Artificial Neural Networks as Models of Human Language Acquisition

by

Alex Warstadt

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Linguistics

New York University

September 2022

Samuel R. Bowman

CHAPTER 6

The Role of Indirect Evidence in Grammar Learning: Investigations with Causal Manipulations of the Learning Environment

---

Abstract

Progress in the study of human language acquisition has been limited by our ability to conduct experiments to draw causal inferences about the effects of variables in the input. This is due to the impracticality of manipulating the input to children acquiring language, and the ethical implications of conducting any manipulation that could impede L1 acquisition. This limitation has been especially obvious in the case of Poverty of the Stimulus claims, such as those surrounding structure dependence in subject auxiliary inversion. Decades of debates on this topic have fixated on the untested assumption that direct evidence against a linear subject auxiliary inversion

rule is the most important factor in its acquisition. More recent work that recognizes the potential importance of indirect evidence has failed to conduct experiments on the full scale of human language acquisition.

In this study, we provide a proof-of-concept for a large-scale controlled ablation study on the input to model learners,[1] while also testing the sufficiency of indirect evidence for acquiring a hierarchical bias. We adopt a top-down approach to constructing a fully controlled training environment. Starting with a naturalistic corpus, we use a statistical parser to systematically filter out direct evidence for the hierarchical rule for subject auxiliary inversion. After training language models in both filtered and unfiltered environments, we test them on a new hand-crafted set of test cases for complex subject auxiliary inversion using an unsupervised forced-choice acceptability judgment paradigm. Our experiments show that direct evidence—while often helpful for acquiring hierarchical rules—is not always necessary, and set the groundwork for subsequent experiments that take advantage of artificial neural networks to address previously untestable hypotheses about human language learning.

6.1 Introduction

6.1.1 Indirect Evidence and the Poverty of the Stimulus

Poverty of the stimulus claims are claims that the input to typical children is insufficient to explain learning of some target phenomenon without assuming some *substantive innate advantage*. Many shortcomings of the input have been identified:

[1] Wei et al. (2021) anticipate this kind of design in a study that alters the frequency of specific verb forms in language model pretraining data. However, the goal of their study is to control for a confound affecting our ability to determine whether LMs apply grammatical rules systematically, not to test the necessity of some environmental stimulus for grammar learning.

small quantity, noise, lack of negative evidence. But the one we focus on in this chapter is the lack of direct evidence against competing hypotheses. Native speakers make consistent and predictable acceptability judgments for novel sentence types. This must be the case since finite experience cannot give evidence for all of the infinite combinatorial possibilities of syntax which we theoretically have command over.

Most often, poverty of the stimulus claims are invoked as a premise in service of the conclusion that humans have substantive innate advantages. But in many cases these claims have themselves become the conclusion: The end-goal of many acquisition studies is to prove the insufficiency of the input. A common approach is to conduct a corpus study in which researchers focus on a target phenomenon where the input is thought to be underspecified, and count instances a particular form of input that would disambiguate the correct human-like generalization from counterfactual generalizations which are not observed. The argument goes that the poverty of the stimulus claim is supported if this particular kind of input is absent, or too rare or noisy to provide a usable learning signal.

A common rebuttal to such studies is that they do not rule out the existence of all forms of disambiguating evidence. While direct counterexamples are probably highly relevant for ruling out an incorrect generalization, there may be less explicit sources of evidence in the input. This position, which we dub the *Indirect Evidence Hypothesis*, holds that even the absence of direct evidence for a target generalization, a learner can still rely on indirect evidence to consistently arrive at that generalization. Just as the poverty of the stimulus is not a single claim, but a whole family of claims, indirect evidence is a schema which can be applied to any number of learnability targets.

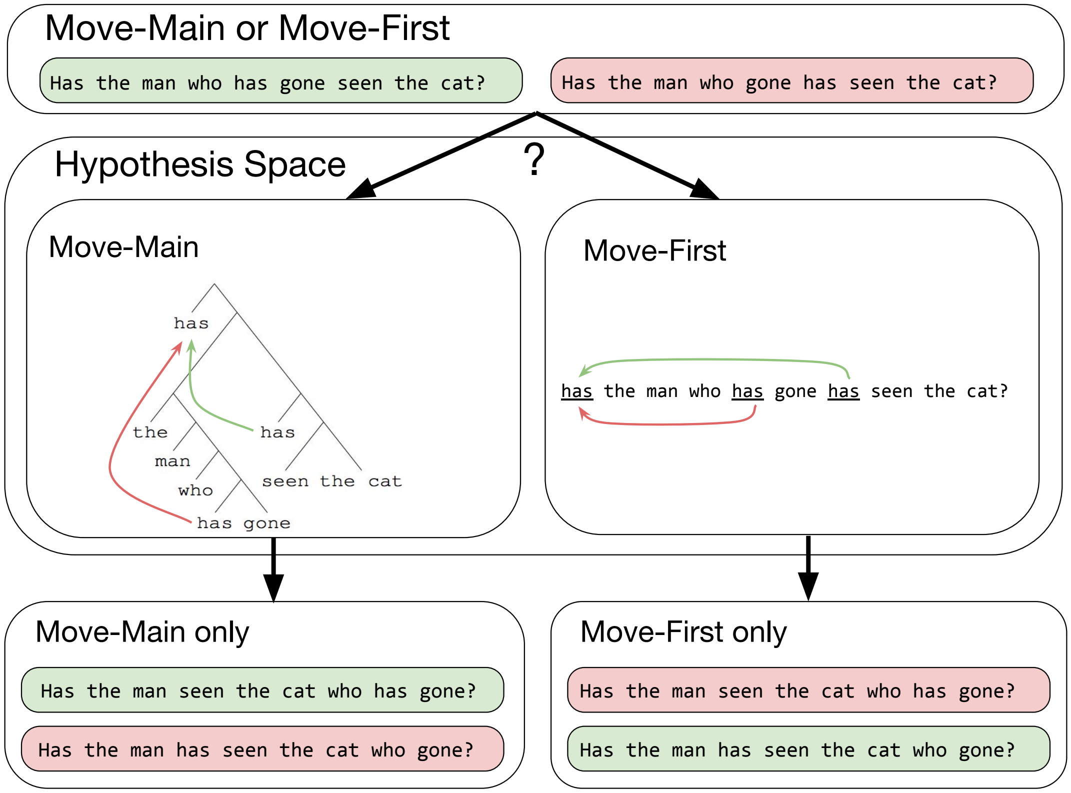

Figure 6.1: Illustration of the MOVE-MAIN and MOVE-FIRST hypotheses for subject auxiliary inversion in English.

6.1.2 The Subject Auxiliary Inversion Puzzle

The rule for subject auxiliary inversion in English is the focus of one of the longest debated poverty of the stimulus claims, summarized in Figure 6.1. Informally, English forms interrogatives (14b) from declaratives (14a) by moving an auxiliary verb before the subject (in the case of *wh*-interrogatives (14c), a *wh*-word also moves before the auxiliary). The tricky part is in deciding *which* auxiliary is fronted in sentences containing more than one. The vast majority of interrogative clauses contain only a single auxiliary verb, thereby providing no direct evidence about what to do when presented with multiple auxiliaries. To make matters worse, when multiple auxiliaries are present, the one which is fronted is nearly always both the main auxiliary of the interrogative clause, and the first auxiliary. How then is the learner to decide between two candidate rules: A structural rule (MOVE-MAIN), or a linear one (MOVE-FIRST)?

(14) a. The string quartet was composed by Mozart.
b. Was the string quartet composed by Mozart?
c. Who was the string quartet composed by?

If we observed variation in adult grammars regarding these two rules, there would be no puzzle. However, native English speakers universally prefer the MOVE-MAIN hypothesis, as evidenced by the fact that they all accept (15b) and not (15c) as the interrogative form of (15a).

(15) a. The string quartet he is rehearsing was composed by Mozart.
b. Was the string quartet he is rehearsing composed by Mozart?
c. *Is the string quartet he rehearsing was composed by Mozart?

To account for the consistency of this learning outcome two common hypotheses are commonly considered: The *Direct Evidence Hypothesis* holds that examples like (15b) are universally present in the input to children in sufficient quantities, while the *Innateness Hypothesis* holds that children universally possess an innate bias that leads them to prefer structural rules over linear ones.

Given the significance of the conclusions hanging on this phenomenon, there has been some back-and-forth about the plausibility Direct Evidence Hypothesis over the years. In early writings on the topic, Chomsky (1965, 1971) mentions it only as a straw man hypothesis, dismissing it on speculation that examples like (15b) are rare. Pullum and Scholz (2002) challenge this assumption. They conduct an informal corpus search from which they claim (without providing their reasoning) that approximately 1% of interrogatives in typical corpora constitute direct evidence against the MOVE-FIRST rule (we more or less replicate this finding in Section 6.2.3).

In a response article, Legate and Yang (2002) push this estimate down to 0.07%. They further make an interesting argument about the precise quantity of direct evidence required to learn the phenomenon. They argue that for a low-bias data-driven learner, any two learning targets representing a binary decision (e.g. MOVE-FIRST vs. MOVE-MAIN) will require roughly equal amounts of direct disambiguating evidence. Therefore, if two targets have the same age-of-acquisition, they must have the same frequency of direct evidence. They cite 3;2 as the age-of-acquisition for MOVE-MAIN (Crain and Nakayama, 1987), which they note is roughly the same as two other targets where the frequency of direct evidence is known to be 1.2%. On this basis, they argue that the frequency of direct evidence for MOVE-MAIN is too low for it to be learned without innate bias.

However, this argument overlooks the potentially significant role of indirect evidence in the acquisition of MOVE-MAIN (Reali and Christiansen, 2005). This mechanism for indirect evidence is explained by Perfors et al. (2011):

> While a child may not receive direct evidence about the correctness of a particular hierarchical phrase structure rule for analyzing some particular set of sentences such as the aux-fronting examples, there is vast indirect evidence for the general superiority of syntax with that structure throughout language. A learner who adopts a hierarchical phrase

> structure framework for describing the syntax of English will arrive at a much simpler, more explanatory account of her observations than a learner who adopts a linear framework.
>
> (Perfors et al., 2011: p. 310)

Thus, the Indirect Evidence Hypothesis depends in part on the presence of a simplicity bias in human learning (Chater and Vitányi, 2003; Hsu et al., 2013). What counts as indirect evidence on this view is potentially extremely broad. Any linguistic generalization that is sensitive to hierarchical structures and not linear order provides some degree of indirect evidence favoring MOVE-MAIN over MOVE-FIRST. This suggests there is a cline of directness of evidence, ranging from non-interrogative uses of subject-auxiliary inversion such as negative inversion, to structural transformations on elements other than auxiliaries such as passivization, to other generalizations that are sensitive to hierarchy rather than linear order such as subject-verb agreement.

Unfortunately, the Indirect Evidence Hypothesis is not testable using corpus studies. As long as direct evidence is present in any quantity, there is no obvious way to determine whether it does or does not play a necessary role in learning MOVE-MAIN. And for obvious reasons, there is no way remove all direct evidence from the input to children during language acquisition.

6.1.3 Artificial Learners and Subject Auxiliary Inversion

For these reasons, a number of researchers have turned to simulations with artificial learners to try to sway the debate about subject auxiliary inversion. In one study, Reali and Christiansen (2005) show that simple statistical learners including bigram language models trained on child-directed speech assign higher likelihood to grammatical MOVE-MAIN sentences (15b) than to ungrammatical MOVE-FIRST sentences (15c). However, the fact that a bigram model can do so well suggests not that there is evidence for MOVE-MAIN in co-occurrence data, but that the specific test cases under investigation are too easy. Indeed, Kam et al. (2008) show that these models fail to generalize when the test cases are modified minimally to exclude a specific high-probability bigram appearing only in the grammatical sentence.

Subsequently, Perfors et al. (2011) found evidence that a Bayesian grammar induction model prefers context-free grammars over non-hierarchical hypotheses. While they were not specifically interested in subject auxiliary inversion, their work articulated a version of the Indirect Evidence Hypothesis. However, their experiments were limited due to the fact that their learner does not lack an innate structural bias. In fact, it is explicitly presented with fully formulated context-free grammars as part of finite hypothesis space. While the learner also considers other hypotheses such as a set of regular grammars, it almost certainly puts greater prior probability on hierarchical generalizations than a standard artificial neural network, constituting a significant language-specific advantage.

More recently, McCoy et al. (2018, 2020) and Petty and Frank (2021) have all approached this question by training modern sequence-to-sequence models to transform declaratives to interrogatives, using a training set that supports both MOVE-FIRST and MOVE-MAIN. These works find that neither RNNs nor Transformers have a systematic hierarchical bias, though certain architectures adopt a generalization consistent with MOVE-MAIN. However, these studies all probe models that are trained end-to-end on highly simplified synthetic languages, meaning they cannot really exclude the possibility that more substantial exposure to natural language induces a systematic bias towards MOVE-MAIN.

In fact, is exactly what Warstadt and Bowman (2020) and Mueller et al. (2022) find. They take Transformer language models pretrained on large quantities of natural language text, and fine-tune them on ambiguous datasets supporting both MOVE-MAIN and MOVE-FIRST. Both find that these models are overwhelmingly successful at rejecting MOVE-FIRST for English subject auxiliary inversion, and their outputs are consistent with the systematic application of MOVE-MAIN. However, these studies still leave some questions. First, the models they evaluate are trained on billions of words of input, and so they have an unfair advantage over human learners. Second, as shown by corpus studies by Pullum and Scholz (2002) and Legate and Yang (2002), as well as later in this chapter, it is likely that these large language models have been exposed to about a hundred thousand instances of direct evidence against MOVE-FIRST which could sway their behavior on the downstream task.[2]

To convincingly address the learnability of MOVE-MAIN we need models like those trained by McCoy et al. (2018, 2020) and Petty and Frank (2021) that have never been exposed to direct evidence against MOVE-FIRST. However, even if indirect evidence turned out to be sufficient to induce hierarchical generalization in the absence of direct counterexamples to MOVE-FIRST, what counts as helpful evidence could be difficult to hypothesize about, and may be highly distributed across natural

[2]Mueller et al. (2022) give a much lower estimate of less than 4 instances of direct evidence against MOVE-FIRST per 100B tokens of English. We believe this is a significant underestimate due to an excessively narrow definition of direct evidence. Their estimate reports only the frequency of sequences of an interrogative followed immediately by the corresponding declarative, as in: *Has the man who has gone seen the cat? The man who has gone has seen the cat.* Defining direct evidence in this way is more defensible (though still arguably too narrow) for their task of generating the interrogative given a declarative, but for our acceptability judgment task, a complex interrogative alone is sufficient to rule out a competing surface generalization.

language, making it impractical to try to build all the necessary indirect evidence into a synthetic language bottom-up. Indeed, Mulligan et al. (2021) attempt this bottom-up approach, training models in a multitask setting including unambiguous evidence for some structural generalizations other than MOVE-MAIN. While this indirect evidence sometimes reduced reliance on MOVE-FIRST, it was never sufficient to induce systematic hierarchical generalization.

Given a lack of success on the bottom-up approach, it makes sense to attempt a top-down approach to constructing a training set with only indirect evidence against MOVE-FIRST. This means taking a naturalistic corpus which should already contain all or most of the kinds of indirect evidence which could conceivably be relevant to subject auxiliary inversion, and systematically ablating the direct evidence against MOVE-FIRST. What follows is our attempt at doing just this.

6.2 Syntactic Filtering

To deliver on our top-down approach to constructing a pretraining dataset with only indirect evidence, we implement a *syntactic filter* that uses a neural network-based dependency parser to identify direct evidence for the subject auxiliary inversion rule in naturally occurring text. The syntactic filter serves two purposes: First, it aids in doing a corpus study—presented in the current section—on the kinds of direct evidence in different domains of text. Second, it allows us to filter out direct evidence from the input to model learners, which is the main manipulation in our subsequent experiments. The goal of the filter is to catch any sentences that illustrate which verb should be targeted by the subject auxiliary inversion transformation. Thus, the filter

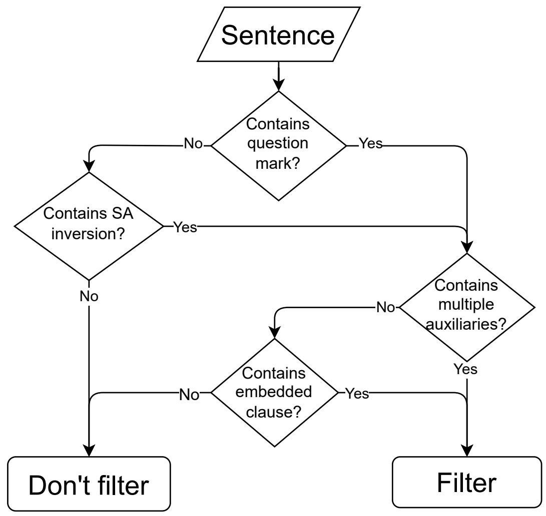

Figure 6.2: Logic for the syntactic filter.

should have high recall potentially at the cost of precision, which we confirm later in this section.

6.2.1 Implementation of the Syntactic Filter

The logic for the filter is shown in more detail in Figure 6.2. Broadly speaking, the filter should catch any sentences that both (a) include subject auxiliary inversion and (b) contain multiple verbs that could be targeted by the transformation. For condition (a), the filter detects subject auxiliary inversion if there is an auxiliary verb anywhere in the sentence that precedes its subject.[3] To increase recall, the filter also treats any sentence that contains a question mark as having subject auxiliary inversion. For condition (b), the filter first checks whether the sentence contains multiple auxiliaries.

[3]The dependency labeling schema makes the implementation a bit more complicated. Unlike in generative syntax, the lexical verb, not the auxiliary verb, is labeled as the head. For this reason, if the auxiliary is a dependent of a lexical verb, we check for any subject of the lexical verb whether or not it precedes the auxiliary verb.

If not, it checks whether there is a dependency arc anywhere in the sentence that indicates the presence of an embedded clause.

The filter uses spaCy's Transformer-based dependency parser,[4] which is built on top of RoBERTa$_{\text{BASE}}$. The spaCy parser is trained on the Penn Treebank converted to dependency graphs.[5] It has near state-of-the-art performance on Penn Treebank, achieving accuracy of 0.98 on part-of-speech tagging, 0.95 on unlabeled dependencies, 0.94 on labeled dependencies, and 0.90 F1 on sentence segmentation.[6]

6.2.2 Evaluating the Syntactic Filter on Universal Dependencies

Since direct evidence is relatively rare in text (by any estimate), we use automated methods to up-sample data which should be non-trivial for the filter. In this round of validation, we compare our filter to a silver standard obtained by applying the filter logic using human-annotated dependency parses from the English training set of the Universal Dependencies treebank (Nivre et al., 2015). We then obtain gold standard annotations by manually reviewing all examples where the filter and the silver standard disagree on whether a sentence should be excluded, and a subset of examples of where they agree. This allows us to better estimate the error rate of the filter without having to manually review disproportionately many correct predictions.

For our gold standard annotations, we devise a more refined set of criteria that an instance of complex subject auxiliary inversion must meet to count as direct

[4] https://github.com/explosion/spacy-models/releases/tag/en_core_web_trf-3.2.0

[5] https://github.com/clir/clearnlp-guidelines/blob/master/md/components/dependency_conversion.md

[6] In a subset of experiments, we used a different parser from spaCy: https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-2.3.0. These were experiments conducted using OpenSubtitles data prior to the release of the Transformer-based parser. The performance of the model is still strong: 0.92 accuracy on unlabeled dependencies and 0.90 on labeled dependencies.

Gold	Silver	Filter	# Extrap	# Annot	Freq. (%)	Example	Explanation
Yes	Yes	Yes	215	29	1.88%	Is that Microwave that you gave Dan really expensive?	
						But so was it unlikely that a small group of Arab mujahidin would virtually take over Afghanistan.	Non-interrogative inversion; distractor verb in comp. clause.
Yes	Yes	No	2	2	0.018%	Also, would those of you who have not responded to me via email confirming your acceptance of the terms upon which our four companies have agreed to assume cost responsibility for the TCA work on this.	
Yes	No	Yes	1	1	0.009%	Never in history, not even in the Nazi period, was there such total disregard of all of the above as we observe now.	
Yes	No	No	0	0	0%		
No	Yes	Yes	170	23	1.49%	Also Do you want me to give it to you now or would you rather wait until tomorrow?	Inversion is clause bounded.
						"Are we made?" Reid suggested.	Inversion is in a quote.
						You think you're going to get it back?	No inversion (echo question).
						When common sense takes a back seat to politics and legal mumbo jumbo what have we become?	Inversion is clause bounded.
No	Yes	No	65	65	0.570%	Did anything happen before losing his trust?	Distractor verb is non-finite.
No	No	Yes	40	40	0.351%	Are you the right lawyer to look at this?	
No	No	No	10918	502	95.7%	Is it safe to go to Rotarua, since the earthquakes?	Distractor verb is non-finite.
						How long to save up for a canon t3i?	No inversion (Fragment question), distractor verb is non-finite.

Table 6.1: Confusion matrix for the spaCy-based filter ("Filter"), the silver-standard Universal Dependencies-based filter ("Silver"), and human annotations ("Gold"), with annotated examples from Universal Dependencies. The column "# Extrap" shows the estimated quantity, extrapolating from the partial annotations shown in column "# Annot".)

evidence for MOVE-MAIN. Table 6.1 provides examples. To a first approximation, we define direct evidence as all instances where subject auxiliary inversion occurs in a sentence that contains multiple candidate verbs that could potentially be inverted. A candidate verb is any auxiliary verb or a finite lexical verb. We make an exception for cases where subject auxiliary inversion occurs within embedded position (such as in a quotation or a tag question) and the distractor occurs outside of the embedded constituent. We do not consider these cases of direct evidence since it is possible to formulate both MOVE-MAIN and MOVE-FIRST in such a way that they apply only within the embedded clause.

Based on these criteria, not all instances of direct evidence for MOVE-MAIN are instances of direct evidence against MOVE-FIRST. There are many examples of complex subject auxiliary inversion where the main verb is the first verb and the distractor follows it. We adopt these more inclusive criteria since they also cover direct counterexamples to other surface generalizations (e.g., MOVE-LAST) that could not be ruled out by direct counterexamples to MOVE-FIRST alone.

The full confusion matrix for the filter, silver standard, and gold standard—along with examples—is given in Table 6.1. From this, we can compute recall, precision, and overall accuracy (Table 6.2). First, we confirm that the filter has very high recall: By our best estimate, it catches 99% of the direct evidence we identified by manual annotation. Also, as expected, precision is quite a bit lower: Only 51% of the sentences caught by the filter actually constituted direct evidence.

We can also estimate what proportion of misclassified sentences are due to parsing errors as opposed to the filter logic. For direct evidence that gets past the filter, this is difficult to estimate due to their sparsity (we only identified 3 such ex-

	Recall	Precision	Accuracy
Filter vs. Gold	99%	51%	98%
Filter vs. Silver	85%	90%	99%
Silver vs. Gold	99.5%	48%	98%

Table 6.2: Agreement metrics for the spaCy-based filter ("Filter"), the silver-standard Universal Dependencies-based filter ("Silver"), and human annotations ("Gold").

amples). However, for examples caught by the filter that did not constitute direct evidence, we observe that the silver standard also filters 84%. Assuming there are no parsing errors in the treebank, this implies that the large majority of misclassified sentences are due to overly aggressive filter logic, rather than parser errors.

6.2.3 Corpus Study

We also use these annotations to make estimates about the prevalence of direct evidence for the subject auxiliary inversion rule in written text. Based on our extrapolations in Table 6.1, we estimate that approximately 218 sentences are instances of complex subject auxiliary inversion, out of a total of 11,411. This means that about 2% of sentences contain both subject auxiliary inversion and multiple finite verbs or auxiliaries. However, most of these examples lack a complex subject, meaning they are still consistent with MOVE-FIRST. In fact, of the 32 sentences that were manually marked as containing complex subject auxiliary inversion, only 2 of them had a complex subject. While extrapolating from such a small sample comes with great uncertainty, this puts our best estimate for the prevalence of evidence against MOVE-FIRST (before filtering) at 0.1% of all sentences, which is within an order of magnitude of estimates by both Pullum and Scholz (2002) and Legate and Yang (2002).

Even with this uncertainty, there is little doubt that direct evidence against MOVE-MAIN is present in ordinary samples of written text. However, our goal was never to test directly whether the actual quantity of direct evidence was sufficient to reject MOVE-MAIN. Instead, we aim to reduce this quantity even further through syntactic filtering. With the recall of the filter at 99%, this means we estimate the prevalence of direct evidence after filtering to be 0.001%, or one in one hundred thousand sentences. Such a large reduction in an already rare phenomenon should substantially lower the chance that models' predictions can be swayed by these examples. Put in another way, assuming the average sentence is 10 words in length (this is probably an underestimate), our filtered models with 100M words of training data (at the upper range of what a child is exposed to) will have been exposed to only about 100 instances of direct evidence against MOVE-FIRST. As future work, one could attempt to counteract the affect of such direct evidence against MOVE-FIRST with corrupted examples providing evidence against MOVE-MAIN.

6.3 Language Model Training

We train all the language models for the main experiment from scratch as described in this section.

6.3.1 Conditions Overview

A summary of all models and conditions is given below. For RoBERTa-style models, there are 16 conditions total. For each condition, we evaluate three separate instances of model trained in that condition with randomly sampled hyperparameters. This gives a population of 48 models. For 5-gram baselines, we only train one instance per condition, and the only condition we vary is the size of the training data.

- RoBERTa-style models (3 instances / condition)
  - Two treatments: FILTERED, CONTROL
  - Two domains: WRITTEN, SPOKEN
  - Four input volumes (# of words): 1M, 10M, 100M, 1B
- 5-gram baselines (1 instance / condition; CONTROL treatment, WRITTEN data only)
  - Four input volumes (# of words): 1M, 10M, 100M, 1B

6.3.2 Input Volumes

To investigate how learnability is affected by the scale of the input, we train language models on different volumes of input data in four tiers: 1M words, 10M words, 100M words, and 1B words. The same tiers are used by Warstadt et al. (2020b) to train the miniBERTas. To downsample, we first separate the pretraining corpora into documents. Then we randomly select documents until the target number of words is exceeded. Since the training corpora contain long documents such as novels and movie scripts, the actual number of words varies from the target slightly.

6.3.3 Architecture and Training Details

6.3.3.1 RoBERTa

RoBERTa (Liu et al., 2019b) is one of the most widely used Transformer-based masked language model architectures. It is based closely on the influential BERT family of models (Devlin et al., 2019). The models are trained in fairseq[7] with hyper-parameters sampled following Warstadt et al. (2020b). For each of the 16 conditions, we initially train five instances from scratch. Using development set perplexity, we then select the three best instances from that condition to study in subsequent experiments.

6.3.3.2 5-gram baseline

We train 5-gram language models using KenLM (Heafield, 2011). The model implements modified Kneser-Ney smoothing with backoff (Heafield et al., 2013b). For these baselines, we choose to train them only on data from the WRITTEN CONTROL condition.

The purpose of this baseline is to place an upper bound on performance on our evaluation data using only shallow co-occurrence features. An n-gram model cannot make interesting abstractions like MOVE-MAIN or MOVE-FIRST, or generalize meaningfully to cases where four more words intervene between the auxiliary and the main verb. Thus, we can only attribute the use of such abstractions to our neural LMs if their behavior differs systematically from this baseline.

[7] Link to our fork: https://github.com/YianZhang/fairseq

Source	Tokens	Tokens Excluded	Tokens Included	% Excluded
Books	1.18B	58.6M	1.12B	5.0%
Wikipedia	1.45B	8.1M	1.44B	0.6%
Written training	0.97B	16.1M	0.95B	1.7%

Table 6.3: Quantity of data (measured by word tokens) filtered from the written training data. Books and Wikipedia refer to the entire preprocessed corpora, not just portions sampled for training. Written training refers to just the data used for training, which was sampled in a 3:1 ratio of Wikipedia to Books, following Devlin et al. (2019).

6.3.4 Treatments: Filtered vs. Control

We obtain filtered data by first applying the syntactic filter described above to the control data. We use the spaCy sentence segmenter to split the data into sentences, and we retain the original order of sentences within a document post-filtering. Since filtering lowers the number of words, we supplement the data post-filtering with data from the same domains, to ensure that datasets from both treatments contain approximately the same number of words. The supplemental data has also been filtered.

Statistics about filtered data are given in Table 6.3. We observe that a much greater proportion of sentences from the Books domain are filtered compared to the Wikipedia domain. This is likely due to the fact that interrogatives are most likely to be present in dialogue, which is far more common in books than in encyclopedia articles.

6.3.5 Domains: Written vs. Spoken

The training data for our models is from one of two domains: written English or spoken English. Three-fourths of our written data is from English wikipedia, and the remaining fourth is self-published books scraped from Smashwords. This combination was shown to be effective in the training of BERT (Devlin et al., 2019) and the miniBERTas (Warstadt et al., 2020b). However, from a cognitive modeling perspective, spoken data is preferable.[8] Since we are not aware of any corpora of transcribed speech on the scale of 100M or 1B of words, we use the English portion of the OpenSubtitles corpus (Lison and Tiedemann, 2016), which consists of over 1B words of scripted and unscripted subtitles from television and film.

6.4 Experiments: The Effect of Syntactic Filtering on Unsupervised Acceptability Judgments

Our primary experiments use the targeted syntactic evaluation paradigm (Marvin and Linzen, 2018; Warstadt et al., 2020a; Hu et al., 2020) to study which kinds of acceptability judgments are impacted by the application of the syntactic filter. Targeted syntactic evaluation is a method for extracting acceptability judgments from language models without task-specific supervision. Evaluation data consists of sentences in minimal pairs of the form $(S_{\text{good}}, S_{\text{bad}})$. The language model is used to estimate probabilities for each sentence, and we consider its prediction correct if the following inequality holds:

$$P_{LM}(S_{\text{good}}) > P_{LM}(S_{\text{bad}}).$$

[8] One could take this argument quite far, and attempt to train models on transcripts of child directed speech, transcriptions of environmental speech, or even audio recordings. Since large datasets from these domains are limited, as discussed in greater detail in Chapter 1, we leave these possibilities to future work.

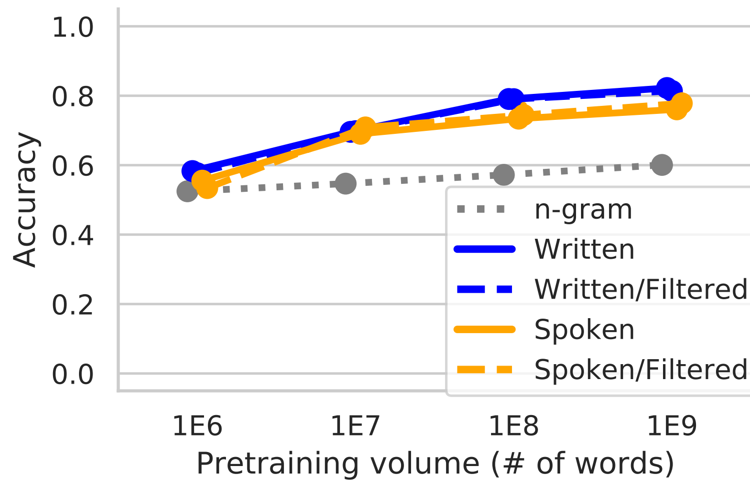

Figure 6.3: Performance of models from all conditions on BLiMP overall.

For masked language models like the RoBERTa-style models we evaluate, a sentence can be scored using by sequentially masking one token at a time, following the approximation used by Wang and Cho (2019) and Salazar et al. (2020):

$$P_{MLM}(S) = \prod_{i=1}^{|S|} P_{MLM}(t_i | S_{[i \backslash \text{MASK}]}).$$

6.4.1 Evaluation on BLiMP

Prior to testing our models on subject auxiliary inversion, we test them on BLiMP (Warstadt et al., 2020a) as a control. BLiMP contains 67 different minimal pair types representing many phenomena in English morphosyntax, syntax, and semantics. The idea behind this control is to test whether the syntactic filtering manipulation had widespread effects on acceptability judgments in our models. We see no reasons *a priori* why the removal of complex subject auxiliary inversion sentences should have

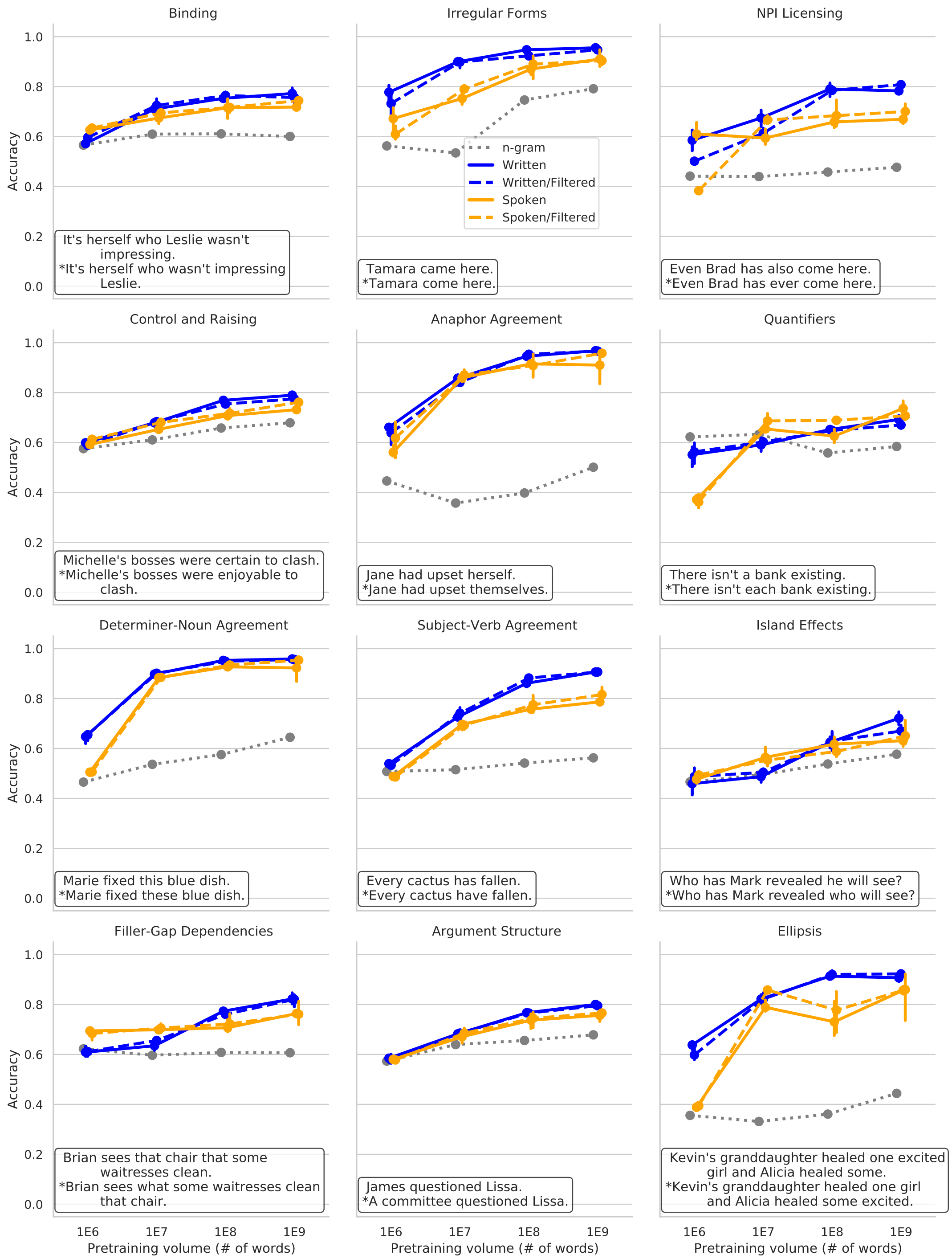

Figure 6.4: Performance of models from all conditions on BLiMP by category.

any impact on the ability of models to detect acceptability contrasts related to phenomena in BLiMP like argument structure, determiner-noun agreement, or reflexive binding. Still, there could be some reasons for doubt. First, the filter has low precision, and the needless removal of a large number of interrogatives is a confound that might harm syntactic generalization. Second, the indirect evidence hypothesis itself suggests that removing one kind of evidence from the input might have unexpected consequences in other domains of grammar.[9]

Results on BLiMP show quite clearly that filtering indeed had little to no effect on general acceptability judgments. Figure 6.3 shows overall BLiMP performance. While we do observe that models trained in the written domain tend to do better than models trained on spoken data, we see no effect of filtering on overall performance. We also note, encouragingly, that n-gram performance is only slightly above chance.

Looking at performance on individual phenomena in BLiMP (Figure 6.4) bolsters the evidence that filtering had little impact on general grammatical generalization in the models. As before we see a consistent and expected difference between the spoken and written domains, and between the neural LMs and the n-gram models. But we do not see any noteworthy difference due to filtering in any of the twelve categories in BLiMP. Among the twelve categories, filler-gap dependencies and island effects would have been the most likely to be affected. However, the general reduction in frequency of interrogatives due to filtering does not seem to have made these contrasts any more difficult to learn.

[9]On the whole though, this goes against the spirit of the indirect evidence hypothesis, which is usually invoked to argue that learners are robust in the absence of evidence. It is reasonable to think that systematically ablating certain kinds of evidence could have unintended harmful consequences, but I hypothesize that such effects will be difficult to observe when the ablated evidence is a rare, specific construction, as in the current experiment.

Name	Template/Example
MOVE-FIRST or MOVE-MAIN A	⟨NP⟩ ⟨Aux⟩$ ⟨V/Pred⟩ ⟨NP⟩ ⟨Rel⟩ ⟨Aux⟩* ⟨VP/pred⟩ Ants are a curiosity to the children that are playing outside.
MOVE-FIRST or MOVE-MAIN B	⟨pron_3sg⟩ has* ⟨V_t⟩ ⟨N⟩ ⟨N⟩ was$ ⟨V_t⟩ ⟨PP⟩ He has stolen the money his father was storing in the safe.
MOVE-FIRST or MOVE-MAIN C	Every ⟨N_sg⟩ is* still ⟨V_t⟩ ⟨N⟩ ⟨Pro_3sg⟩ has$ been ⟨V_t⟩ since ⟨VP/N⟩ Every child is still eating the sandwich she has been nibbling on since 2pm.
MOVE-FIRST or MOVE-MAIN D	⟨NP_pl⟩ ⟨Modal⟩$ ⟨V_bare⟩ ⟨NP_pl⟩ ⟨Rel⟩ (do)* ⟨V_pres⟩ The auditors will go after people who (do) cheat on their taxes.
Only MOVE-MAIN A	⟨NP⟩ ⟨Rel⟩ ⟨Aux⟩* ⟨VP/pred⟩ ⟨Aux⟩$ ⟨VP/pred⟩ Plants that couldn't adapt to climate change have died out.
Only MOVE-MAIN B	The ⟨Noun_sg⟩ ⟨pron_3sg⟩ is$ ⟨V_t⟩ was* ⟨VP/pred⟩. The string quartet he is rehearsing was composed by Mozart.
Only MOVE-MAIN C	Every ⟨noun_sg⟩ ⟨pron_3sg⟩ has$ ever ⟨V_t⟩ has* ⟨V_i⟩ ⟨adv⟩. Every dog he has ever tried to pick up has barked loudly.
Only MOVE-MAIN D	⟨NP_pl⟩ ⟨Rel⟩ (do)* ⟨V_pres⟩ ⟨Modal⟩$ ⟨V_bare⟩ Heirloom apples that (do) receive special care will taste delicious in a pie.

Table 6.4: Templates for the subject auxiliary inversion evaluation data, with representative examples.

6.4.2 Evaluation on Subject Auxiliary Inversion

Having established that syntactic filtering had no general effects of grammatical generalization, we now now investigate whether it had a localized effect on subject auxiliary inversion, and in particular the learning of MOVE-MAIN over MOVE-FIRST.

6.4.2.1 Evaluation Data

We generate evaluation data in minimal pairs from templates. There are 8 templates, each specifying a fixed ordering of constituents which three of the authors semi-manually populated with content. There are many tradeoffs to consider with such data creation. We opt for a large quantity of high quality, semantically plausible data, at the expense of diversity. Within each of the templates, we initially wrote

20 completely different items. Then, to increase quantity and diversity, we added variations to each of the 20 items. For instance, taking the Cartesian product of the following set of variations, we can generate 36 unique minimal pairs. The templates are described in greater detail in Table 6.4.

$$\begin{array}{ccccccc} NP & Aux & V & NP & Rel & Aux & VP \\ \left\{\begin{array}{c}\text{tomorrow's election}\\ \text{this week's debate}\end{array}\right\} & will & \left\{\begin{array}{c}\text{attract}\\ \text{draw}\\ \text{bring in}\end{array}\right\} & \left\{\begin{array}{c}\text{young voters}\\ \text{participants}\\ \text{live broadcasters}\end{array}\right\} & that & had & \left\{\begin{array}{c}\text{been politically aloof}\\ \text{shown little interest}\end{array}\right\} \end{array}$$

6.4.2.2 The "Classic" Examples

The majority of the discussion about subject auxiliary inversion has focused on examples with a single subject relative clause. The following examples both fit this description, but differ in the position of the relative clause with respect to the main auxiliary, and thus the applicability of the MOVE-FIRST rule.

(16) a. Are centipedes and millipedes a curiosity to the students that are playing outside?

b. *Are centipedes and millipedes are a curiosity to the students that playing outside?

(17) a. Are the pushbuttons that are most frequently used on the control console most vulnerable?

b. *Are the pushbuttons that most frequently used on the control console are most vulnerable?

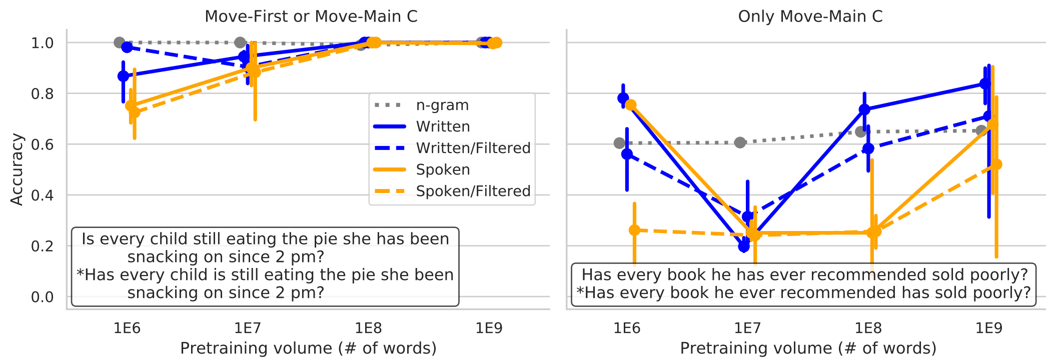

Figure 6.5: Model performance on "classic" subject auxiliary inversion examples with a single subject relative clause. The MOVE-MAIN and MOVE-FIRST rules give the same predictions for the template on the left, and opposite predictions for the template on the right.

Results for these templates are given in Figure 6.5. Performance is above chance for all models, and above the n-gram baseline for all models with 10M words or more of training data. Furthermore, performance is substantially above baseline for models with 100M words or more. This rules out the possibility that any of these models systematically acquire the MOVE-FIRST generalization. The results are similar for models trained on filtered data, suggesting that even with little to no direct evidence against MOVE-FIRST, none of our model learners preferentially accept that hypothesis as the rule for subject auxiliary inversion.

This alone does not rule out that our models could sometime rely on the erroneous MOVE-FIRST rule as one of an ensemble of strategies. Indeed, comparing the left and right sides of Figure 6.5, we see that similar models consistently perform worse on the template that is inconsistent with MOVE-FIRST. By way of example, this is what we would expect to see if the model relied MOVE-FIRST with some probability $p < 0.5$ and MOVE-MAIN or some other correlated heuristic with probability $1 - p$.

6.4.2.3 The Local n-gram Confound

One possible heuristic is the presence of low probability bigrams common in the ungrammatical sentences from these templates. Specifically, Kam et al. (2008) notice that the bigram of a relativizer followed by a non-finite verb, for instance *that playing* in (16b), is sufficient to deflate the likelihood of the ungrammatical sentence for a model that lacks the representations to identify the main auxiliary.

This next set of results remove the spurious correlation between acceptability and these low probability bigrams. The B and C templates from Table 6.4 use object relatives without a relativizer to eliminate the spurious bigram correlation. The D templates get around this by having a third person plural subject and a present tense verb for the relative clause (e.g. *the articles that offend people*). This allows the complex NP to have the same string in both the grammatical and ungrammatical example.

As shown in Figure 6.6, performance is quite different depending on the position of the relative clause. On the left hand side, these examples are consistent with both MOVE-MAIN and MOVE-FIRST. At 100M words or more, performance of all models, regardless of filtering, is near perfect for the B and C templates, and high (though close to the n-gram baseline) for the D template.

For the templates on the right hand side, MOVE-FIRST gives the incorrect prediction. Notably, the models are quite often worse than the n-gram baseline. This means that whatever features the neural LMs use to score sentences are less reliable than shallow co-occurrence features. Again, this finding is consistent with the explanation that the neural models are using an ensemble of heuristics, of which MOVE-FIRST is one. In fact, for the C and D templates, some neural models perform

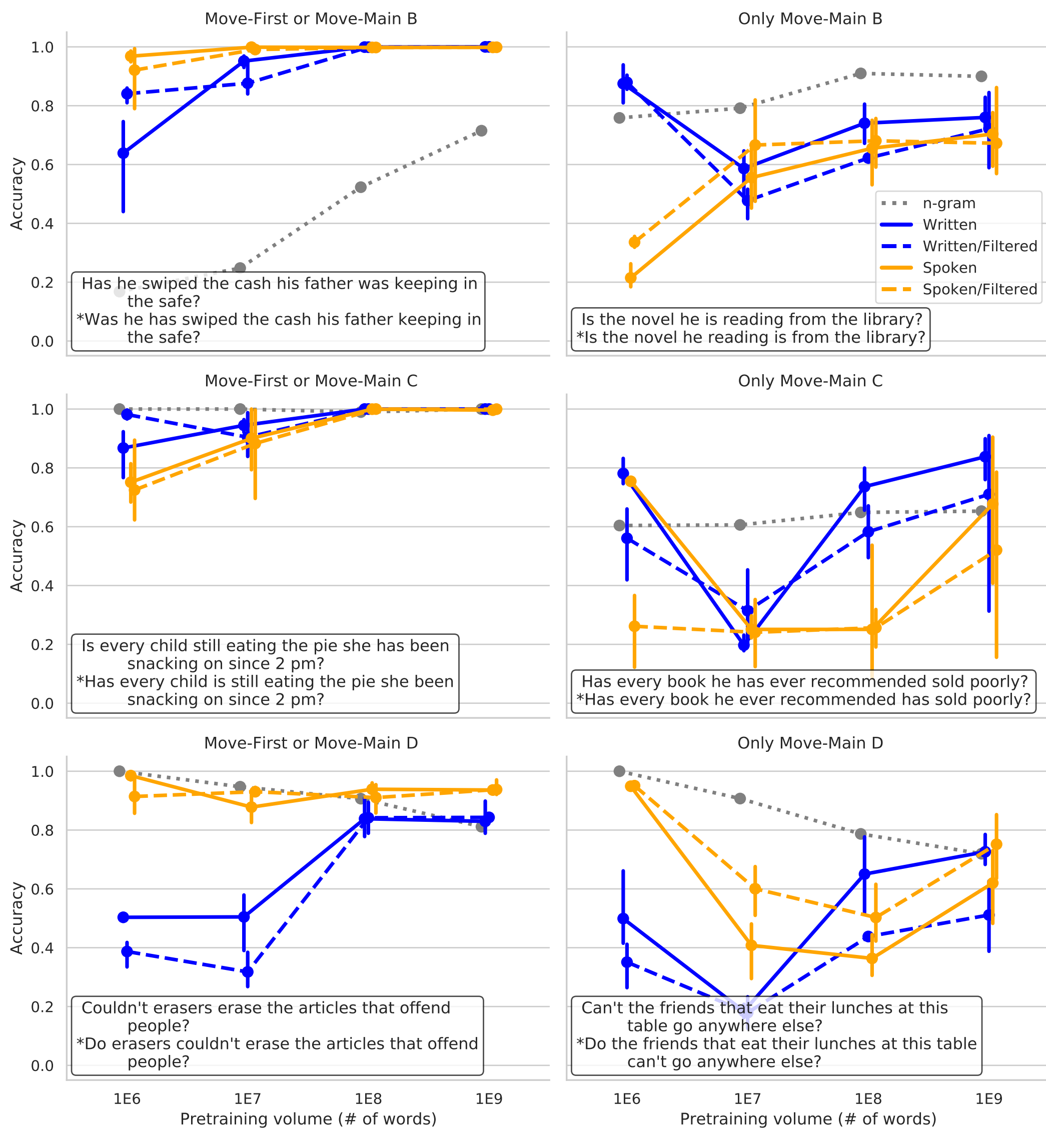

Figure 6.6: Model performance on examples of complex subject auxiliary inversion without the spurious bigram correlation identified by Kam et al. (2008).

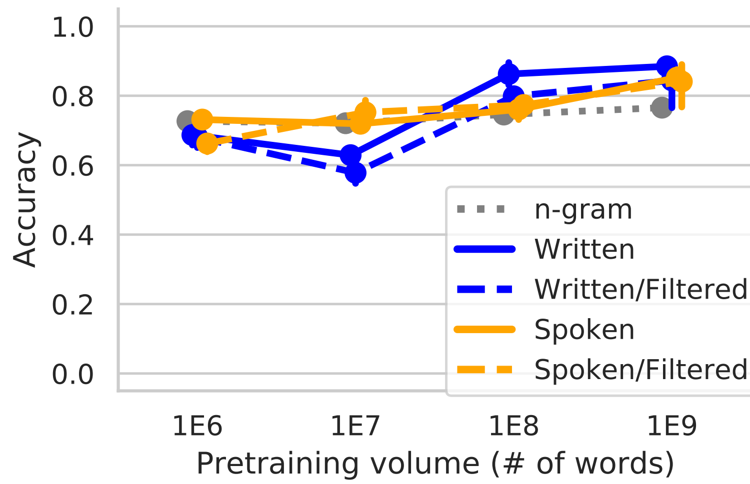

Figure 6.7: Model performance overall on all 8 subject auxiliary inversion test cases.

close to floor, consistent with MOVE-FIRST being the dominant determinant of the model's predictions (though only for this one template).

We also observe a somewhat inconsistent difference between the filtered and control treatments for the ONLY MOVE-MAIN templates. The models trained in the filtered environment often show a behavior consistent with applying the MOVE-FIRST strategy more often—especially for those trained in the written domain.

6.4.2.4 The Effect of Filtering and Domain

From these results, there appears to be a generally negative affect of filtering on models trained in the written domain, and little effect (if not a slightly positive one) of filtering in the spoken domain. We confirm this impression by plotting overall performance on the eight subject auxiliary inversion test cases in Figure 6.7.

We also observe that the performance of models trained in the written domain increases dramatically between 10M and 100M words—a finding consistent with the

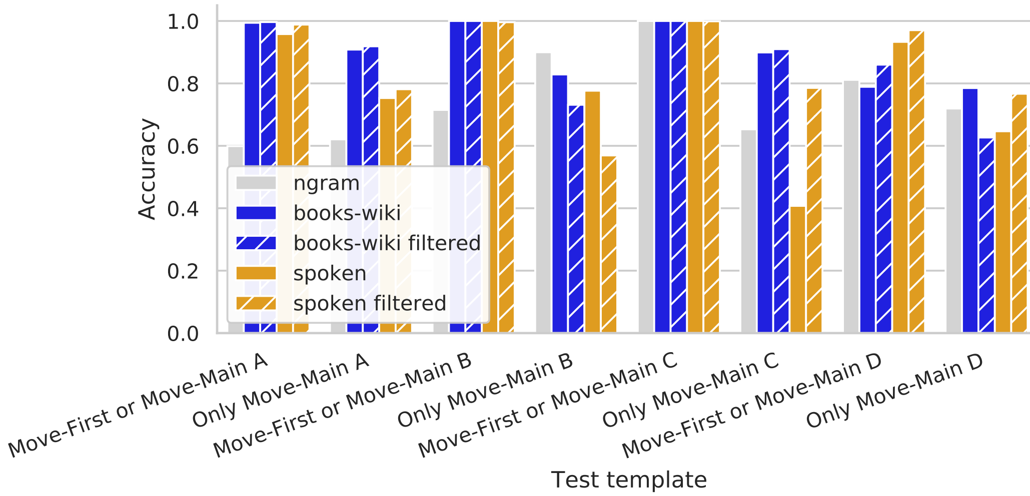

Figure 6.8: Performance of 1B-word models with highest performance on BLiMP from the given condition.

BLiMP learning curves reported in Chapter 5—while models trained in the spoken domain show more gradual signs of improvement, and generally lie closer to the n-gram baseline.

6.4.2.5 Best-Case Performance

The results above all focus on average-case performance over the three model instances trained for each condition. However, results are qualitatively different when we consider only best-case performance. In Figure 6.8 we select the best 1B word model from each condition based on overall BLiMP performance.[10] Again, there is a tendency for performance to be lower on templates that are inconsistent with MOVE-FIRST, suggesting this strategy might be applied in a subset of cases. But for all templates, best-case performance for test cases consistent only with MOVE-MAIN

[10] We do model selection based on BLiMP to avoid biasing the results towards models that are good at subject auxiliary inversion at the expense of other grammatical knowledge.

is nearly always above chance, meaning MOVE-FIRST cannot be a dominant strategy under any condition. Furthermore, for models trained in the written domain, performance is usually near or above the n-gram baseline. Finally, there is no clear sign that filtering had an effect on best-case performance at the 1B word scale. In most cases, filtered and unfiltered models are comparable in performance, and in cases where they differ, filtering does not always have a harmful effect.

6.5 Discussion

Our primary goal was to determine the role of indirect evidence in rejecting the MOVE-FIRST rule for subject auxiliary inversion. On this question, our results are somewhat nuanced. We found that in the best case, models could achieve similarly strong performance on nearly all subject auxiliary inversion test cases whether or not direct evidence against MOVE-FIRST was filtered out of their input. On top of this, we did not observe any models clearly adopting MOVE-FIRST, even in the filtered conditions. In the subset of test cases and data quantities where model predictions were consistent with applying MOVE-FIRST as a dominant strategy, this was observed in both the filtered and unfiltered conditions, and models in all conditions and test cases rejected this hypothesis given sufficient pretraining data. These findings support the Indirect Evidence Hypothesis that a learner without a prior hierarchical bias can rule out linear generalizations simply through an abundance of indirect evidence that language is hierarchical.

However, this conclusion comes with several caveats. First, we only obtain a clearly positive result for the 1B word models. For models trained on less data, performance falls at or below the n-gram baseline, making it conceivable that they actually rely mainly on simple co-occurrence heuristics rather than MOVE-MAIN. The likelihood of this explanation goes down as we see success in a wider variety of test cases, as different test cases share fewer coincidental surface cues.

Another caveat is that we found that the syntactic filtering manipulation did have a consistent negative effect on subject auxiliary inversion predictions for models trained in the written domain. No such effect was observed on BLiMP examples, suggesting that the removal of direct evidence did in fact have a causal and targeted effect on the subject auxiliary inversion. On the other hand, it is not clear that this effect is disproportionately large for ONLY MOVE-MAIN test cases. Unless there is such an interaction, it is more likely that the harm caused by filtering is due to the removal of a large number interrogatives of all kinds (i.e. false positives caught by the filter).

Yet another caveat is that we do observe that accuracy is generally lower for ONLY MOVE-MAIN than for MOVE-FIRST OR MOVE-MAIN test cases. One explanation for this finding is that all or most models are applying an ensemble of strategies, of which MOVE-FIRST is one. Under this interpretation, while MOVE-FIRST is rarely a dominant strategy, if there is any probability that this heuristic will be applied, it should only reduce performance on the ONLY MOVE-MAIN test cases. An alternative explanation is that it is the increased linear distance between the fronted auxiliary and the main verb, rather than the presence of an invervening distractor auxiliary, that results in lower performance on these cases. This hypothesis can be tested in future work with additional test cases.[11]

[11]For example, in examples (ia) and (ib) below, the difference in condition is correlated with a difference in the distance between the fronted auxiliary and the main verb. Examples such as (ic) eliminate this confound: Despite the longer distance between the auxiliary and the main verb, this example is still consistent with move-first because the added modifier on the subject does not contain a finite distractor verb.

A final caveat is that the success of the models trained on 1B words in the filtered environment might be due to exposure to a greater amount of direct evidence that was not captured by the filter. According to our earlier estimate, these models could be exposed to approximately 1000 instances of direct evidence against MOVE-FIRST. While these examples do not make up a greater proportion of the input at this scale, there may be some absolute threshold over which the evidence is sufficient to reject MOVE-FIRST, as suggested by Legate and Yang (2002). However, this explanation predicts that the unfiltered models would reject MOVE-FIRST with just 1% of the input given to the filtered models, since they have 100 times the quantity of direct evidence. This is not what we observe. For example, in test template ONLY MOVE-MAIN C in Figure 6.6, we observe some filtered models are clearly able to reject MOVE-FIRST with 1B words of input. By the threshold argument, then, we should expect some unfiltered models to reject MOVE-FIRST with just 10M words. Instead, we find that the filtered and unfiltered models show similar predictions at this scale, suggesting that the absolute quantity of direct evidence is not the main factor driving these changes in performance, though it likely still plays some role. In future work, we plan to counteract the affect of direct evidence against MOVE-FIRST that passes through the filter by injecting comparable evidence in favor of MOVE-FIRST.

(i) a. Has the man who John is helping seen the cat? — MOVE-FIRST only; distance=6
    b. Has the man seen the cat who John is helping? — MOVE-FIRST or MOVE-MAIN; distance=2
    c. Has the man being helped by John seen the cat? — MOVE-FIRST or MOVE-MAIN; distance=6

6.6 Conclusion

Where does this leave us on the Poverty of the Stimulus and the innateness hypothesis? We have not shown that humans learn MOVE-MAIN without the benefit of an innate hierarchical bias. However, we have shown that MOVE-FIRST is not a particularly attractive hypothesis for data-driven learners trained in a naturalistic setting. It seems likely that the input actually *does* provide evidence that favors MOVE-MAIN, even if it is difficult to pin the notion of evidence to a cohesive set of examples, and even if the evidence is largely still consistent with MOVE-FIRST. In other words, the stimulus may be richer than is often acknowledged. This is still consistent with direct evidence being helpful. However, it highlights the importance of looking beyond direct evidence in deciding whether the input is sufficient for learning a particular target phenomenon.

While the results from this study are not totally conclusive, it makes a more substantial contribution. It is a first step towards proving the viability of a new methodology with the potential to give new decisive evidence on long-standing questions in the study of language acquisition. We have argued that debates in language acquisition no longer need to rely on speculation about what constitutes sufficient evidence for or against a hypothesis. As our experiments show, we have the tools to conduct experiments on model learners trained on the quantity and variety of linguistic input available to children, and to make targeted manipulations of their input to draw causal inferences about the effects of variables in the input on grammatical generalization.

Bibliography

Omri Abend, Tom Kwiatkowski, Nathaniel J. Smith, Sharon Goldwater, and Mark Steedman. Bootstrapping language acquisition. *Cognition*, 164:116–143, July 2017. ISSN 00100277. doi: 10.1016/j.cognition.2017.02.009. URL https://linkinghub.elsevier.com/retrieve/pii/S0010027717300495.

David Adger. *Core Syntax: A Minimalist Approach*. Oxford University Press Oxford, 2003.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR Conference Track. Toulon, France.*, 2017.

Aixiu An, Peng Qian, Ethan Wilcox, and Roger Levy. Representation of Constituents in Neural Language Models: Coordination Phrase as a Case Study. *arXiv preprint arXiv:1909.04625*, 2019.

JL Austin. *How to Do Things With Words*. Oxford University Press, 1962.

Carl Lee Baker. *Introduction to Generative-Transformational Synta*. Prentice-Hall, Englewood Cliffs, NJ, 1978.

Mark R. Baltin. A Landing Site Theory of Movement Rules. *Linguistic Inquiry*, 13 (1):1–38, 1982. Publisher: JSTOR.

Mark R. Baltin and Chris Collins, editors. *Handbook of Contemporary Syntactic Theory*. Blackwell Publishing Ltd, 2001. doi: 10.1111/b.9781405102537.2003.x. URL https://doi.org/10.1111%2Fb.9781405102537.2003.x.

Roy Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second PASCAL recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, 2006. URL http://u.cs.biu.ac.il/~nlp/RTE2/Proceedings/01.pdf.

Marco Baroni. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *arXiv:2106.08694 [cs]*, June 2021. URL http://arxiv.org/abs/2106.08694. arXiv: 2106.08694.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstral, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. 2018.

Yonatan Belinkov and James R. Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7: 49–72, 2019a. URL https://www.aclweb.org/anthology/Q19-1004.pdf.

Yonatan Belinkov and James R. Glass. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019b. URL https://www.aclweb.org/anthology/Q19-1004.pdf.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. The Fifth PASCAL Recognizing Textual Entailment Challenge. In *Textual Analysis Conference (TAC)*, 2009. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.232.1231.

Robert C Berwick, Paul Pietroski, Beracah Yankama, and Noam Chomsky. Poverty of the stimulus revisited. *Cognitive Science*, 35(7):1207–1242, 2011. Publisher: Wiley Online Library.

Steven Bird and Edward Loper. NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics, 2004.

Kathryn Bock and Carol A. Miller. Broken agreement. *Cognitive psychology*, 23(1): 45–93, 1991. Publisher: Elsevier.

Diane Bouchacourt and Marco Baroni. How agents see things: On visual representations in an emergent language game. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 981–985, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1119. URL https://aclanthology.org/D18-1119.

Joan W. Bresnan. Syntax of the comparative clause construction in English. *Linguistic Inquiry*, 4(3):275–343, 1973.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, 2020. URL https://papers.nips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Andrew Carnie. *Syntax: A Generative Introduction*. John Wiley & Sons, 2013.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. Anti-efficient encoding in emergent communication. *Advances in Neural Information Processing Systems*, 32, 2019.

Rahma Chaabouni, Florian Strub, Florent Altché, Eugene Tarassov, Corentin Tallec, Elnaz Davoodi, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot. Emergent communication at scale. In *International conference on learning representations*, 2022.

Tyler A. Chang and Benjamin K. Bergen. Word Acquisition in Neural Language Models. *Transactions of the Association for Computational Linguistics*, 10:1–16, January 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00444. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00444/109271/Word-Acquisition-in-Neural-Language-Models.

Nick Chater and Paul Vitányi. Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7(1):19–22, January 2003. ISSN 13646613. doi: 10.1016/S1364-6613(02)00005-0. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661302000050.

Rui P. Chaves. What Don't RNN Language Models Learn About Filler-Gap Dependencies? In *Proceedings of the third meeting of the Society for Computation in Linguistics (SCiL)*, 2020.

Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394, 1999. Publisher: Elsevier.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. In *European conference on computer vision*, pages 740–755. Springer, April 2015. URL http://arxiv.org/abs/1504.00325. arXiv: 1504.00325.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal image-text representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer vision – ECCV 2020*, pages 104–120, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58577-8.

Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. Pretrained Language Model Embryology: The Birth of ALBERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6813–6828, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.553. URL https://www.aclweb.org/anthology/2020.emnlp-main.553.

Gennaro Chierchia. *Logic in Grammar*. Oxford University Press, July 2013. doi: 10.1093/acprof:oso/9780199697977.001.0001. URL https://doi.org/10.1093%2Facprof%3Aoso%2F9780199697977.001.0001.

Noam Chomsky. *Syntactic Structures.* Mouton, 1957.

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

Noam Chomsky. Problems of knowledge and freedom: The Russell lectures. 1971.

Noam Chomsky. *Lectures on government and binding*. 1981.

Noam Chomsky. *The Minimalist Program*. MIT press, 1995.

Noam Chomsky. Biolinguistic Explorations: Design, Development, Evolution. *International Journal of Philosophical Studies*, 15(1):1–21, January 2007. ISSN 0967-2559, 1466-4542. doi: 10.1080/09672550601143078. URL http://www.tandfonline.com/doi/abs/10.1080/09672550601143078.

Noam Chomsky and Howard Lasnik. The theory of principles and parameters. In *The minimalist program*. MIT Press, 1993.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. The grammar-learning trajectories of neural language models. *arXiv preprint arXiv:2109.06096*, 2021.

Michelle M Chouinard and Eve V Clark. Adult reformulations of child errors as negative evidence. *Journal of child language*, 30(3):637–669, 2003. Publisher: Cambridge University Press.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and others. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Shammur Absar Chowdhury and Roberto Zamparelli. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th international conference on computational linguistics*, pages 133–144, 2018.

Shammur Absar Chowdhury and Roberto Zamparelli. An LSTM adaptation study of (un) grammaticality. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 204–212, 2019.

Morten H Christiansen and Nick Chater. *Creating language: Integrating evolution, acquisition, and processing*. MIT Press, 2016.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.

arXiv:1412.3555 [cs], December 2014. URL http://arxiv.org/abs/1412.3555. arXiv: 1412.3555.

Sandra Chung, William A. Ladusaw, and James McCloskey. Sluicing and Logical Form. *Natural Language Semantics*, 3(3):239–282, 1995. Publisher: Springer.

Alexander Clark and Shalom Lappin. *Linguistic Nativism and the Poverty of the Stimulus*. John Wiley & Sons, 2011.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. 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 2924–2936, 2019a. URL https://www.aclweb.org/anthology/N19-1300.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019b.

Ariel Cohen and Manfred Krifka. Superlative quantifiers and meta-speech acts. *Linguistics and Philosophy*, 37(1):41–90, 2014. Publisher: Springer.

Chris Collins. A Smuggling Approach to the Passive in English. *Syntax*, 8(2):81–120, 2005. Publisher: Wiley Online Library.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Em-

pirical Methods in Natural Language Processing, pages 670–680. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1070. URL http://aclweb.org/anthology/D17-1070. event-place: Copenhagen, Denmark.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single &!#* vector: Probing sentence embeddings for linguistic properties. In *ACL 2018-56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2126–2136. Association for Computational Linguistics, 2018.

G. G. Coulton. The Princes of the World. In *From St. Francis to Dante*, Translations from the Chronicle of the Franciscan Salimbene, 1221-1288, pages 239–256. University of Pennsylvania Press, 2 edition, 1972. ISBN 978-0-8122-7672-5. URL https://www.jstor.org/stable/j.ctv4t8279.25.

Stephen Crain and Mineharu Nakayama. Structure dependence in grammar formation. *Language*, pages 522–543, 1987. Publisher: JSTOR.

Alejandrina Cristia, Emmanuel Dupoux, Michael Gurven, and Jonathan Stieglitz. Child-Directed Speech Is Infrequent in a Forager-Farmer Population: A Time Allocation Study. *Child Development*, 90(3):759–773, May 2019. ISSN 0009-3920, 1467-8624. doi: 10.1111/cdev.12974. URL https://onlinelibrary.wiley.com/doi/10.1111/cdev.12974.

Peter W. Culicover and Ray Jackendoff. The view from the periphery: The English comparative correlative. *Linguistic Inquiry*, 30(4):543–571, 1999.

Jillian K. Da Costa and Rui P. Chaves. Assessing the ability of Transformer-based Neural Models to represent structurally unbounded dependencies. In *Proceedings of the third meeting of the Society for Computation in Linguistics (SCiL)*, 2020.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer, 2006.

Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in neural information processing systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/7137debd45ae4d0ab9aa953017286b20-Paper.pdf.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 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, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285. URL https://www.aclweb.org/anthology/P19-1285.

Mark Davies. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, 14(2):159–190, 2009. Publisher: John Benjamins.

Veneeta Dayal. Any as inherently modal. *Linguistics and Philosophy*, 21(5):433–476, 1998.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. The CommitmentBank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124, 2019. URL https://semanticsarchive.net/Archive/Tg3ZGI2M/Marneffe.pdf. Issue: 2.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019. URL https://www.aclweb.org/anthology/N19-1423.

Bhuwan Dhingra, Qiao Jin, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. Neural Models for Reasoning over Multiple Mentions Using Coreference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 42–48, 2018.

Emmanuel Dupoux. Cognitive Science in the era of Artificial Intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59, April 2018. ISSN 00100277. doi: 10.1016/j.cognition.2017.11.008. URL http://arxiv.org/abs/1607.08723. arXiv: 1607.08723.

Gabe Dupre. (What) Can Deep Learning Contribute to Theoretical Linguistics? *Minds and Machines*, September 2021. ISSN 0924-6495, 1572-8641. doi: 10.1007/s11023-021-09571-w. URL https://link.springer.com/10.1007/s11023-021-09571-w.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A Smith. Recurrent Neural Network Grammars. In *Proceedings of NAACL-HLT*, pages 199–209, 2016.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021a.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021b.

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

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018. URL https://www.aclweb.org/anthology/L18-1544.

Allyson Ettinger. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 2020. doi: 10.1162/tacl_a_00298. URL https://doi.org/10.1162/tacl_a_00298.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 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, pages 134–139, 2016. URL https://www.aclweb.org/anthology/W16-2524.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801. Association for Computational Linguistics, 2018. Journal Abbreviation: arXiv preprint arXiv:1809.03992.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. Causal Analysis of Syntactic Agreement Mechanisms in Neural Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.144. URL https://aclanthology.org/2021.acl-long.144.

Janet Dean Fodor and Carrie Crowther. Understanding stimulus poverty arguments. *The Linguistic Review*, 18(1-2), January 2002. ISSN 0167-6318, 1613-3676. doi: 10.1515/tlir.19.1-2.105. URL https://www.degruyter.com/document/doi/10.1515/tlir.19.1-2.105/html.

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

Michael C. Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A. Marchman. Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, 44(3):677–694, May 2017. ISSN 0305-0009, 1469-7602.

doi: 10.1017/S0305000916000209. URL https://www.cambridge.org/core/product/identifier/S0305000916000209/type/journal_article.

Robert Frank and Donald Mathis. Transformational networks. *Models of Human Language Acquisition*, page 22, 2007.

Victoria Fromkin, Stephen Krashen, Susan Curtiss, David Rigler, and Marilyn Rigler. The Development of Language in Genie: a Case of Language Acquisition beyond the "Critical Period". *Brain and Language*, 1:81–107, 1974.

Richard Futrell and Roger P Levy. Do RNNs learn human-like abstract word order preferences? *Proceedings of the Society for Computation in Linguistics*, 2(1): 50–59, 2019.

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

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55(1):63–77, March 2021. ISSN 1574-020X, 1574-0218. doi: 10.1007/s10579-020-09503-7. URL https://link.springer.com/10.1007/s10579-020-09503-7.

Kanishk Gandhi and Brenden M Lake. Mutual exclusivity as a challenge for neural networks. *arXiv preprint arXiv:1906.10197*, 2019.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. SyntaxGym: An Online Platform for Targeted Evaluation of Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.10. URL https://aclanthology.org/2020.acl-demos.10.

Gerald Gazdar. Unbounded dependencies and coordinate structure. In *The Formal Complexity of Natural Language*, pages 183–226. Springer, 1981.

Dedre Gentner. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S Kuczaj, editor, *Language development: Language cognition and culture.*, page 48. 1982.

Bart Geurts and Rick Nouwen. 'At least' et al.: The semantics of scalar modifiers. *Language*, pages 533–559, 2007. Publisher: JSTOR.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Association for Computational Linguistics, 2007. URL https://www.aclweb.org/anthology/W07-1401.

Edward Gibson and Evelina Fedorenko. Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences*, 14(6):233–234, 2010.

Jill Gilkerson, Jeffrey A Richards, Steven F Warren, Judith K Montgomery, Charles R Greenwood, D Kimbrough Oller, John HL Hansen, and Terrance D Paul. Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology*, 26(2):248–265, 2017. Publisher: ASHA.

Lila Gleitman and Eric Wanner. Language Acquisition: The State of the Art. In Lila Gleitman and Eric Wanner, editors, *Language Acquisition: The State of the Art*. Cambridge University Press, 1982.

E Mark Gold. Language identification in the limit. *Information and control*, 10(5): 447–474, 1967. Publisher: Elsevier.

Adele E. Goldberg and Ray Jackendoff. The English resultative as a family of constructions. *Language*, 80(3):532–568, 2004.

Peter Gordon. Level-ordering in lexical development. *Cognition*, pages 73–93, 1985.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34, 2003.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. *Proceedings of the Society for Computation in Linguistics*, 2(1):363–364, 2019.

Kristina Gulordava, Thomas Brochhagen, and Gemma Boleda. Which one is the dax? achieving mutual exclusivity with neural networks. *arXiv preprint arXiv:2004.03902*, 2020.

Rebecca L Gómez and LouAnn Gerken. Infant artificial language learning and language acquisition. *Trends in cognitive sciences*, 4(5):178–186, 2000. Publisher: Elsevier.

John Hale. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001. URL https://aclanthology.org/N01-1021.

Chung-hye Han, Julien Musolino, and Jeffrey Lidz. Endogenous sources of variation in language acquisition. *Proceedings of the National Academy of Sciences*, 113(4):942–947, January 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1517094113. URL https://pnas.org/doi/full/10.1073/pnas.1517094113.

Harry F Harlow. The formation of learning sets. *Psychological review*, 56(1):51, 1949.

Betty Hart and Todd R. Risley. American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental Psychology*, 28(6):1096, 1992. URL https://psycnet.apa.org/fulltext/1993-09151-001.pdf. Publisher: American Psychological Association.

David Haussler. Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. *Artificial intelligence*, 36(2):177–221, 1988. Publisher: Elsevier.

David Haussler. Probably approximately correct learning. In *Proceedings of the eighth national conference on artificial intelligence*, pages 1101–1108. AAAI Press, 1990.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International conference on learning representations*, 2020.

Kenneth Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL https://aclanthology.org/W11-2123.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, 2013a.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 690–696, Sofia, Bulgaria, August 2013b. Association for Computational Linguistics. URL https://aclanthology.org/P13-2121.

Andrew Heathcote, Scott Brown, and Douglas JK Mewhort. The power law repealed: The case for an exponential law of practice. *Psychonomic bulletin & review*, 7(2): 185–207, 2000. URL https://link.springer.com/content/pdf/10.3758/BF03212979.pdf. Publisher: Springer.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. Predicting grammaticality on an ordinal scale. In *Proceed-*

ings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), volume 2, pages 174–180, 2014a.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 174–180, 2014b.

Irene Heim and Angelika Kratzer. *Semantics in generative grammar*, volume 1185. Blackwell Oxford, 1998.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/S10-1006.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. URL https://www.aclweb.org/anthology/D19-1275. event-place: Hong Kong.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. 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 4129–4138, 2019. URL https://www.aclweb.org/anthology/N19-1419.

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

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

Jeremy Howard and Sebastian Ruder. 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, 2018.

Steve R. Howell, Damian Jankowicz, and Suzanna Becker. A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language*, 53(2):258–276, August 2005. ISSN 0749596X. doi: 10.1016/j.jml.2005.03.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0749596X05000495.

Anne S. Hsu, Nick Chater, and Paul Vitányi. Language Learning From Positive Evidence, Reconsidered: A Simplicity-Based Approach. *Topics in Cognitive Science*, 5(1):35–55, January 2013. ISSN 17568757. doi: 10.1111/tops.12005. URL https://onlinelibrary.wiley.com/doi/10.1111/tops.12005.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. A Systematic Assessment of Syntactic Generalization in Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.158.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1243. URL https://aclanthology.org/D19-1243.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th conference on computational natural language learning*, pages 624–646, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.49. URL https://aclanthology.org/2021.conll-1.49.

Taichi Iki and Akiko Aizawa. Effect of Visual Extensions on Natural Language Understanding in Vision-and-Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2189–2196, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.167. URL https://aclanthology.org/2021.emnlp-main.167.

Ray S. Jackendoff. Gapping and related rules. *Linguistic Inquiry*, 2(1):21–35, 1971. Publisher: JSTOR.

Rohan Jha, Charles Lovering, and Ellie Pavlick. When does data augmentation help generalization in nlp? *arXiv preprint arXiv:2004.15012*, 2020.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, 2017.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.

Jaap Jumelet and Dieuwke Hupkes. Do Language Models Understand Anything? On the Ability of LSTMs to Understand Negative Polarity Items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, 2018.

Nirit Kadmon and Fred Landman. Any. *Linguistics and Philosophy*, 16(4):353–422, August 1993. doi: 10.1007/bf00985272. URL https://doi.org/10.1007%2Fbf00985272. Publisher: Springer Nature.

Xuân-Nga Cao Kam, Iglika Stoyneshka, Lidiya Tornyova, Janet D. Fodor, and William G. Sakas. Bigrams and the Richness of the Stimulus. *Cognitive Science*, 32(4):771–787, 2008. ISSN 1551-6709. doi: 10.1080/03640210802067053. URL https://onlinelibrary.wiley.com/doi/abs/10.1080/03640210802067053. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1080/03640210802067053.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. Verb Argument Structure Alternations in Word and Sentence Embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 287–297, 2019. doi: 10.7275/q5js-4y86. URL https://www.aclweb.org/anthology/W19-0129.

Slava Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing*, 35(3):400–401, 1987.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring Compositional Generalization: A Comprehensive Method on Realistic Data. *arXiv:1912.09713 [cs, stat]*, June 2020. URL http://arxiv.org/abs/1912.09713. arXiv: 1912.09713.

Jong-Bok Kim and Peter Sells. *English Syntax: An Introduction*. CSLI Publications, 2008.

Najoung Kim and Tal Linzen. COGS: A Compositional Generalization Challenge Based on Semantic Interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.731. URL https://aclanthology.org/2020.emnlp-main.731.

John P. Kimball. *The Formal Theory of Grammar*. Prentice-Hall, Englewood Cliffs, NJ, 1973.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2014. Journal Abbreviation: arXiv preprint arXiv:1412.6980.

Simon Kirby. *Function, selection, and innateness: The emergence of language universals*. OUP Oxford, 1999.

Ryan Kiros, Yukun Zhu, Ruslan R. Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3294–3302, 2015.

Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. Natural language does not emerge 'naturally' in multi-agent dialog. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2962–2967, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1321. URL https://aclanthology.org/D17-1321.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73, May 2017. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-016-0981-7. URL http://link.springer.com/10.1007/s11263-016-0981-7.

Dave Kush, Terje Lohndal, and Jon Sprouse. Investigating variation in island effects. *Natural language & linguistic theory*, 36(3):743–779, 2018. Publisher: Springer.

Brenden M Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, pages 2879–2888, 2018.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

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

Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Adelrahman Mohamed, and Emmanuel Dupoux. Generative Spoken Language Modeling from Raw Audio. *arXiv:2102.01192 [cs]*, September 2021. URL http://arxiv.org/abs/2102.01192. arXiv: 2102.01192.

Thomas K Landauer and Susan T Dutnais. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2):211–240, 1997.

Jey Han Lau, Alexander Clark, and Shalom Lappin. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241, 2016.

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

Marvin Lavechin, Maureen de Seyssel, Marianne Métais, Florian Metze, Abdelrahman Mohamed, Hervé BREDIN, Emmanuel Dupoux, and Alejandrina Cristia. Early phonetic learning from ecological audio: domain-general versus domain-specific mechanisms. 2022. Publisher: PsyArXiv.

Steve Lawrence, C. Lee Giles, and Sandiway Fong. Natural language grammatical inference with recurrent neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 12(1):126–140, 2000.

Angeliki Lazaridou and Marco Baroni. Emergent Multi-Agent Communication in the Deep Learning Era. *arXiv:2006.02419 [cs]*, July 2020. URL http://arxiv.org/abs/2006.02419. arXiv: 2006.02419.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining Language and Vision with a Multimodal Skip-gram Model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado, May 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1016. URL https://aclanthology.org/N15-1016.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-Agent Cooperation and the Emergence of (Natural) Language. In *International Conference on Learning Representations*, March 2017. URL http://arxiv.org/abs/1612.07182. arXiv: 1612.07182.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. Emergence of linguistic communication from referential games with symbolic and pixel input. In *International conference on learning representations*, 2018.

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. Multi-agent Communication meets Natural Language: Synergies between Functional and Structural Language Learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7663–7674, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.685. URL https://aclanthology.org/2020.acl-main.685.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521 (7553):436, 2015. Publisher: Nature Publishing Group.

Julie Anne Legate and Charles D Yang. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review*, 18(1-2):151–162, 2002. Publisher: Walter de Gruyter.

Nathaniel Leibowitz, Barak Baum, Giora Enden, and Amir Karniel. The exponential learning equation as a function of successful trials results in sigmoid performance. *Journal of Mathematical Psychology*, 54(3):338–340, 2010. URL https://www.sciencedirect.com/science/article/abs/pii/S0022249610000179. Publisher: Elsevier.

Fred Lerdahl, Ray S Jackendoff, and Ray Jackendoff. *A Generative Theory of Tonal Music*. MIT Press, 1983.

Beth Levin. *English Verb Classes and Alternations: A preliminary investigation*. University of Chicago Press, 1993.

Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, March 2008. ISSN 00100277. doi: 10.1016/j.cognition.2007.05.006. URL https://linkinghub.elsevier.com/retrieve/pii/S0010027707001436.

Fushan Li and Michael Bowling. Ease-of-teaching and language structure from emergent communication. *Advances in neural information processing systems*, 32, 2019.

Jeffrey Lidz, Sandra Waxman, and Jennifer Freedman. What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition*, 89(3):295–303, October 2003. ISSN 00100277. doi: 10.1016/S0010-0277(03)00116-1. URL https://linkinghub.elsevier.com/retrieve/pii/S0010027703001161.

Tal Linzen. What can linguistics and deep learning contribute to each other? Response to Pater. *Language*, 95(1):e99–e108, 2019. Publisher: Linguistic Society of America.

Tal Linzen and Marco Baroni. Syntactic Structure from Deep Learning. *Annual Reviews of Linguistics*, 2021.

Tal Linzen and Yohei Oseki. The reliability of acceptability judgments across languages. *Glossa: a journal of general linguistics*, 3(1), 2018. Publisher: Ubiquity Press.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

Pierre Lison and Jorg Tiedemann. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, page 7, 2016.

Haokun Liu, William Huang, Dhara Mungra, and Samuel R. Bowman. Precise task formalization matters in Winograd schema evaluations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8275–8280, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.664. URL https://www.aclweb.org/anthology/2020.emnlp-main.664.

Leo Z. Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. Probing Across Time: What Does RoBERTa Know and When? *CoRR*, abs/2104.07885, 2021. URL https://arxiv.org/abs/2104.07885. _eprint: 2104.07885.

Nelson Liu, Roy F Schwartz, and Noah A Smith. Challenge. *manuscript*, 2019a.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019b. URL http://arxiv.org/abs/1907.11692.

Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. Predicting Inductive Biases of Fine-tuned Models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=mNtmhaDkAr.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html.

Brian MacWhinney. *The CHILDES project: Tools for analyzing talk, Volume II: The database.* Psychology Press, 2014.

Kyle Mahowald, Peter Graff, Jeremy Hartman, and Edward Gibson. SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language*, 92(3):619–635, 2016. Publisher: Linguistic Society of America.

Jessica Mankewitz, Veronica Boyce, Brandon Waldon, Georgia Loukatou, Dhara Yu, Jesse Mu, Noah D Goodman, and Michael C Frank. Multi-party referential communication in complex strategic games. Publisher: PsyArXiv.

Christopher D. Manning. Computational Linguistics and Deep Learning. *Computational Linguistics*, 41(4):701–707, December 2015. ISSN 0891-2017, 1530-9312. doi: 10.1162/COLI_a_00239. URL https://direct.mit.edu/coli/article/41/4/701-707/1512.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48): 30046–30054, December 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1907367117. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1907367117.

Alec Marantz. Verbal argument structure: Events and participants. *Lingua*, 130: 152–168, 2013. Publisher: Elsevier.

Gary F Marcus. Negative evidence in language acquisition. *Cognition*, 46(1):53–85, 1993. Publisher: Elsevier.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.645. URL https://www.aclweb.org/anthology/2020.acl-main.645.

Rebecca Marvin and Tal Linzen. Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, 2018.

Brian W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

R Thomas McCoy, Robert Frank, and Tal Linzen. Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society.*, 2018.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *Proceedings of the Association for Computational Linguistics*, 2019.

R. Thomas McCoy, Robert Frank, and Tal Linzen. Does Syntax Need to Grow on Trees? Sources of Hierarchical Inductive Bias in Sequence-to-Sequence Networks. *Transactions of the Association for Computational Linguistics*, 8:125–140, December 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00304. URL https://direct.mit.edu/tacl/article/43542.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *CoRR*, abs/1609.07843, 2016. URL http://arxiv.org/abs/1609.07843.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. On the importance of pre-training data volume for compact language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.632.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černock, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.

George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 1995. URL https://www.aclweb.org/anthology/H94-1111. Publisher: Association for Computing Machinery.

Jim Miller. *An Introduction to English Syntax*. Edinburgh University Press, 2002.

Junghyun Min, R Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. Syntactic Data Augmentation Increases Robustness to Inference Heuristics. *arXiv preprint arXiv:2004.11999*, 2020.

Kanishka Misra. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*, 2022.

Tom M Mitchell. *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research ..., 1980.

Richard Montague. The proper treatment of quantification in ordinary English. In *Approaches to natural language*, pages 221–242. Springer, 1973.

Aaron Mueller, Robert Frank, Tal Linzen, Luheng Wang, and Sebastian Schuster. Coloring the blank slate: Pre-training imparts a hierarchical inductive bias to sequence-to-sequence models. In *Findings of the association for computational linguistics: ACL 2022*, pages 1352–1368, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-acl.106.

Karl Mulligan, Robert Frank, and Tal Linzen. Structure here, bias there: Hierarchical generalization by jointly learning syntactic transformations. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 125–135, Online, February 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.scil-1.12.

Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. *arXiv:2011.11588 [cs, eess]*, December 2020. URL http://arxiv.org/abs/2011.11588. arXiv: 2011.11588.

Joakim Nivre, Željko Agić, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Cristina Bosco, Sam Bowman, Giuseppe G. A. Celano, Miriam Connor, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Daniel Galbraith, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Berta Gonzales, Bruno Guillaume, Jan Hajič, Dag Haug, Radu Ion, Elena Irimia, Anders Johannsen, Hiroshi Kanayama, Jenna Kanerva, Simon Krek, Veronika Laippala, Alessandro Lenci, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Shunsuke Mori, Hanna Nurmi, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Prokopis Prokopidis, Sampo Pyysalo, Loganathan Ramasamy, Rudolf Rosa, Shadi Saleh, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Kiril Simov, Aaron Smith, Jan Štěpánek, Alane Suhr, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Sumire Uematsu, Larraitz Uria, Viktor Varga, Veronika Vincze, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. Universal dependencies 1.2, 2015. URL http://hdl.handle.net/11234/1-1548. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Emin Orhan, Vaibhav V Gupta, and Brenden M Lake. Self-supervised learning through the eyes of a child. page 12.

Ludovica Pannitto and Aurélie Herbelot. Recurrent babbling: evaluating the acquisition of grammar from limited input data. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 165–176, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.13. URL https://aclanthology.org/2020.conll-1.13.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume*, pages 2522–2532, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.215. URL https://aclanthology.org/2021.eacl-main.215.

Joe Pater. Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, 95(1):e41–e74, 2019. Publisher: Linguistic Society of America.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

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

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. doi: 10.18653/v1/N18-1202.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language Models as Knowledge Bases? 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 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://www.aclweb.org/anthology/D19-1250.

Jackson Petty and Robert Frank. Transformers Generalize Linearly. *arXiv:2109.12036 [cs]*, September 2021. URL http://arxiv.org/abs/2109.12036. arXiv: 2109.12036.

Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics, 2019. URL https://arxiv.org/abs/1808.09121.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-Theoretic Probing for Linguistic Structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.420. URL https://www.aclweb.org/anthology/2020.acl-main.420.

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

Geoffrey K. Pullum and Barbara C. Scholz. Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 18(1-2):9–50, 2002. Publisher: Walter de Gruyter.

Laura Pérez-Mayos, Miguel Ballesteros, and Leo Wanner. How much pretraining data do language models need to learn syntax? September 2021. URL https://arxiv.org/abs/2109.03160v2.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1 (8), 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and others. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, and others. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Altaf Rahman and Vincent Ng. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/D12-1071.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the*

2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, 2016. URL https://www.aclweb.org/anthology/D16-1264.

Malka Rappaport Hovav and Beth Levin. The English dative alternation: The case for verb sensitivity. *Journal of Linguistics*, 44(1):129–167, 2008.

Ezer Rasin and Athulya Aravind. The nature of the semantic stimulus: the acquisition of every as a case study. *Natural Language Semantics*, 29(2):339–375, June 2021. ISSN 0925-854X, 1572-865X. doi: 10.1007/s11050-020-09168-6. URL https://link.springer.com/10.1007/s11050-020-09168-6.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. Studying the Inductive Biases of RNNs with Synthetic Variations of Natural Languages. In *Proceedings of NAACL-HLT*, pages 3532–3542, 2019.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://www.aclweb.org/anthology/2020.acl-main.647.

Florencia Reali and Morten H Christiansen. Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, 29(6): 1007–1028, 2005. Publisher: Wiley Online Library.

Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B Cohen, and Simon Kirby. Compositional languages emerge in a neural iterated learning model. In *International conference on learning representations*, 2019.

Cinjon Resnick, Abhinav Gupta, Jakob Foerster, Andrew M Dai, and Kyunghyun Cho. Capacity, bandwidth, and compositionality in emergent language learning. In *Proceedings of the 19th international conference on autonomous agents and MultiAgent systems*, pages 1125–1133, 2020.

Eric Reuland. Grammar of binding in the languages of the world: Unity versus diversity. *Cognition*, 168:370–379, November 2017. ISSN 00100277. doi: 10.1016/j.cognition.2016.01.020. URL https://linkinghub.elsevier.com/retrieve/pii/S0010027716300208.

Mathieu Rita, Rahma Chaabouni, and Emmanuel Dupoux. "LazImpa": Lazy and Impatient neural agents learn to communicate efficiently. In *Proceedings of the 24th conference on computational natural language learning*, pages 335–343, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.conll-1.26.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, 2011. URL https://people.ict.usc.edu/~gordon/publications/AAAI-SPRING11A.PDF.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A Primer in BERTology: What we know about how BERT works. In *Findings of EMNLP*, 2020. URL https://www.aclweb.org/anthology/2020.tacl-1.54.

John Robert Ross. *Constraints on Variables in Syntax.* PhD Thesis, MIT, 1967.

Rachel Rudinger, Adam Teichert, Ryan Culkin, Sheng Zhang, and Benjamin Van Durme. Neural-Davidsonian semantic proto-role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 944–955, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1114. URL https://www.aclweb.org/anthology/D18-1114.

Ivan A. Sag. English relative clause constructions. *Journal of Linguistics*, 33(2): 431–483, 1997.

Ivan A. Sag, Gerald Gazdar, Thomas Wasow, and Steven Weisler. Coordination and how to distinguish categories. *Natural Language & Linguistic Theory*, 3(2):117–171, 1985. Publisher: Springer.

Ivan A. Sag, Thomas Wasow, and Emily M. Bender. *Syntactic Theory: A Formal Introduction*. CSLI Publications, 2 edition, 2003.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Wino-Grande: An Adversarial Winograd Schema Challenge at Scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(05), pages 8732–8740, 2020. doi: https://doi.org/10.1609/aaai.v34i05.6399.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked Language Model Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.240. URL https://www.aclweb.org/anthology/2020.acl-main.240.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Naomi Saphra and Adam Lopez. Understanding Learning Dynamics Of Language Models with SVCCA. 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 3257–3267, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1329. URL https://www.aclweb.org/anthology/N19-1329.

Karin Kipper Schuler. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. PhD Thesis, University of Pennsylvania, 2005. URL http://verbs.colorado.edu/~kipper/Papers/dissertation.pdf. ISBN: 0-542-20049-X.

Carson T. Schütze. *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press, 1996.

John R. Searle. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press, 1969.

Xing Shi, Inkit Padhi, and Kevin Knight. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, 2016.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/1089_Paper.pdf.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482*, 2021.

Melanie Soderstrom, Amanda Seidl, Deborah G Kemler Nelson, and Peter W Jusczyk. The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language*, 49(2):249–267, 2003. Publisher: Elsevier.

Robert Speer and Catherine Havasi. Representing General Relational Knowledge in ConceptNet 5. In *LREC*, pages 3679–3686, 2012. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/1072_Paper.pdf.

Dominique Sportiche, Hilda Koopman, and Edward Stabler. *An Introduction to Syntactic Analysis and Theory*. John Wiley & Sons, 2013.

Jon Sprouse and Diogo Almeida. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*, 48(3):609–652, 2012.

Jon Sprouse and Diogo Almeida. Setting the empirical record straight: Acceptability judgments appear to be reliable, robust, and replicable. *Behavioral and Brain Sciences*, 40, 2017. Publisher: Cambridge University Press.

Jon Sprouse, Carson T. Schütze, and Diogo Almeida. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua*, 134:219–248, 2013.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SygXPaEYvH.

Jessica Sullivan, Michelle Mei, Andrew Perfors, Erica Wojcik, and Michael C. Frank. SAYCam: A Large, Longitudinal Audiovisual Dataset Recorded From the Infant's Perspective. *Open Mind*, 5:20–29, May 2021. ISSN 2470-2986. doi: 10.1162/opmi_a_00039. URL https://direct.mit.edu/opmi/article/doi/10.1162/opmi_a_00039/97495/SAYCam-A-Large-Longitudinal-Audiovisual-Dataset.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975, 2020.

Hao Tan and Mohit Bansal. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. 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 5100–5111, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514.

Adam Teichert, Adam Poliak, Benjamin Van Durme, and Matthew Gormley. Semantic proto-role labeling. In *AAAI Conference on Artificial Intelligence*, 2017. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14997.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. In *Proceedings of ICLR*, 2019b.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. Small and Practical BERT Models for Sequence Labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3623–3627, 2019.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984. ISSN 0001-0782, 1557-7317. doi: 10.1145/1968.1972. URL https://dl.acm.org/doi/10.1145/1968.1972.

Annmarie Van Dooren, Anouk Dieuleveut, Ailís Cournane, and Valentine Hacquard. Figuring out root and epistemic uses for modals: The role of the input. *Journal of Semantics*.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5831–5837, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1592. URL https://www.aclweb.org/anthology/D19-1592.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, \Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in neural information processing systems*, volume 33, pages 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Elena Voita and Ivan Titov. Information-Theoretic Probing with Minimum Description Length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.14.pdf.

Joachim Wagner, Jennifer Foster, and Josef van Genabith. Judging grammaticality: Experiments in sentence classification. *CALICO Journal*, 26(3):474–490, 2009.

Alex Wang and Kyunghyun Cho. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the workshop on methods for optimizing and evaluating neural language generation*, pages 30–36, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2304. URL https://aclanthology.org/W19-2304.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018. URL https://www.aclweb.org/anthology/W18-5446.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *33rd Conference on Neural Information Processing Systems*, 2019a. URL https://proceedings.neurips.cc/

paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf. Journal Abbreviation: arXiv preprint arXiv:1905.00537.

Alex Wang, Ian F. Tenney, Yada Pruksachatkun, Katherin Yu, Jan Hula, Patrick Xia, Raghu Pappagari, Shuning Jin, R. Thomas McCoy, Roma Patel, Yinghui Huang, Jason Phang, Edouard Grave, Haokun Liu, Najoung Kim, Phu Mon Htut, Thibault F'evry, Berlin Chen, Nikita Nangia, Anhad Mohananey, Katharina Kann, Shikha Bordia, Nicolas Patry, David Benton, Ellie Pavlick, and Samuel R. Bowman. `jiant` 1.2: A software toolkit for research on general-purpose text understanding models. http://jiant.info/, 2019b.

Gregory Ward and Betty Birner. Definiteness and the English existential. *Language*, pages 722–742, 1995. Publisher: JSTOR.

Alex Warstadt and Samuel R. Bowman. Linguistic Analysis of Pretrained Sentence Encoders with Acceptability Judgments. *arXiv preprint arXiv:1901.03438*, 2019.

Alex Warstadt and Samuel R Bowman. Can neural networks acquire a structural bias from raw linguistic data? In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society.*, 2020.

Alex Warstadt and Samuel R Bowman. What artificial neural networks can teach us about human language acquisition. In Shalom Lappin and Jean-Philippe Bernardy, editors, *Algebraic Structures in Natural Language*. Taylor and Francis, to appear.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Corpus of Linguistic Acceptability, 2018. URL http://nyu-mll.github.io/cola.

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 Jeretič, and Samuel R. Bowman. Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs. In *Proceedings of EMNLP-IJCNLP*, pages 2870–2880, 2019a. URL https://www.aclweb.org/anthology/D19-1286.

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

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8, 2020a. doi: 10.1162/tacl_a_00321. URL https://doi.org/10.1162/tacl_a_00321.

Alex Warstadt, Yian Zhang, Haau-Sing Li, Haokun Liu, and Samuel R Bowman. Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, November 2020b. Association for Computational Linguistics.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 932–948, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational

Linguistics. doi: 10.18653/v1/2021.emnlp-main.72. URL https://aclanthology.org/2021.emnlp-main.72.

Ralph Weischedel, Martha Palmer, Marcus Mitchell, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. OntoNotes release 5.0 LDC2013T19., 2013. URL https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf. Published: Linguistic Data Consortium.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005, 2017.

William F Whitney, Min Jae Song, David Brandfonbrener, Jaan Altosaar, and Kyunghyun Cho. Evaluating representations by the complexity of learning low-loss predictors. *arXiv preprint arXiv:2009.07368*, 2020. URL https://arxiv.org/abs/2009.07368.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 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, 2018.

Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. 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*, pages 3302–3312, 2019.

Ethan Wilcox, Pranali Vani, and Roger Levy. A Targeted Assessment of Incremental Processing in Neural Language Models and Humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 939–952, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.76. URL https://aclanthology.org/2021.acl-long.76.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL 2018*, volume 1, pages 1112–1122, 2018.

Edwin Williams. Predication. *Linguistic Inquiry*, 11(1):203–238, 1980. Publisher: JSTOR.

Andrew G Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

Colin Wilson. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive science*, 30(5):945–982, 2006. Publisher: Wiley Online Library.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the*

16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2784–2790, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.242. URL https://aclanthology.org/2021.eacl-main.242.

Yuan Yang and Steven T. Piantadosi. One model for the learning of language. *Proceedings of the National Academy of Sciences*, 119(5):e2021865119, February 2022. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2021865119. URL https://pnas.org/doi/full/10.1073/pnas.2021865119.

Tian Yun, Chen Sun, and Ellie Pavlick. Does Vision-and-Language Pretraining Improve Lexical Grounding? In *Proceedings of EMNLP*, September 2021. arXiv: 2109.10246.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 93–104, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1009. URL https://aclanthology.org/D18-1009.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. When Do You Need Billions of Words of Pretraining Data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.90. URL https://aclanthology.org/2021.acl-long.90.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.