonce sentences (83%), and that error rate was independent of the number of "distractor" words between the subject and verb; Yu et al. (2020) showed that language models do not exhibit better grammatical knowledge of more frequent nouns. Other work has found that BERT's performance is sensitive to factors that may suggest item-specific learning; Chaves and Richter (2021) found that BERT's performance on number agreement is sensitive to the verb, across seven different verbs, and Newman et al. (2021) found that language models performed better on verbs that they predicted were likely in context. The focus on frequency effects also relates to a more general line of work on understanding the effect of training size and distribution on neural language models' generalization (Warstadt et al., 2020; Lovering et al., 2021). To our knowledge, our present study is the first to investigate these questions via controlled interventions on the model's pre-training data, making it possible to draw stronger conclusions.

Our formulation of the SVA task also relates to work which investigates neural networks' abilities to learn lexical abstractions (Chronis and Erk, 2020; Kim and Smolensky, 2021) and to reason systematically (Lake and Baroni, 2018; Yanaka et al., 2019; Kim and Linzen, 2020; Goodwin et al., 2020). These studies on systematicity, however, run controlled experiments by training small models on toy data. Our work studies the widely-used BERT model, trained on real data and at scale.

4 Experimental Setup

4.1 Model

Differentiating between the hypotheses presented in §2 requires analyzing model performance on individual items as a function of frequencies in the training data. The original BERT model was trained on both English Wikipedia and BooksCorpus (Zhu et al., 2015). However, BooksCorpus is not publicly available (Bandy and Vincent, 2021), so when we pre-train our BERT models, we use only the Wikipedia data (2.3 billion tokens). Despite this difference in training data, our models

Natural	**Addition** of such minor characters **seem/seems** more promotional than encyclopedic.
	Other popular trade **items** of the area **include/includes** sandalwood, rubber, and teak.
	The **party** that originally buys the securities effectively **act/acts** as a lender.
Nonce	The **astronomer** of the first session of the court during that year **perform/performs** a...
	The **isometry** in the gulf **market/markets** santa catalina island.
	The **sheepdog** of basic needs providers ... **review/reviews** a damaging effect.

Table 1: Examples of natural and nonce stimuli. Target verbs and their subjects are bolded. The model takes as input the sentence with the verb masked, and is evaluated based on which verb inflection it scores more highly.

achieve performances comparable to the public BERT-Base release on GLUE (Wang et al., 2018) (see Appendix A).

4.2 Evaluation Stimuli

We evaluate the model's SVA ability on two classes of stimuli: (1) *natural* sentences, which are generally both syntactically and semantically coherent, and (2) *nonce* sentences (following Gulordava et al. 2018) which are grammatically valid but not necessarily semantically coherent ("*colorless green ideas sleep furiously*", Chomsky, 1956). Examples of each are shown in Table 1.

Evaluating on natural sentences provides a measurement of how well the model can be expected to perform in realistic settings, but these sentences are not ideal for a targeted SVA evaluation since they often contain additional cues relevant to verb inflection, such as other plural verbs or plural determiners, as in "*two* `[SUBJECT]` *and their dogs* `[VERB]`," making it difficult to discern whether a model has chosen a particular verb inflection based on the subject. In contrast, performance on synthetic nonce sentences allows us to ensure that the only source of information about the verb's correct inflection is the subject itself.

Natural Stimuli. Following Goldberg (2019), for natural stimuli we use the dataset from Linzen et al. (2016), which comprises 23,298 sentences from Wikipedia. The target verb is plural in 16,232 of these sentences and singular in 7,064 of these sentences. These evaluation sentences span 176 verb lemmas and 329 verb forms.

Nonce Stimuli. For our nonce stimuli, we compiled a list of 200 nouns, 336 verbs, and 56 *sentential contexts*—sentence templates where we remove the original subject and verb—such that any given (noun, verb, sentential context) triplet yields a grammatically correct nonce sentence. E.g., given the sentence "*the investigation of chaperones has a long history*", we can create a sentential context: "*the* `[SUBJECT]` *of chaperones* `[VERB]` *a long history.*" We can then randomly chose a noun and verb from our noun and verb lists (e.g., *cities* and *play*) to construct a nonce sentence: "*The cities of chaperones play a long history.*" Considering all possible combinations of nouns, verbs, inflections, and contexts yields a dataset of 7,526,400 sentences which is is both large (c.f., 383 sentences in Gulordava et al. (2018)) and balanced in terms of number form (50% singular and 50% plural).

To ensure that constructed sentences are grammatically correct, we apply several manual filters (e.g., removing verbs that have ambiguous inflections), which are described in detail in Appendix B (with a list of all nouns, verbs, and sentential contexts in Appendix D). To verify the quality of the resulting stimuli, one of the authors manually examined 154 randomly generated nonce sentences in the same way that they would be presented to the model. The verb inflection was correctly predicted in all but one of the instances (with the single error attributed to carelessness), and the annotator confirmed that all generated sentences were grammatically correct. Our stimuli and code are available at `https://github.com/google-research/language/tree/master/language/bertology/frequency_effects`.

5 Exploratory Analyses: What Factors Correlate with Error Rates?

We first perform an exploratory analysis of how the model's abilities on the SVA task vary as a function of pre-training frequency. As discussed in §2, we consider generalization to unseen subject-verb pairs to be evidence of symbolic reasoning (H1 or H3), and strong frequency effects to suggest item-specific learning (H2 or H3). Note that in these experiments, it is not the individual lexical items—the subject and verb—that are unseen, only the combination of them in a single sentence. Therefore, this analysis evaluates the model's ability to perform abstract reasoning about individual items for which it has learned representations.

	Natural		Nonce	
	Seen	Unseen	Seen	Unseen
argmax$_V$ P(V)	39.1	39.0	50.0	50.0
argmax$_V$ P(SV)	22.9	50.0	41.2	50.0
BERT	3.3	8.8	15.6	17.6

Table 2: Error rate on natural and nonce evaluation stimuli, stratified by whether the subject–verb pair was seen (frequency ≥ 1) or unseen (frequency = 0) during pre-training. Heuristics (argmaxes) show performance for a item-specific learner that memorizes probabilities of specific verbs (V) or subject–verb (SV) pairs. BERT's performance degrades on unseen pairs, but is significantly better than these heuristics.

5.1 Overall Performance

Overall, the model's error rate is 3.2% on natural stimuli and 16.8% on nonce stimuli. This is similar to Goldberg (2019)'s reported 3% error on natural stimuli from Linzen et al. (2016) and a 17% error on nonce stimuli from Gulordava et al. (2018).[3]

5.2 Unseen Subject–Verb Pairs

Table 2 stratifies error rate by seen and unseen subject–verb pairs. Compared with subject–verb pairs seen at least once during training, error rates on unseen subject–verb pairs are 5% higher on natural sentences and 2% higher on nonce sentences. This degradation, however, is minimal compared with what we might expect from a naive item-specific learner (H2), represented by the heuristic baselines in Table 2. These results thus suggest that BERT reasons over representations that abstract to some degree over individual words, though it does not meet the definition of a fully-abstract symbolic learner (H1), which would have no degredation in performance.

5.3 Frequency of the Target Form

To further examine the effect of frequency, we draw inspiration from the human language processing literature. One of the most widely-observed phenomena in such research is that high-frequency words are learned better (Ambridge et al. 2015):

Reduce Error Hypothesis. *High-frequency forms reduce errors in contexts where they are the target.*

Figure 1 stratifies error rate by the training frequency of (1) subject–verb pairs and (2) verbs (independent of subject). On both natural and nonce stimuli, error rate decreases for more-frequent subject–verb pairs and more-frequent verbs, consistent with the Reduce Error hypothesis.

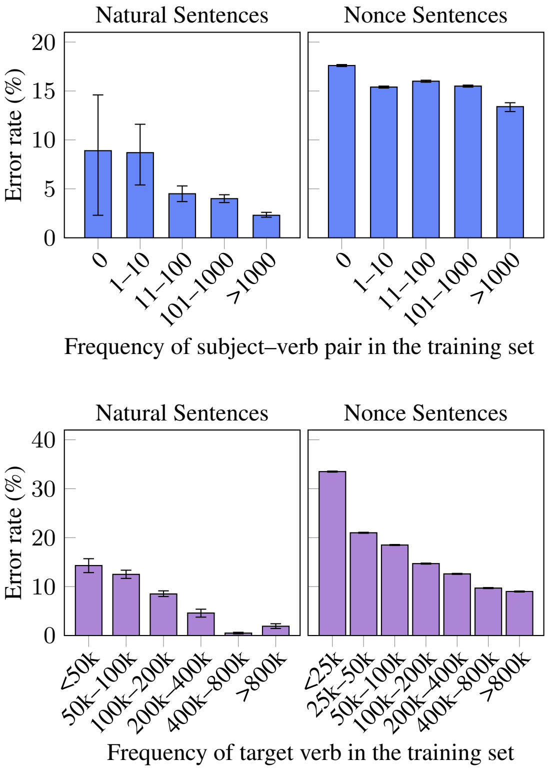

Figure 1: Error rate stratified by how often subject–verb pairs appeared in the same sentence in BERT's training set (top) and how often verbs appeared in the training set (bottom). Error rate was lower for subject–verb pairs and verbs that were more frequent.

5.4 Frequency of the Competing Form

Although seeing a verb more often in training often reduces errors when that verb is the target, when high-frequency forms are not the target, they can act as distractors and reduce accuracy:

Cause Error Hypothesis. *High-frequency forms cause errors when a competing, lower-frequency form is the target (Ambridge et al., 2015).*

Is BERT's error rate similarly affected by distractor frequency effects? For instance, the word *"combat,"* which is not only the plural form of the verb *"combat"* but also a fairly frequent noun, appears 102× more often in the training set than *"combats."* If word frequency influences BERT's predictions, then such asymmetries may cause a high error rate when the target form is *"combats."*

As Figure 2 shows, error rate is lower when the target form is more frequent relative to the competing form. For nonce sentences, for example, error rate was only 2.2% when the target form was 16

[3] The Gulordava et al. (2018) stimuli slightly differ in that all content words (not just the subject and verb) were replaced.

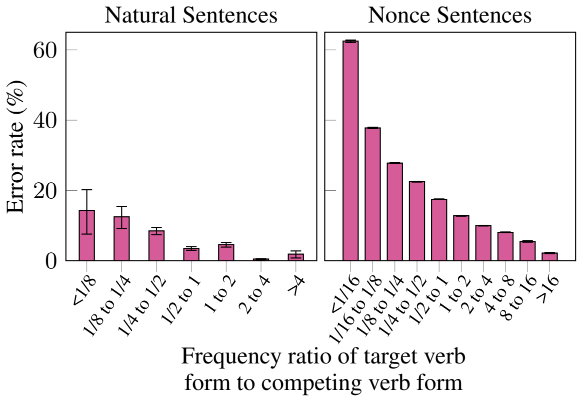

Figure 2: Error rate stratified by the ratio between the frequency of the target verb form versus the competing verb form. BERT's error rate was higher when the competing verb form occurred more frequently in the training set than the target verb form.

times or more as frequent than the competing form, compared with 62.5% when the competing form was 16 times or more frequent than the target form.

5.5 Takeaways

The above exploratory analyses suggest that BERT is influenced by both the absolute frequency of the target form (Reduce Error Hypothesis), as well as the frequency of the target relative to the competing form (Cause Error Hypothesis). Although these results are strong correlational evidence, absolute and relative frequency are highly correlated with one another (i.e., as the absolute frequency of a word increases, so does its frequency relative to other words).[4] Thus, more controlled studies are needed to establish which effects have a causal effect on BERT's rule-learning.

6 Confirmatory Analysis: Manipulating the Training Data

To better understand the above trends, we design a set of experiments in which we manipulate one variable (absolute or relative frequency of a verb form in pre-training) while holding the other fixed.

6.1 Experimental Setup

We select 60 verbs of interest (VOIs) and manipulate their training set frequencies. We choose the VOIs by taking 60 transitive verbs for which both singular and plural forms of each verb occur at least 10^4 times in the corpus (the full list is shown in Appendix D.4). We remove all sentences containing these VOIs from the training set, and, based on the experiment, add them back in such that VOIs appear at a specified (absolute or relative) frequency. We evaluate the model's performance on these VOIs by using both a natural dataset of approximately 100 examples per VOI, as well as by inserting the VOIs into the nonce (noun + sentential context) constructions from §4.2.

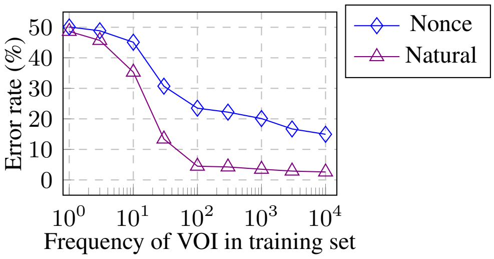

Figure 3: Effect of absolute frequency of a verb of interest (VOI) when the ratio between singular and plural forms is held constant at 1:1. The error rate for sixty VOI is shown for BERT models that have seen the sixty VOI at different frequencies in the pre-training dataset.

We note that the exact size of the training set in our manipulations changes depending on how many sentences containing VOIs are added in (e.g., models which see 10,000 examples per VOI see more total training examples than models that see only 1 example per VOI). The difference in absolute terms, however, is small (less than 1% of the total training set). Thus, we consider it unlikely that any observed difference in performance is due to a difference in the total size of the training corpus.[5]

6.2 Absolute Frequency of Verb Form

We first examine how the absolute frequency of a verb form affects the model's number agreement ability on that form. For each of nine frequencies n = 1, 3, 10, 30, 100, 300, 1,000, 3,000, and 10,000, we train a new BERT model that sees the verbs n times each during training. To isolate the effect of absolute frequency, we fix the relative frequency to be balanced—for each VOI, n instances are singular and n are plural.

The results of this experiment are shown in Figure 3. When the occurrences of a VOI are balanced between inflections (singular and plural), error rate decreases monotonically when target form is more frequent in training.

[4] This correlation is not only intuitive but also empirical—see Figure 10 in Appendix C.3.

[5] As one measure, masked language model accuracy (on the same dev set) was 59.96% for a model with VOIs appearing at frequency 10,000, versus 59.91% for a model with VOIs appearing at frequency 1.

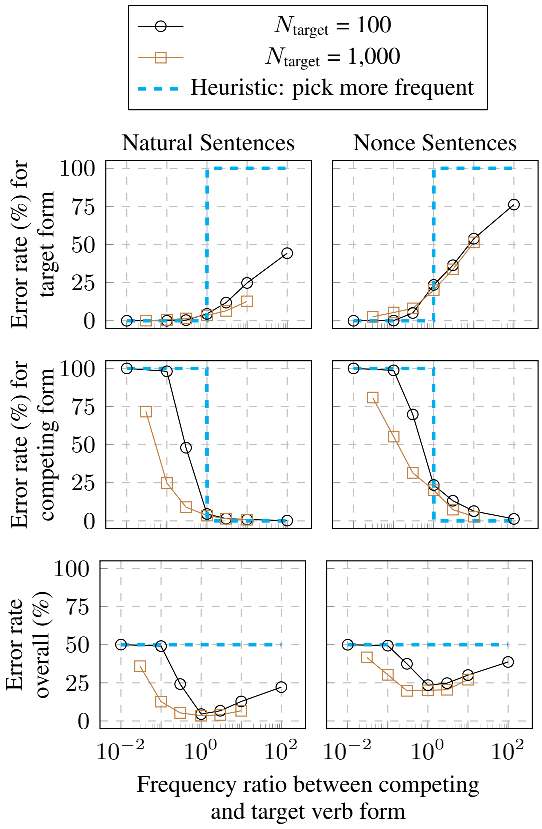

Figure 4: When the absolute frequency of the target verb form is held constant, increasing the relative frequency of the competing verb form increases error for the target form and decreases error for the competing form. This behavior is in the same direction as a heuristic that, at inference time, picks the more frequent verb.

6.3 Relative Frequency of Verb Form

We next analyze whether the frequency ratio between a target verb form v and its competing form v' affects the model's ability to produce v in context. To balance how often the target v is singular vs. plural, we use the following procedure. We randomly split our 60 VOI into two groups of 30 verbs each, which we denote as $\mathcal{S}$ and $\mathcal{P}$. In each experiment, we set the frequency of the singular verbs in $\mathcal{S}$ to N_{vary}, while holding the frequency of the plural forms of the verbs in $\mathcal{S}$ constant at N_{constant}. Likewise, we set the frequency of the plural verbs in $\mathcal{P}$ to N_{vary}, and hold the frequency of the singular form of these verbs constant at N_{constant}. We run experiments with $N_{\text{vary}} = \{1, 10, 30, 100, 300, 1{,}000, 3{,}000, 10{,}000\}$, and do this twice for N_{constant} set to 100 and 1,000. As our evaluation stimuli are balanced such that both v and v' occur as the target in every template for every VOI, we are able to analyze the effect of the v:v' frequency ratio—holding the absolute frequency of v fixed—for v:v' ranging from 1:100 to 100:1.

Figure 4 shows the results. When the competing form occurs more frequently (with respect to the target form), error rate increases for the target form and decreases for the competing form.

7 Disentangling Sources of Error

7.1 Setup

Our goal is to characterize BERT's rule-learning behavior in terms of the three hypotheses H1–H3 described in §2. The frequency effects observed in §6 rule out H1 (Idealized Symbolic Learner). However, BERT's generalization to unseen noun-verb pairs (§5.2) is too good to be explained by H2 (Item-Specific Learner). Hence, the hybrid H3 (Symbolic Learner with Noisy Observations) seems like the most plausible candidate.

H3 is not simply a catch-all compromise between rule-based and item-specific learners—H3 makes specific predictions about the nature of the errors BERT will make. Under H3, BERT represents the SVA rule and the concept of agreement features, and follows the rule as long as it identifies the number of the subject and verb correctly. Thus, H3 predicts that observed errors are due to failures to identify the number of either the subject or verb.

Given such a model, we might observe frequency effects because training frequency influences the model's ability to predict the agreement feature for a given verb form. That is, we might observe a trend like the following: if a verb v occurs in fewer than some n training examples, BERT mispredicts the agreement feature (e.g., predicting v to be singular when v is plural); if v occurs more than n times, BERT correctly predicts v's agreement feature and correctly produces v in context. In this scenario, we expect that SVA errors will correlate with frequency, but the frequency of these errors should not exceed the error rate in predicting agreement features.

7.2 Predicting Agreement Feature

To test whether the above predicted pattern holds, we use two probing classifiers (see Veldhoen et al., 2016; Ettinger et al., 2016, on probing) which we describe below.

Subject agreement feature probe. Our first probe evaluates whether, given a sentence with the verb masked, the embedding at the masked position contains information as to whether a singular or plural verb is required. This setup actually evaluates two subtasks: identifying the subject of the verb (i.e., parsing the sentence) and predicting the agreement feature of the identified subject. For simplicity, however, we use a single probing classifier because our interpretation does not hinge on differentiating these subtasks. We use our sentential templates for this experiment, for which the only cue for the number of the subject (and hence the verb) is the subject itself. Hence, if the embedding at the masked position can be used to predict number, it follows that both the subject has been identified correctly and that the agreement feature of the subject was identified correctly.

We feed our nonce sentences from §4.2 with the verb masked into the model, retrieve the final hidden state representation of the masked token, and train an MLP to classify the desired verb form (singular or plural). We train the probe using cross-validation, using sentences constructed from 150 subjects × 50 sentential for training, and the remainder for evaluation. Subjects and sentential contexts in evaluation sentences are not seen by the probe during training.

Verb agreement feature probe. Our second probe predicts the number of a verb from its contextual word embedding. If a probe can predict the number of a verb given its contextual word embedding, we can conclude that the model represents the agreement feature of that verb form and thus its predictions about SVA can, in principle, depend on the agreement feature. We obtain contextual embeddings of verbs by inserting them into the nonce sentential contexts, with the noun masked so that there are no external clues aside from the form of the verb that indicate whether the verb is singular or plural. We then train an MLP to, given the embedding, classify the verb as singular or plural. We use the 331 verbs from our nonce stimuli that were not VOI (×56 sentential contexts per verb) to train the probe and the 60 VOI (×56 sentential contexts per verb) to evaluate it.

Results. We evaluate these two probes as a function of both absolute frequency (using models from §6.2) and relative frequency (using models from §6.3). Results are shown in Figure 5 and Figure 6, respectively.

For absolute frequency (Figure 5), we see that the accuracy of the verb agreement feature probe is highly dependent on absolute frequency of the VOI. The probe has lower error for models that saw the verb form more often, implying that seeing forms more frequently in training led to embeddings of those forms that better encode the number agreement feature. In constrast, the accuracy of the subject agreement feature probe is constant, which is expected because identifying the number feature of a subject should not be affected by absolute frequency of VOI. Notably, the combined error rate of our two probes falls close to the model's observed overall error rate on the SVA task, as predicted by our "Symbolic Reasoner with Noisy Observations" hypothesis (H3).

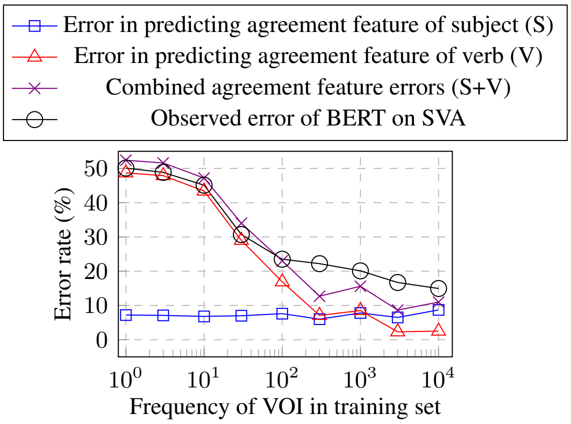

Figure 5: Errors in either identifying the number agreement feature of the subject or identifying the number agreement feature of the verb comprised a large portion of the observed SVA error.

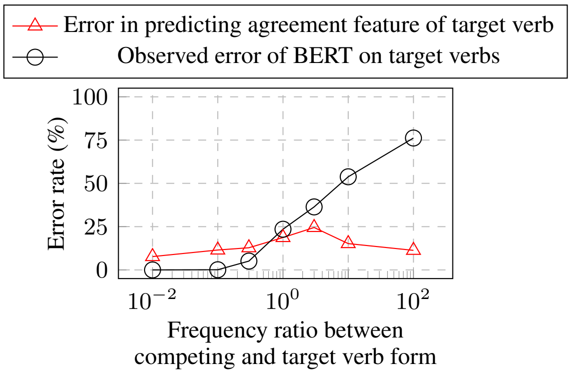

Figure 6: The error rate of a probe that predicts the agreement feature of a verb (red triangle) is not correlated with frequency of competing verb forms. Moreover, the error rate of this probe does not correlate with the observed error of BERT on the target verb, which is highly affected by frequency of competing verb forms.

For relative frequency (Figure 6), on the other hand, we see no clear increase or decrease in the accuracy of predicting the agreement feature for a target verb form v in response to changes in frequency of the competing verb form v'. In other words, when one verb is much more frequent than the other, BERT produces the more common verb form despite having access (in principle) to the information (rule + agreement features) that would allow it to infer the correct form. Such behavior is not explicitly accounted for by the "noisy observations" in H3, and thus appears more as evidence of item-specific learning (in line with H2).

8 Discussion

The goal of this work is to determine whether BERT performs SVA by implicitly representing rules defined over abstract agreement features and characterize the training conditions under which such representations emerge. We differentiate between the representation of the rule (*"if x then y"*) and that of observations (containing the correct agreement features). We draw two conclusions, which suggest a mix of systematic rule-like generalization and unsystematic item-specific inferences.

BERT appears to represent the correct rule, but fails to predict agreement features for low-frequency verb forms. Although the error rate decreases as a function of frequency of target verb v (§6.2), BERT's ability to predict the agreement feature of v (§7.2) follows the same trend. This observed behavior is thus consistent with a model that correctly represents the SVA rule (§2), but makes mistakes at inference time due to noise in the represented observations for low frequency verb forms (for example, producing *"run"* in the context *"the dog run"* because *"run"* is incorrectly encoded as singular), rather than due to a failure to represent the concept of singular altogether.

BERT fails to apply the rule when doing so requires overcoming strong item-specific priors. Similar to the absolute frequency trend, we see that BERT's error rate on SVA also decreases as a function of the frequency of the target verb v relative to its competing form v' (§6.3). Unlike above, however, we see no effect of the frequency ratio $N_v : N_{v'}$ on BERT's ability to predict the agreement feature of v when the frequency of v is fixed (§7.2). These results suggest that BERT is heavily influenced by skewed training distributions, preferring to produce more common verb forms over forms consistent with the rule. Such behavior could either mean that, when $P(v) << P(v')$, (1) BERT represents the correct SVA rule but it is overridden in favor of the prior, or (2) BERT does not represent the rule at all. Teasing apart these possibilities is a valuable direction for future work.

Open questions. Our results on absolute frequency effects indicate that BERT does not infer agreement features until it sees 10–100 examples of a verb, even though it is possible, in principle, to infer agreement features from a single training example (e.g., *"All of the dogs dax"* implies *"dax"* is a plural verb form). Future controlled studies could investigate how the sample efficiency of inferring agreement information depends on factors such as architecture (e.g., access to morphological signals), size of the model, and amount of training data. Analysis of such patterns would elucidate how models like BERT (and by extension, transformers and neural networks more generally) learn and generalize, enabling more principled development and deployment.

9 Conclusions

We have studied whether BERT's performance on subject–verb agreement exhibits rule-governed behavior. We focus on frequency effects, pre-training multiple BERT instances in order to isolate how the model's predictions are affected by absolute and relative verb frequency. Our results suggest that BERT's behavior is consistent with a system that correctly applies the SVA rule in general but struggles to overcome strong training priors and to estimate agreement features (singular vs. plural) on infrequent lexical items.

Acknowledgements

We thank Ian Tenney, Ryan Cotterell, Clara Meister, Charles Lovering, and Slav Petrov for feedback on our manuscript. We thank Christo Kirov, Kellie Webster, and Jacob Eisenstein for helpful discussions about the project.

References

Ben Ambridge, Evan Kidd, Caroline F. Rowland, and Anna L. Theakson. 2015. The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, 42(2):239–273.

Jack Bandy and Nicholas Vincent. 2021. Addressing "documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. *arXiv preprint arXiv:2105.05241*.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Rui Chaves and Stephanie Richter. 2021. Look at that! BERT can be easily distracted from paying attention to morphosyntax. In *Proceedings of the Society for Computation in Linguistics*.

Noam Chomsky. 1956. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124.

Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships. In *Proc. of CoNLL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*.

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

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

Emily Goodwin, Koustuv Sinha, and Timothy J. O'Donnell. 2020. Probing linguistic systematicity. In *Proc. of ACL*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proc. of NAACL*.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proc. of NAACL*.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proc. of EMNLP*.

Najoung Kim and Paul Smolensky. 2021. Testing for grammatical category abstraction in neural language models. *Proceedings of the Society for Computation in Linguistics*.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proc. of ICML*.

Tal Linzen and Marco Baroni. 2020. Syntactic structure from deep learning. *Annual Review of Linguistics*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *TACL*.

Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. 2021. Predicting inductive biases of pre-trained models. In *Proc. of ICLR*.

Alec Marantz. 2013. Words and rules revisited: Reassessing the role of construction and memory in language. In *27th Pacific Asia Conference on Language, Information, and Computation, PACLIC 2013*. National Chengchi University.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proc. of EMNLP*.

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021. Refining targeted syntactic evaluation of language models. In *Proc. of NAACL*.

Judea Pearl. 1988. Probabilistic reasoning in intelligent systems: Networks of plausible inference.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *TACL*.

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

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and practical BERT models for sequence labeling. In *Proc. of EMNLP*.

Sara Veldhoen, Dieuwke Hupkes, and Willem Zuidema. 2016. Diagnostic Classifiers: Revealing how Neural Networks Process Hierarchical Structure. In *CEUR Workshop Proceedings*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proc. of EMNLP*.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP*.

Charles Yu, Ryan Sie, Nicolas Tedeschi, and Leon Bergen. 2020. Word frequency does not predict grammatical knowledge in language models. In *Proc. of EMNLP*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proc. of ICCV*.

A BERT Model

To analyze the effect of word frequency on BERT's ability to follow SVA, we need to know the exact number of occurrences of each word in the dataset. The original BERT checkpoint (Devlin et al., 2019) uses both Wikipedia and BooksCorpus (Zhu et al., 2015), but BooksCorpus is no longer publicly available (Bandy and Vincent, 2021). So we train a replicated version of BERT on only wikipedia data.

Our version of BERT largely follows the procedure of the original, differing only in that we use dynamic masking and pre-train for 4 million updates at a learning rate of 3e-4.[6] Table 3 shows the performance of our replicated version of BERT.

	Downstream Task			
Pre-training data	MRPC	CoLA	MNLI	SST2
Wikipedia + BooksCorpus	86.6	82.1	84.4	92.8
Wikipedia only	86.2	78.7	84.3	92.2
	QQP	QNLI	RTE	STS-B
Wikipedia + BooksCorpus	91.2	91.6	68.5	89.3
Wikipedia only	90.8	91.5	64.9	88.8

Table 3: Performance of our replicated BERT checkpoint, which is pre-trained on only Wikipedia data, compared with the original BERT checkpoint, which used both Wikipedia and BooksCorpus.

B Nonce Stimuli Collection Details

This appendix section details our nonce stimuli collection process. Our goal is to create a large set of evaluation stimuli in which we can analyze how properties of certain stimuli (e.g., subjects, verbs, and sentential contexts) affect the model's ability to perform number agreement. Therefore, we create a list of 200 nouns, 336 verbs, and 56 sentential contexts such that any given (noun, verb, sentential context) triplet where the noun is used as the subject yields a grammatically correct nonce sentence. By considering both singular and plural inflections for possible triplet, we analyze a dataset of $2 \cdot 200 \cdot 336 \cdot 56 = 7{,}526{,}400$ sentences. To facilitate further use of our dataset, we make a plain-text version available at `http://anonymized`.

B.1 Nouns

To propose candidate nouns, we first ran a POS tagger (Tsai et al., 2019) over the pre-training dataset, and retrieved all nouns occurring at least 100 times. Then, we randomly sampled 200 nouns from this set of candidate nouns that were evenly distributed into four buckets of training set frequency (100–999, 1,000–9,999, 10,000–99,999, and 100,000+). All nouns were common nouns and were manually validated to have correct, unambiguous singular and plural inflections.

B.2 Verbs

To propose candidate verbs, we similarly retrieved all verbs that occurred at least 100 times in the training set. The masked-LM evaluation procedure for SVA requires that both the singular and plural inflections of the verb exist directly in the model's vocabulary, and so we filtered out verbs that did not match this criteria, leaving us with 379 candidate verbs.

Unlike for nouns, we generally cannot indiscriminately swap out verbs in a sentence while maintaining grammatical correctness, since some verbs are exclusively transitive (used with an object) or intransitive (used without an object). In English, more verbs can be used transitively than intransitively, and so we decided to consider only transitive verbs. We manually filtered out strictly intransitive verbs and ensured that each verb had correct, unambiguous singular and plural inflections, leaving us with 336 verbs that can be used transitively.

B.2.1 Sentential Contexts

Finally, we curated a list of sentential contexts (sentences with the subject and verb removed) that would maintain grammatical validity for both singular and plural forms of any given subject–verb pair from our list of nouns and list of transitive verbs. To get candidate sentential contexts, we randomly sampled 600 sentences from the Linzen et al. (2016) dataset of Wikipedia sentences to be manually examined. We kept only 56 of these 600 candidate sentential contexts, filtering out 544 for the following reasons:

- Sentential contexts that contained additional cues for number outside of the subject and verb inflection cannot form grammatical sentences for both singular and plural subject–verb pairs. For instance, the sentential context "*[SUBJECT], who thinks roses are red, [VERB]* ..." can only be used

[6] The decision to use dynamic masking was made to the availability of code, rather theoretically or empirically motivated. We train for additional updates because the development loss did not converge at 1 million updates (the original number used in the paper).

with singular subjects and verbs because of the modifying clause "*who thinks roses are red*"; and the sentential context "*[SUBJECT] in the park [VERB] ...*" can only be used with plural subjects and verbs because there is no determiner for the subject. 381 sentences like the above had such cues for number inflection and were removed.

- 64 sentential contexts contained verb usages that were hostile to swapping in most transitive verbs (e.g., in "*[SUBJECT] shows that ...*", "*shows*" could not be replaced with most transitive verbs).
- 15 sentential contexts contained noun usages that were hostile to swapping in most nouns (e.g., in *the fact that she likes him ...*, the subject *fact* cannot be replaced with most nouns).
- 21 sentential contexts were ungrammatical or incomprehensible to our human annotator.
- In 36 sentential contexts, the subject and verb parsed by Linzen et al. (2016) was incorrect.
- 27 sentential contexts for which the original verb was used intransitively were removed.

B.2.2 Human evaluation

To check the validity of the test set, the first author manually examined 160 generated nonce sentences in the same fashion that the model would evaluate them. That is, each example comprised either a singular or plural noun inserted into a template, and the first author had to predict the correct number inflection of a given verb. In addition, the first author had to verify that the sentence was grammatical and contained no number inflection cues other than the inflection of the subject. The sentences were presented in random order, with the first author blinded from the ground-truth label.

The first author found that 6 sentences (3 templates, since each template had two inflections) contained additional number inflection cues that were missed in the first round of annotation, and so these sentences were removed. In terms of accuracy, the first author correctly predicted the verb inflection 153 of the 154 instances (99.4% accuracy) and attributed the single error to carelessness. We take these manual evaluation results as evidence that our test set is grammatical and tests a syntactic rule that can be consistently applied by humans.

C Additional Figures and Tables

We show several auxiliary experiments that elucidate BERT's performance with respect to various characteristics of evaluation stimuli.

C.1 Relative Frequency of Forms

Following the results from §5.4, we show the error rate for singular and plural forms of all verbs in our nonce stimuli in Figure 7. Additionally, Table 4 shows the five verbs with the highest and lowest error rates, as well as their frequency ratios.

C.2 Comparison with prior work

Because our work proposes a new set of nonce stimuli (which is larger than prior work, e.g., 383 sentences from Gulordava et al. (2018) or 7 verbs from Chaves and Richter (2021)), we run several analyses from prior work on our dataset. The results are largely consistent with conclusions from prior work.

Attractors. As shown in Table 5, BERT did not perform worse on templates with more attractors (clauses between the subject and verb), corroborating Goldberg (2019).

Noun frequency. We find similar evidence, like Yu et al. (2020), that BERT did not perform better on nouns that were more frequent in the training set. Figure 8 shows these results—for each subject (we consider both singular and plural forms as a single subject), we plot that subject's error rate against its frequency in the training data.

High- and low-confidence predictions. As an auxiliary analysis to compare with Newman et al. (2021), we plot the error rate of BERT with respect to different thresholds for how confident the model was about its prediction. Figure 9 shows error rates for all predictions with confidence above some threshold, and error rates for predictions with confidence below some threshold. This result concurs with Newman et al. (2021)'s finding that model performance drops when testing verbs that the model finds unlikely.

C.3 Absolute versus relative frequency

As additional background, Figure 10 shows the correlation between absolute frequency of a verb form and its frequency relative to its competing form.

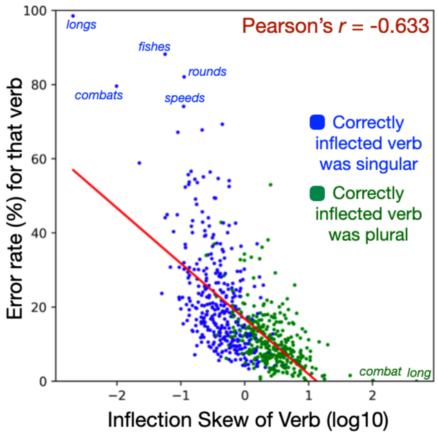

Figure 7: For all 366 verbs (×2 inflections per verb), we plot the error rate against *inflection skew*, which is how much more frequent the correct inflection of the verb occurred than the incorrect inflection in the pretraining data. Several of the most skewed words are shown—for instance, *long* appears 490 times more often in the pre-training data than *longs*, so its inflection skew is $\log_{10}(490) = 2.69$. Conversely, *longs* has an inflection skew of -2.69.

	Inflection Skew (Log10)	Error Rate
Best-performing verbs		
long	2.69	0.0%
speed	0.95	0.2%
combat	0.95	0.2%
round	2.01	0.4%
fish	1.25	0.6%
Worst-performing verbs		
longs	-2.69	98.5%
fishes	-1.25	88.0%
rounds	-0.95	82.1%
combats	-2.01	79.6%
speeds	-0.95	74.1%

Table 4: The verbs for which BERT had the highest and lowest error rates. Inflection skew indicates how much more often a verb appeared in BERT's pre-training data compared with its other number inflection, in log10. For instance, an inflection skew of 1 indicates that a verb appeared $10^1 = 10$ times more often in the training set than its other number inflection.

Stratification of Stimuli	Examples	Error Rate
All examples	10.2M	17.9%
Templates with one attractor	3.9M	15.2%
Templates with two attractors	1.6M	22.3%
Templates with three attractors	1.3M	16.8%
Templates with four attractors	672k	13.0%

Table 5: Performance for templates with different numbers of attractors (distracting clauses between the subject and verb).

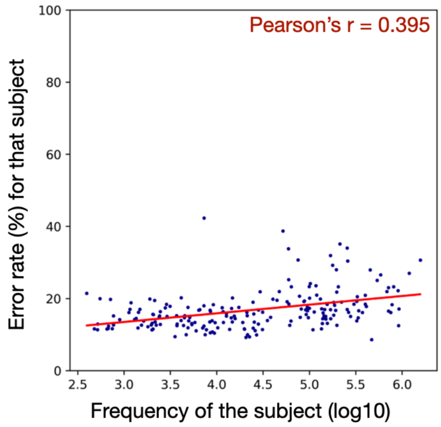

Figure 8: For all 200 subjects (both singular and plural forms are included into a single subject), we plot the error rate against the frequency of that subject in the pre-training data.

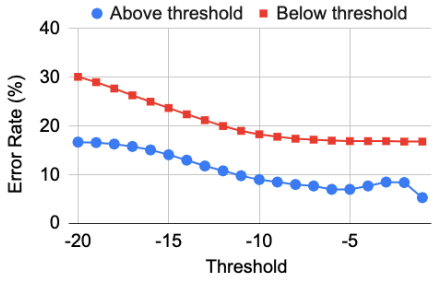

Figure 9: Error rates of examples for which the model's predictions were above and below certain thresholds.

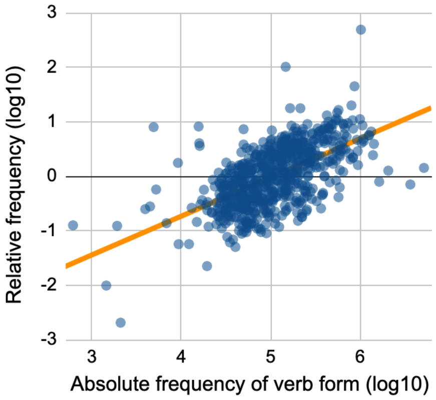

Figure 10: For all 336 verbs (672 verb forms), we plot the absolute frequency of a verb form versus the relative frequency of that verb form compared with its competing form. Pearson's $R^2 = 0.356$.

C.4 Training manipulations: Seen vs. Unseen

Figure 3 in the main body showed the performance of models that have seen the VOI n times in training, where n varies from 1 to 10,000. Table 6 stratifies this performance on the nonce evaluation stimuli by seen and unseen subject–verb pairs in the evaluation stimuli. Note, though, that this stratification differs for each VOI frequency. That is, there will be more unseen subject–verb pairs when the VOIs are less frequent in the training set.

Frequency	Seen SV		Unseen SV	
of VOI	# examples	Error	# examples	Error
1	672	63.4%	1.34M	50.1%
3	2.6k	54.8%	1.34M	48.9%
10	7.4k	43.7%	1.34M	45.1%
30	21.6k	29.6%	1.32M	30.8%
100	59.9k	20.9%	1.28M	23.7%
300	135k	22.1%	1.21M	22.8%
1,000	289k	20.4%	1.06M	20.0%
3,000	456.3k	16.3%	887.7k	16.9%
10,000	662.k	15.4%	682.1k	14.5%

Table 6: Stratifying seen (frequency > 0) and unseen (frequency = 0) subject–verb pairs in the nonce sitmuli for our training manipulation from Figure 3.

D Raw SVA Nonce Stimuli

D.1 Verbs

The 336 verbs used in our nonce stimuli are listed below (only plural/base inflections are shown): add, advance, age, aim, air, allow, analyse, angle, approach, archive, arrive, ask, assist, attack, attempt, award, bar, base, battle, bear, become, begin, believe, benefit, block, board, bond, book, border, branch, brand, break, bridge, bring, call, campaign, carry, cause, center, centre, challenge, champion, change, channel, charge, chart, circle, claim, class, coach, code, color, colour, combat, comment, compound, comprise, concern, connect, consider, contact, contain, continue, contract, control, copy, count, course, cover, create, credit, critique, crop, cross, cycle, date, deal, debate, decide, define, demand, describe, design, detail, develop, discover, display, dispute, distance, document, double, draw, drive, drug, effect, end, enter, equip, estimate, exhibit, experience, explain, extend, eye, face, factor, fan, farm, feature, feel, field, fight, file, fill, finance, find, fire, fish, fly, follow, force, form, frame, fund, gain, get, give, grade, graduate, grant, group, grow, guard, guide, hand, have, head, help, hold, honor, honour, host, house, include, increase, indicate, influence, interview, involve, issue, join, judge, keep, kill, know, label, land, lap, lead, learn, leave, level, light, limit, line, link, list, live, long, look, love, maintain, make, manage, map, mark, market, master, match, matter, mean, measure, meet, mention, minister, model, move, murder, name, need, note, number, object, offer, open, operate, order, own, pair, park, partner, pass, pattern, pay, peak, perform, picture, pilot, place, plan, plant, play, position, post, pound, power, practice, present, print, process, produce, program, project, protest, prove, provide, question, race, raid, range, rank, rate, reach, reason, receive, record, refer, reference, reflect, refuse, release, remain, repair, report, represent, reprise, require, reserve, return, reveal, review, ring, rise, risk, rival, round, route, rule, run, sample, say, scale, score, seat, see, seed, seek, send, serve, service, share, ship, show, sign, signal, single, sketch, slope, sound, source, speak, speed, sport, spot, stage, stand, star, start, state, stop, store, stream, strike, strip, structure, study, style, suggest, supply, support, surface, tackle, take, talk, target, task, tax, tell, term, test, tour, trace, track, trail, train, transport, travel, trend, trouble, try, turn, use, value, vary, vent, view, visit, voice, volunteer, walk, want, wave, win, witness, work, write,

D.2 Nouns

The 200 nouns used in our nonce stimuli are listed below (only singular inflections are shown): abuser, actuary, affiliate, album, application, artefact, articulation, artiste, aspect, astronomer, attempt, attribution, autopilot, ball, barnacle, basalt, batch, battalion, beaker, bettor, bidder, biosensor, blazer, bluebird, brake, brush, bulletin, business, campaign, capital, captor, caretaker, catholic, caveman, charge, chestnut, clarinetist, climate, columnist, command, commando, commuter, comparison, compiler, constant, consul, craftsman, credential, cup, debate, debater, demigod, device, dhole, disorder, distribution, diviner, draft, drum, dynasty, echidna, electron, emoticon, enclave, etymology, exhibition, explosive, faith, fanatic, fantasy, fat, ferret, fiction, foal, forager, form, forwarder, fossil, foundation, franchise, friendship, girl, glass, good, grantee, grapevine, hair, harmonic, headlamp, hedgehog, hotel, hypothesis, imp, impact, instruction, intensifier, interest, intrusion, island, isometry, kabbalist, kind, launch, layer, legionnaire, lioness, loading, locksmith, logarithm, logger, mammoth, martin, matchup, microphone, misfit, motorcyclist, nasal, necessity, officer, ogre, opposition, palace, panchayat, parrot, pioneer, platform, plum, poet, possibility, postposition, potentiometer, president, press, pro, proponent, provider, race, radiologist, rank, rat, reaper, region, relief, remark, repeater, repellent, rescuer, researcher, retriever, ribbon, ride, ring, rogue, role, sage, salaryman, seagull, section, selection, sense, sex, shearer, sheepdog, shoreline, siding, sign, simulation, situation, skateboarder, snowflake, sorcerer, specimen, speech, spill, spiritualist, spore, spring, starling, starship, stingray, stock, street, suffix, switch, tarsier, terrier, town, treaty, truth, tutor, tweeter, undertaker, uniform, vendor, ventilator, view, walker, warlock, watcher, youngster

D.3 Sentential Contexts

The 56 sentential contexts used in our stimuli are listed below:

1. the edwardian semi-detached [SUBJECT] of brantwood road , facing the park [VERB] an art deco style whilst those in ashburnham road include ornate balconies .
2. for protestant denominations , the [SUBJECT] of marriage [VERB] intimate companionship , rearing children and mutual support for both husband and wife to fulfill their life callings .
3. the [SUBJECT] , due to being the same colour green as the shield , [VERB] a green sign with a white inlay border , and a green outer border .
4. wright 's other acting [SUBJECT] on television [VERB] itv 's crossroads and bbc one 's doctors .
5. the [SUBJECT] , where most of the population lives and the majority of activity takes place , [VERB] an expanse of low-lying , flat , and comparatively dry grassland .
6. the [SUBJECT] of the load line with the transistor characteristic curve [VERB] the different values of ic and vce at different base currents .
7. the other [SUBJECT] on the pillar [VERB] only two bolts for sport climbing , making a ground-fall more likely should a mistake be made .
8. the [SUBJECT] honoring saint joseph , the saint patron of the city , [VERB] a part of the city 's culture .
9. furthermore , other scholars have noted how the cryptic dharani [SUBJECT] within the lotus sutra [VERB] a form of the magadhi dialect that is more similar to pali than sanskrit .

10. he may be a wonderful vandal fighter , but i think the [SUBJECT] about this matter at the talk page [VERB] a clear misunderstanding of practice , here .
11. bce) , although no [SUBJECT] of that period [VERB] today .
12. its major american [SUBJECT] in the plastic model kit market [VERB] amt-ertl , lindberg , and testors .
13. his feature [SUBJECT] as a screenwriter [VERB] american hot wax , rafferty and the gold dust twins and where the buffalo roam .
14. hence , some state [SUBJECT] of assault weapon explicitly [VERB] assault rifles .
15. his other research [SUBJECT] in modern cornish history [VERB] cornish emigration ; ethnicity and territorial politics and centre-periphery relations .
16. her [SUBJECT] of interest [VERB] the history of childhood and family , networks , social interactions and reciprocity , poverty , welfare , gift-exchange and the history of the emotions .
17. the [SUBJECT] for approval of an employment visa [VERB] suitable educational qualifications or work experience , a secured employment contract in moldova , provide proof of adequate means of subsistence in moldova , police confirmation that you have no criminal record , and a satisfactory medical examination .
18. the [SUBJECT] of this measure [VERB] not a single reason to advance why this bill should not pass .
19. the control [SUBJECT] at the left hand end of the instrument [VERB] the d6 clavinet mixture controls and a sliding control for the volume .
20. second of all , his [SUBJECT] today [VERB] very high prices , placing him well above the notability minimum for artists .
21. communities bans should be the absolute last resort when an editor 's [SUBJECT] to the project [VERB] a net detrimental effect and lesser sanctions have failed to improve this problem .
22. the [SUBJECT] of the first session of the washington county court during that year [VERB] a call for a road from canon 's mill to pittsburgh .
23. the anaerobic [SUBJECT] in osteomyelitis associated with peripheral vascular disease generally [VERB] the bone from adjacent soft-tissue ulcers .
24. i have taken note of michaelqschmidt 's keep vote , but note that the [SUBJECT] in the second google news link [VERB] nothing non-trivial and the first shows only brief local coverage .
25. the small [SUBJECT] of non-white students in the schools accurately [VERB] the racial and ethnic demographics of the community .
26. however , the [SUBJECT] of cost versus benefit [VERB] an area of ongoing research and discussion .
27. the common [SUBJECT] for dizziness [VERB] vertigo , pre-syncope and disequilibrium .
28. the [SUBJECT] of the increasing lack of physical education [VERB] budgetary pressure which limits the resources provided for this ; the increasing attractiveness of rival pastimes such as video games ; and the increased emphasis upon academic results .
29. the [SUBJECT] of the former route from drysdale to south geelong , along with a walking track adjacent to the queenscliff-drysdale line , now [VERB] the bellarine rail trail , accessible to cyclists and walkers .
30. one 's [SUBJECT] at this level of existence [VERB] a consistency and coherence that they lacked in the previous sphere of existence .
31. the [SUBJECT] in the gulf [VERB] santa catalina island .
32. the oaths themselves talk about the family bond , and we can conjecture that the [SUBJECT] of secrecy [VERB] the family loyalty as well as a sense of self-preservation .
33. the [SUBJECT] on death row [VERB] foreign nationals , many of whom were convicted of drug-related offences .
34. the [SUBJECT] on current routes [VERB] nothing to the info .
35. the [SUBJECT] of the buildings [VERB] commercial space , including two restaurants , a dental office (pinnacle dental) , a medical clinic , and spa , while the surrounding area will consist of public parks , shops and recreation spaces .
36. the genetic [SUBJECT] among the viruses isolated from different places (7-8) [VERB] the difficulty of developing vaccines against it .
37. the [SUBJECT] of shipwrecks in 1980 [VERB] all ships sunk , foundered , grounded , or otherwise lost during 1980 .
38. the [SUBJECT] of these techniques to humans [VERB] moral and ethical concerns in the opinion of some , while the advantages of sensible use of selected technologies is favored by others .
39. under chairwoman agnes gund , the moma ps1 's [SUBJECT] of directors [VERB] the artists laurie anderson and paul chan , art historian diana widmaier-picasso , fashion designer adam kimmel , and art collectors richard chang , peter norton , and julia stoschek .
40. the first [SUBJECT] of " cities of the plain " [VERB] a detailed account of a sexual encounter between m .
41. the diocese 's [SUBJECT] of arms [VERB] a red field in honor of the sacred heart of jesus .
42. the [SUBJECT] from local schools , like west lafayette junior-senior high school , also [VERB] high academic standards .
43. the [SUBJECT] of chaperones [VERB] a long history .
44. his [SUBJECT] in rabies [VERB] multiple studies investigating efficacy and side effects of tissue culture derived rabies vaccines , as well as leading clinical trials as primary investigator in collaboration with the who .
45. the [SUBJECT] of glaciers on people [VERB] the fields of human geography and anthropology .
46. the object is to eliminate as many stars as possible before the [SUBJECT] of blocks [VERB] the top of the screen ; a hand raises up the set of blocks , introducing a new row .
47. the [SUBJECT] of all preordered sets with monotonic functions as morphisms [VERB] a category , ord .
48. the [SUBJECT] of goods and services chosen [VERB] changes in society 's buying habits .
49. the [SUBJECT] of binding partners to induce conformational changes in proteins [VERB] the construction of enormously complex signaling networks .
50. the [SUBJECT] of university of toledo people [VERB] notable alumni , former students , and faculty of the university of toledo .
51. romayne 's film [SUBJECT] scoring independent features and documentaries [VERB] the screamfest crystal skull winner h .

52. the [SUBJECT] of basic needs providers emigrating from impoverished countries [VERB] a damaging effect .
53. the [SUBJECT] of extensive deposits of fishbones associated with the earliest levels also [VERB] a continuity of the abzu cult associated later with enki and ea .
54. in addition , the [SUBJECT] of secondary measures [VERB] applications in quantum mechanics .
55. the [SUBJECT] of large quantities [VERB] specific precautions to prevent the release of the vapour into the environment .
56. the [SUBJECT] with the highest votes [VERB] the deputy mayor and may proxy for the mayor .

D.4 Verbs of Interest

The 60 verbs of interest (VOI) used in our pretraining manipulation experiments are listed below: emphasize, threaten, announce, utilize, propose, translate, confront, portray, prefer, declare, denote, admit, conclude, inform, imply, relate, derive, suffer, constitute, employ, possess, attract, assume, resemble, depict, demonstrate, incorporate, celebrate, generate, realize, collect, enjoy, deliver, introduce, explore, prepare, depend, recognize, encourage, contribute, hear, publish, retain, discuss, enable, prove, spend, comprise, define, marry, affect, teach, argue, survive, choose, identify, lose, vary, raise, reveal