Contents

CHAPTER 1

What Artificial Neural Networks Can Tell Us About Human Language Acquisition

Alex Warstadt

New York University, Department of Linguistics

Samuel R. Bowman

New York University, Department of Linguistic, Department of Computer Science, Center for Data Science

CONTENTS

1.1 ABSTRACT

We discuss the possibilities and limitations of studying human language acquisition using artificial learners from both a theoretical and practical perspective. While this possibility has become increasingly discussed in recent years [51, 61, 102]—and in some cases doubted [3]—there has been no precise characterization of the scope of the possible evidence and the conditions that must be met. On the theoretical side, we put forward that a *deprivation experiment* in which a learner acquires language in an intentionally restricted environment can provide proof that some advantage hypothesized to be enjoyed by humans is, in fact, not necessary for language acquisition. This strong result is only obtained under the strict conditions that the learner's environment is no stronger than the human learning environment, and that the learner's innate inductive is no stronger than humans'. On the practical side, we argue these conditions are difficult to meet. However, there are many practical opportunities for strengthening the kinds of evidence that studies with artificial neural networks can give us about humans. Limiting the quantity and type of linguistic data available to learners, while enriching the non-linguistic input in their environments, can increase the chance of obtaining strong evidence about the necessary conditions for humans to acquire language.

1.2 INTRODUCTION

In the 13th century, the Holy Roman Emperor Frederick II conducted a troubling experiment. He arranged for children to be raised from infancy without any human language to determine which language children know from birth: Hebrew, Latin, Greek, or their mother's native tongue [17]? This experiment, like similar ones reportedly conducted by the Pharaoh Psamtik two millennia earlier and by Scotland's King James IV two centuries later [27], was deeply unethical and yielded no legitimate conclusions. But the possibility of language deprivation experiments has appealed to people throughout history because of the potential to better understand human language acquisition by manipulating variables during learning.

In the last decade, this possibility has begun to come within reach—without any of the ethical baggage—through the study of artificial neural networks. Since the "deep learning tsunami" in computational linguistics [69], we have gained access to artificial neural networks (ANNs) that largely learn to compose high quality multi-paragraph prose, to answer reading comprehension questions, and to make human-like acceptability judgments [65, 37, 5]. Crucially for the purpose of studying human language acquisition, as we will argue, we also have a high degree control over the internal learning mechanism and the learning environment of these systems.

In this time, researchers have begun to investigate the grammatical knowledge of generalized neural language models [63, 13, 32, 104, 98, 99, 100, 41, 6]. We will refer to these models simply as LMs, but they learn from several self-supervised training objectives such as next-word prediction (as in traditional language modeling) or the cloze task. While many results show that neural networks remain far from human-like language understanding, massive progress in that direction has been made through both technical innovations and increases in scale over the last few years [70, 62].

Many authors suggest that, to the extent that models succeed, this can help settle debates about humans' innate biases [51, 101, 13, 61, 78, 99]. However, most studies in this vein use artificial learners trained on convenient but un-human-like datasets like Wikipedia. As a result, these studies are not optimized to answer questions about *human* language learning, and so while their findings might be a useful stepping stone, their direct relevance to language acquisition is limited. At the same time, others have questioned the value of using neural networks to study human language acquisition at all, arguing that their inductive biases are too strong for this to be successful [3].[1]

The goal of this paper is to characterize what we can (and cannot) hope to learn about human language acquisition from studying artificial learners, and how best to maximize the relevance of studies on ANNs to questions of human learning. We agree with many others who contend that artificial neural networks are especially well suited to determine which hypothesized advantages (i.e., innate biases or types of stimuli in the environment) are *unnecessary* for human language learning [51, 101, 13, 61, 78, 99]. We make this claim more precise by showing how the relevant conclusions follow deductively from a specific kind of experimental result under very strict conditions. The way to accomplish this is through a *deprivation experiment* in which a learner is deprived during language acquisition of some advantage and then tested for some target knowledge. If the learner passes the test, this results in an existence proof that a hypothesized innate bias or some kind of linguistic input is not needed to successfully acquire the target. This becomes a proof about what is *human-learnable* as long as the model learner does not have any additional advantage over human learners. These experiments could prove to be a valuable tool for testing many long-standing hypotheses regarding what innate language-specific biases are needed to explain human language learning [9, 12], as well as for evaluating claims that the input to the learner lacks key evidence for acquiring certain forms of linguistic knowledge [10, 56, 59, 4, 87].

The structure of the paper is as follows: Section 2 articulates what an ideal deprivation experiment in this style would look like, and what such an experiment can and cannot tell us about humans. The next three sections each focus on ingredients for designing the ideal experiment, and discuss the ideal setup, the state of current experiments with model learners, and how to deal with obstacles in achieving this ideal: Section 3 discusses tests for linguistic knowledge, Section 4 discusses the con-

[1] On a related note, Dupre has argued that research on ANNs cannot contribute to a theory of linguistic competence [23]. However, they note that their argument is "consistent with recent work...that has argued that DL may provide insight into the mechanisms by which linguistic competence is acquired."

siderations about the learning environment, and Section 5 discusses considerations about the learner.

1.3 THE IDEAL EXPERIMENT

In the best case scenario, a deprivation experiment with an artificial learner can give an existence proof that some linguistic knowledge is human-learnable without some hypothesized advantage such as an innate bias or a kind of stimulus. An example of such a proof is given below:

1. *Let* there exist a test T such that any learner L' can pass T if and only if L' has knowledge of a target generalization K.
2. *Let* there exist a learning environment E such that E is no richer than the learning environment of a typical human.
3. *Let* there exist a learner L such that L has no stronger inductive bias than a typical human.
4. *Let* there be some environmental (or innate) advantage A that is not initially present in E (or L) and is hypothesized to be necessary for L to acquire K.
5. *Assume* that, if the richness of environment E_1 is greater than that of E_2, then everything that is learnable in E_2 is learnable in E_1.
6. *Assume* that, if learner L_1 has stronger inductive bias than L_2, then every generalization K' within the hypothesis space for both L_1 and L_2 that is learnable for L_2 is learnable for L_1.
7. *Hypothesis*: L can pass T after exposure to E.
8. *Conclusion*: K is learnable for a typical human in a typical learning environment without A.

We will step through the logic of this proof in detail, but to make things more concrete, we do so using Chomsky's classic example of the acquisition of the subject-auxiliary inversion rule in English [10], which we review briefly here: Through analogy of strings like (1-a) with declaratives, a learner without a hierarchical bias could discover empirically that questions in English are formed by moving an auxiliary to the front; but without examples like (1-b), such the learner could not determine that the correct rule targets the structurally highest auxiliary, rather than the linearly first auxiliary. In this example, the target knowledge is the hierarchical rule for subject-auxiliary inversion. It also includes two advantages hypothesized to play a role in the acquisition of this rule that we can deprive an artificial learner of. One is an innate bias towards hierarchical rules, and the other is the disambiguating evidence from examples like (1-b). The argument from a deprivation experiment states that, if a model learner deprived of both advantages (as well as any other advantages over humans) can learn the subject-auxiliary inversion rule, it follows that humans do not need these advantages, either.

(1) a. Is the man __ happy?
b. Is the man who is tall __ happy?

The argument above begins in lines (1)-(3) with supposing the existence of three things that each meet their own strict conditions: a test, a learning environment, and a learner. In the case of subject-auxiliary inversion, the test is some kind of task that, if completed correctly by a learner, gives proof that the learner has acquired knowledge of the hierarchical rule for subject-auxiliary inversion in English. If the test does not have this property, we cannot draw conclusions with any certainty.

The learning environment is some set of stimuli, such as texts of English, that is not richer than a typical human's environment. If the learner's environment is richer *in any respect* than a human's, human-learnability no longer follows from learner success, since it is possible that the model may have only succeeded by virtue of advantages in its environment not available to humans. There are many ways in which the environment could be richer, such as containing far more data than a human learner is exposed to, or containing numerous examples of subject-auxiliary inversion in complex sentences.

The learner is some data-driven learning algorithm, such as an ANN. A similar logic applies to the learner's bias as to the environment. If the learner possesses some inductive bias not innate to humans—for example if the learner has innately programmed into it the ability to parse English sentences—then even if it passes an adequate test in an impoverished environment, the possibility remains that it would have failed if restricted to human-like inductive bias.

Line (4) of the argument supposes the existence of some particular advantage that we know to be omitted from the artificial learning setup. This advantage could be environmental—for example, the presence of examples of subject-auxiliary inversion in sentences with embedded clauses like (1-b). It could also be an innate bias—for example, a bias towards acquiring hierarchical rules.

Lines (5) and (6) introduce two monotonicity assumptions about learnability. Namely, they suppose that learnability is non-decreasing with respect to richness of the environment and learner inductive bias. For the environment, this assumption implies that, if subject-auxiliary inversion is learnable in some environment E, then enriching E with many examples like (1-b) would not cause the rule to become unlearnable.[2] When it comes to learner inductive bias, the logic is similar, but we must add the condition that the target generalization remains in the learner's hypothesis space for learnability to be preserved, since one way to strengthen inductive bias is to remove hypotheses from the learner's hypothesis space.

Line (7) simply states that the experiment has yielded a positive result. For example, the learner has passed the test for subject-auxiliary inversion without having access to helpful examples like (1-b) or a hierarchical bias. If this result is obtained, the conclusion in (8) follows. By lines (2) and (3), typical humans enjoy at least as many environmental and innate advantages as the learner. By lines (5) and (6),

[2]This assumption becomes subtle, but still tenable, when one considers the idea that learning might be easier if the learner is exposed to the simple examples first [24]. Arguably, building this kind of curriculum into the environment is a form of richness itself.

it would not harm learnability of the subject-auxiliary inversion rule to enrich the learner's environment and inductive bias in such a way that they are identical to a human's but also lack examples like (1-b) and a hierarchical bias.

To summarize, this argument is useful mainly for the purpose of falsifying a hypothesis that some advantage, either innate to the learner or in the environment, is necessary for humans acquiring the target knowledge. There are some differences between these two cases. In the case of testing the necessity of innate knowledge, it is generally a matter of contention what innate biases humans have. An existence proof that some innate bias would be superfluous is evidence against humans having that bias. In the case of testing the necessity of some stimulus, there are easy ways to test what is in the input to a typical learner. Numerous corpus studies exist demonstrating the presence or absence of some kind of stimulus that is hypothesized to be necessary [60, 87, 95], including some specifically targeting subject-auxiliary inversion examples like (1-b) [82, 56]. However, merely demonstrating that some kind of example is present or not does not settle whether the related knowledge is human-learnable from those examples. A deprivation experiment with an artificial learner *can* settle this issue.

What a deprivation experiment cannot do is settle whether humans actually make use of a particular form of innate knowledge or stimulus. Even if the target knowledge does not *require* some hypothesized advantage, that advantage might still play a role in human language acquisition. There are several reasons this could be the case. First, there might be some other form of linguistic knowledge that *does* require the advantage in question. Second, the advantage might make acquiring the target knowledge more efficient, conferring an evolutionary advantage to a human that makes use of it. Still, the simplest theory of human language acquisition is one that does not posit superfluous advantages. If we find that subject auxiliary inversion is learnable without a hierarchical bias, then we should assume humans do not have a hierarchical bias absent additional evidence to the contrary.

Another limitation of a deprivation experiment is that it only gives an existence proof when the learner succeeds. If the learner fails, this does not imply that the hypothesized advantage is necessary. Instead, failure could be due to selecting too low of a lower bound for model inductive bias or the environment. For this reason, the best strategy for achieving an existence proof is to aim for a *tight* lower bound on human advantage for both these factors. The stronger the model's bias or the richer the environment (while not exceeding those of humans), the more likely it is to succeed.

1.3.1 A Less Ideal Experiment

In reality, the conditions outlined above are quite extreme, and a deductive proof of human-learnability may be extremely difficult. In the less ideal setting where the environment or learner have some advantages not enjoyed by humans, there is still evidence to be gained from deprivation experiments, but the strategies are somewhat different. One strategy is to limit how much the learner's environment and biases exceed humans', as the evidence becomes stronger to the extent that it is possible.

Another strategy is to undershoot other aspects of human bias and the richness of the stimulus as severely as possible. For example, if the learning environment includes examples like (1-b), but has been augmented with similar but ungrammatical examples like **Is the man who tall is happy?*, this could strengthen evidence about learnability, assuming the learner is successful.

1.3.2 Examples of Targets for Existence Proofs

The literature on language acquisition contains many claims about what kinds of innate knowledge are necessary, and in what ways the input to human learners is too impoverished. These are the kinds of things that can be tested.

First, there are claims about the kinds of innate advantages humans have. [9, pp. 55-56] argues that humans have a restricted hypothesis space that excludes generalizations based on linear order, counting, and other surface features. This view has been expanded by many others (e.g., [107]). Additionally, there are claims about what kinds of universals and degrees of freedom are present in language acquisition, many of which were articulated in the Principles and Parameters framework [12].

Second, there are numerous examples of specific phenomena where the input to learners is thought to be too impoverished to learn the correct generalization without some innate bias. [10] discussed subject auxiliary inversion. [82] compile a list of additional cases where key data is argued to be absent from the stimulus: plurals in noun-noun compounds [31], auxiliary sequence ordering [44], and anaphoric one [2]. [87] argue that environmental data also does not sufficiently constrain the meaning of the quantifier *every*. Many other impoverishment claims can be found in the language acquisition literature, and deprivation experiments would subject them to a more stringent empirical test.

1.4 TESTS OF HUMAN-LINGUISTIC KNOWLEDGE

The formal argument in Section 1.3 requires the existence of a test. In this section, we discuss in theory what it means to test an artificial learner for human-like linguistic competence, before surveying existing resources that can be used as tests of linguistic performance.

1.4.1 Tests in Theory

The learnability proof we have outlined outlined concludes with proving the learnability of some form of linguistic *knowledge*. In theory, a deprivation experiment could target human-like linguistic competence *if* a fool-proof test for competence exists.[3] However, in practice, most tests are behavioral, and therefore really target linguistic performance. In addition to being easily observable, performance is more theory-neutral than competence. Competence is a theoretical construct even for humans, so a test of competence would always be subject to our degree of belief in the theory. Additionally, there is a good deal about knowledge that can be inferred from per-

[3]Though see [23] for arguments that the notion of linguistic competence cannot apply to ANNs.

Phenomenon	N	Acceptable Example	Unacceptable Example
Anaphor agr.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
Arg. structure	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
Binding	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
Control/raising	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
Det.-noun agr.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
Ellipsis	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
Filler-gap	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
Irregular forms	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
Island effects	8	*Which bikes is John fixing?*	*Which is John fixing bikes?*
NPI licensing	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
Quantifiers	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
Subj.-verb agr.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

TABLE 1.1 Minimal pairs from each of the twelve categories covered by BLiMP. Differences between sentences are underlined. N is the number of minimal pair types within each broad category. (Table from Warstadt et al. [100] reprinted with permission.)

formance alone. Although this has its limitations—two systems that have identical behavior in some respects could have very different internal functioning—we can construe performance very broadly to include many aspects of behavior, ranging from acceptability judgments to order of acquisition and processing difficulty.

1.4.2 Existing Tests

There are now numerous well-motivated, controlled, and challenging tests for different aspects of neural networks' grammatical knowledge.

These tests fall roughly into two main categories: supervised and unsupervised. Unsupervised tests do not rely on task-specific training beyond a self-supervised training objective such as language modeling. Thus any linguistic knowledge revealed by these methods can only have been acquired through self-supervised exposure to the learning environment, or to innate abilities of the learner. While supervised or weakly supervised methods do provide models with task-specific instruction, supervised tasks can be constructed to answer complementary questions to unsupervised tests, much like artificial language learning experiments on humans [33].

1.4.2.1 Unsupervised Tests

Language model scoring over minimal sets of sentences, sometimes referred to as targeted syntactic evaluation, is one of the mostly widely adopted methods. Introduced by [63], this method relies on the assumption that language models should assign higher probability to a grammatical sentence[4] than to a minimally different ungrammatical one. This is a necessary—though not sufficient—condition of encoding whatever grammatical concepts are responsible for the reported contrast in humans.

BLiMP (The Benchmark of Linguistic Minimal Pairs) [100] is the largest-scale re-

[4] ...or substring of a sentence, conditioned on the prefix.

source for language model scoring. It tests 67 minimal pair types in English, each consisting of 1k pairs, organized into 12 broad categories. These categories cover morphology (e.g. subject verb agreement and determiner-noun agreement), syntax (e.g. argument structure, island effects and binding), and semantics phenomena (e.g. quantification and negative polarity items). Table 1.1 shows examples from BLiMP of one minimal pair type for each of these categories. Closely related is SyntaxGym [28, 41], which adopts a version of the LM scoring paradigm in which the model's predictions must conform to more than one hypothesized inequality over a set of sentences, rather than just a minimal pair.

The tests above focus on offline acceptability judgments, but other measures of performance are possible. For example, [105] test LMs' predictions against humans' online processing difficulty using SyntaxGym. Under the theoretically motivated assumption that there should be a log-linear relationship between a word's online processing time in humans and a LM's predicted probability for the word in context [34, 58], it is possible to test the conditions under which human-like processing can be acquired.

1.4.2.2 Supervised Tests

Another kind of evaluation uses constrained supervision to probe how neural networks generalize. In this approach, what is under investigation is not knowledge of a particular phenomenon in the training data, but whether models extend knowledge to unseen cases in ways that we expect humans to.

This approach can tell us the extent to which models show rule-governed behavior. For example, COGS (Compositional Generalization Challenge based on Semantic Interpretation) [43] is a semantic parsing dataset in which certain semantic configurations in the test data are systematically held-out from the training data. If a model is able to learn that semantics, syntax, and surface form are related by a set of general compositional and phrase-structure rules, then it should correctly parse a noun in any syntactic position, even if it has only seen that noun in object position during training.

This approach is also useful for probing the inductive biases of neural networks. The Poverty of the Stimulus experimental design [106] provides a paradigm for doing so. Figure 1.1 gives an example from [103] of an experiment following this design. A learner is trained to perform a task given data that is ambiguous between (at least) two hypotheses, and tested on data where the hypotheses make divergent predictions. For example, numerous studies have used this design to test whether ANNs prefer a generalization based syntactic structure or one based on linear order for subject auxiliary inversion [26, 71, 72, 99].

One large-scale dataset making use of this design is MSGS (The Mixed Signals Generalization Set) [103], which tests whether a learner has a bias towards linguistic or surface generalizations. MSGS consists of 20 ambiguous tasks, each pairing one of four linguistic generalizations (e.g. *labels indicate whether the main verb of the sentence is in the progressive form*) with one of five surface generalizations (e.g.

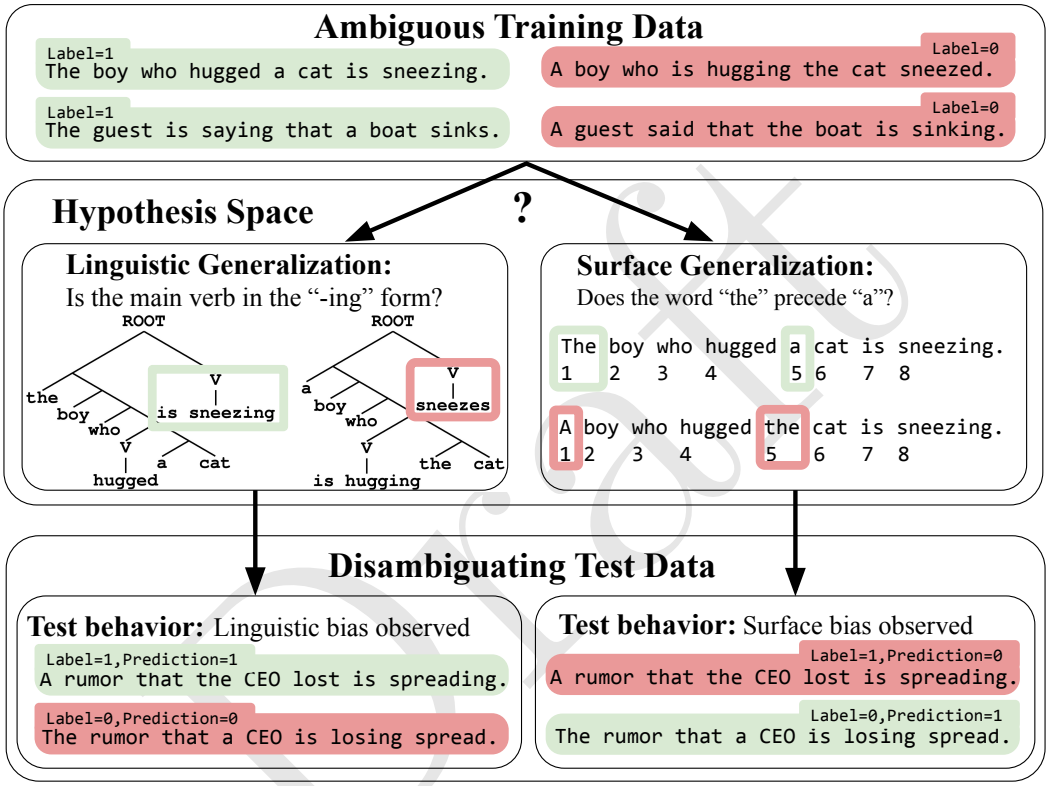

Figure 1.1 Example of an experiment following the Poverty of the Stimulus design (reprinted from [103] with permission). A model is trained on ambiguous data whose labels are consistent with either a linguistic or a surface generalization, and tested on disambiguating data whose labels support only the linguistic generalization. Light green and dark red shading represents data or features associated with the positive and negative labels/predictions, respectively.

labels indicate whether the sentence is longer than 10 words). [66] also construct a similar dataset.

1.4.2.3 What Do Out-of-Domain Tests Tell Us About Learnability?

The goal of a learnability existence proof is to draw some conclusion about an *unstructured* learning environment. Perhaps surprisingly, training on a supervised task with labeled data can still tell us something of use here. The key is to use already trained language models to provide the initial weights for the main component of a task-specific models. This is, of course, a straightforward application of the popular pretrain and fine-tune paradigm in NLP [19, 39, 85, 21].

Following this setup, the experiment can tell us whether an inductive bias, such as a hierarchical bias or a compositionality bias, can be *acquired* through exposure to the unstructured learning environment [99, 103]. An acquired inductive bias, though not present innately in the learner, can still influence how the learner forms generalizations about sub-problems encountered during the learning process. One caveat is that it is difficult to tell whether the target learning mechanism is actually implicated in learning the relevant phenomenon.

1.5 THE LEARNING ENVIRONMENT

To achieve an existence proof of human-learnability, it is necessary to create a learning environment for the artificial learner that represents a lower bound on the richness of the input to human learners. The designer of a deprivation experiment must take care that the learner's environment does not exceed the quantity or quality of data available to humans. Of course, if learning succeeds in an environment that is *far poorer* than a human's—for instance, one containing only a few thousand words—this would make the point strongly. One can go too far in this direction, though, for if the learner fails to acquire the target there is no existence proof. This is why a *tight* lower bound gives the greatest chance of an existence proof. Furthermore, although ANNs are exposed to large amounts of text data, their learning environment is still impoverished compared to humans in terms of non-linguistic input, such as visual stimuli, and inter-agent interaction. Thus there is ample room to enrich the learning environment of ANNs in certain directions without exceeding the conditions of a strong learnability proof.

There is a lot we do not yet know about how altering the learning environments of LMs to better resemble humans' will affect grammar learning. Much of the discussion of the human learning environment has emphasized how *impoverished* it is [9]. On the other hand, a realistic environment may turn out to be richer than it seems. ANNs have proven in the last few years that there is a surprising amount of signal in plain text data. The problem with most of these results is that these models are trained in an environment that is very unlike the human learning environment, and neither strictly richer nor strictly more impoverished. In this section, we will step through the ways in which the learning environments of ANNs are unlike the human learning

environment, and discuss a growing body of work which has sought to address these points of divergence.

1.5.1 Data Quantity

Most popular ANNs for NLP have been trained on far more words than a human learner. While this was not the case only a few years ago, this trend has only been increasing. Thus, researchers interested in questions about human language acquisition have already begun to evaluate models trained on more human-scale datasets [96, 100, 109, 41, 77, 83].

However, it is not trivial to determine how many words a typical human learner is exposed to. The best-known figures come from Hart & Risley's study on American English-speaking children's linguistic exposure in the home [35]. They find that children are exposed to anywhere from 11M words per year to as little as 3M words, depending on a number of factors such as family income. These figures include all speech in the home environment, not just child-directed speech. Choosing the beginning of puberty as a somewhat arbitrary cutoff point for language acquisition, and assuming that Hart & Risley's numbers extrapolate linearly, a child will acquire language with anywhere from tens of millions of words to over a hundred million words.

By comparison, popular neural language models are trained on corpora consisting of far more: BERT [21] is trained on about 3.3 billion words, RoBERTa [65] is trained on about 30 billion, and GPT-3 [5] is trained on about 300 billion. Thus, the most impoverished of these models has linguistic experience equivalent to about 300 human years, and the most enriched is at 30,000 human years.

We can already begin to draw some conclusions about how linguistic performance of LMs scales with the quantity of available data. One study by Zhang, Warstadt, Li, and Bowman [109] uses BLiMP to evaluate models trained in the style of RoBERTa [65] on datasets ranging from 1M words to 1B words. Figure 1.2 summarizes their results, showing the growth in sensitivity to acceptability contrasts as a function of the amount of training data available to an LM.

They find that language models do learn many human-like generalizations given abundant data when tested using unsupervised LM scoring. $\text{RoBERTa}_{\text{BASE}}$ which is trained on about 30B words [65] achieves near-human performance (which we define as accuracy within 2% points of humans or better) on 6 out 12 BLiMP categories. Among these categories are phenomena involving long distance syntactic dependencies such as filler-gap dependencies and island effects, which have been previously found be challenging [100].

On the other hand, language models generally fail to reach human-level accuracy when restricted to human-scale data quantities. According to the same study, RoBERTa models trained at human-scale on 100M words only achieve near-human performance in at most 2 BLiMP categories. Models trained on 10M words are unsurprisingly even worse, reaching near-human performance in at most 1 BLiMP category. This finding is corroborated by several other studies that report results from models trained at similar data scales [96, 100].

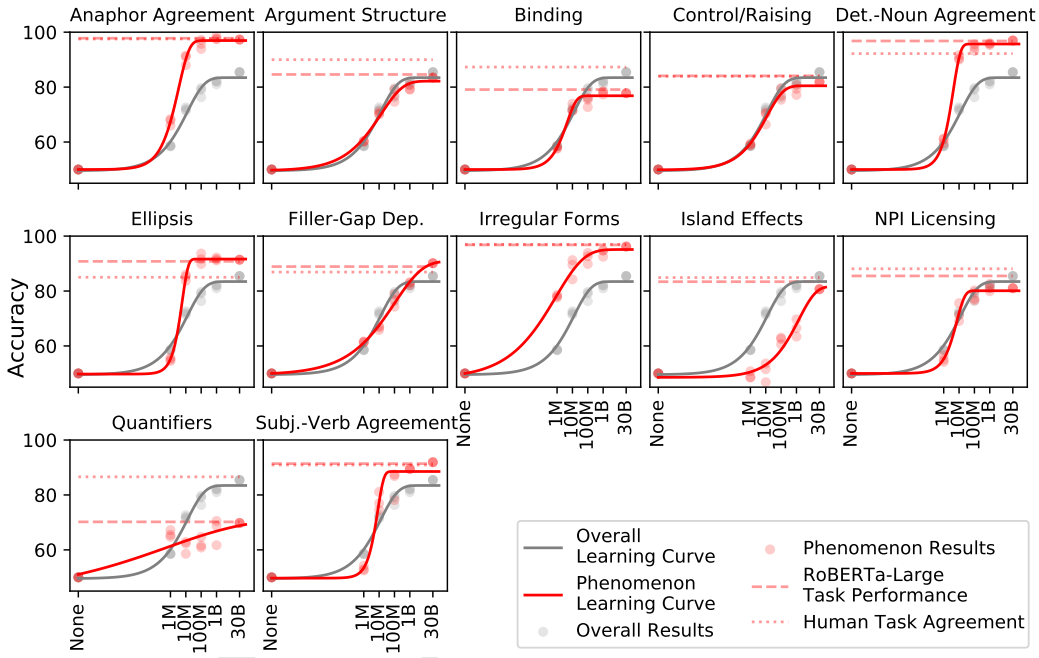

Figure 1.2 Learning curves from Zhang, Warstadt, Li, and Bowman [109] (printed here with permission) showing LM improvement in BLiMP performance as a function of the number of words of training data available to the model. RoBERTa-Large performance was originally reported from Salazar et al. [90], and human task agreement was originally reported in the release of BLiMP [100].

Zhang et al. also find that most improvement in language models on BLiMP occurs as training corpus volume increases from 1M to 10M words. This can be seen in both the repeated gray curve in Figure 1.2 representing overall performance, as well is in many phenomenon-specific learning curves. Substantial improvements are still possible between 10M and 100M words, but after this point, performance starts to plateau. Similar findings have been reported for LSTMs as well [96, 100].

1.5.2 Data Source

Another point of divergence between human and model learning environments is the source of language data. One of the main distributional differences is that all the linguistic input to a pre-literate child is spoken or signed. Ideally, the model learner's environment should consist of unstructured *audio* of spoken language (or video of signing). While there have been some first steps towards training LMs on such data [75, 50], these techniques are still not developed enough to draw conclusions from.[5]

As long as text-based training remains the only viable option for training LMs, the most ecologically valid text domain is transcribed speech. One source of such data is CHILDES, a database of transcribed parent-child discourse [68]. Indeed, such infant-directed speech is a major source of input to many child learners, and some go so far as to train model learners exclusively on child-directed speech [89, 79]. This is probably overkill: Child-directed speech makes up only a small fraction of the linguistic input to child learners, and in some communities it is vanishingly rare [18].

Another large-scale source of transcribed speech is COCA [20], which includes 83M words of transcribed speech from unscripted radio and TV programs. One step down in terms of ecological validity is OpenSubtitles [64], which contains over 2B words of English subtitles from scripted and unscripted television and radio, as well as over 100M words of subtitles in numerous other languages. While these datasets are ultimately not what is needed to obtain strong learnability proofs, they can give stronger evidence than training datasets currently used to train popular language models such as Wikipedia, news, and web data.

1.5.3 Prosody

There is substantial linguistic information in speech not present in text, especially prosody. Prosodic bootstrapping is thought to play a major role in syntactic acquisition [30, 92], so LMs are at a distinct disadvantage in this respect. On the other hand, text data has punctuation and white space, and is tokenized prior to input into an LM, which provides an *advantage* to models when it comes to detecting word, phrase, and sentence boundaries. Again, if practical limitations are not an issue, it is best to study models trained mostly on audio. But since this is not totally practical at the moment, there is still a lot to learn from LMs trained on text. Text exceeds the

[5] For example, current state-of-the-art performance of audio-trained LMs on a modified audio version of BLiMP is only 58% accuracy—just 8% points above chance—compared to over 79% accuracy for RoBERTa models trained on 100M words.

richness of speech in fairly limited ways, meaning that results from text-trained LMs still give suggestive evidence about humans.

1.5.4 Non-linguistic Input

Despite some advantages in the linguistic environments of typically studied ANNs, they have severe non-linguistic disadvantages compared to humans. Whereas most ANNs studied in this literature learn in a text-only environment with a simple LM training objective, humans learn in a multifaceted environment with many forms of sensory input, other agents to interact with, and complex risks and rewards. The effects of these differences in non-linguistic input on grammar learning are likely to be more indirect than changes in linguistic input. Still, they may turn out to be substantial, especially when it comes to the quantity of linguistic input the learner requires.

1.5.4.1 Multimodal Input

Theories of language acquisition have long hypothesized a substantial role for sensorimotor input. Concepts learned through multimodal inputs in pre-linguistic infants and later on in learning can enable word learning and potentially even grammar learning [29, 40].

Standard LMs like BERT must achieve all learning of this kind from text alone. Indeed, these models can acquire some semblance of world knowledge. They have been shown to be somewhat effective as knowledge bases for retrieval of encyclopedic knowledge [80], and they now achieve strong performance on Winograd schema challenge sets [86]. On the other hand, there is still reason for skepticism regarding how much world knowledge is needed to succeed on these evaluations [46]. Also, the quantity of training data needed to achieve strong performance on even these limited benchmarks is on the order of one billion of words [109]. Thus, whatever limited world knowledge language models can acquire is not likely to be useful for language acquisition at the same early stages of learning as in humans.

On the other hand, there is a growing inventory of ANNs trained jointly on vision and language data [54, 67, 94, 8]. However, evidence from linguistic evaluation of several models suggests that the visual input they receive does little to help with grammar learning [108], though more recent contrastively trained models may be different in this respect [84]. Most of these models learn from extending self-supervised objectives like the cloze task to the vision domain. Currently, one limitation of the models is that their linguistic input is even farther from that of human learners than language-only models. Most are trained entirely on image caption datasets such as MS COCO or Visual Genome [7, 47], which lack extended discourses and dialogues and contain a non-representative sample of sentence types. Furthermore, visual input to humans is continuous and moving, and thus richer than still images. Video and language models do not achieve a more realistic training environment. For example, VideoBERT is trained on YouTube cooking videos and text from automatic speech recognition [93].

1.5.4.2 Interactive Learning

Another ingredient missing from the input to most available model learners is interaction with an environment containing other conversational agents. While the objective of LMs is to reproduce the distribution of words and phrases in the language as faithfully as possible, human learners have a much more complex and varied objective function. We use language to share information, to make queries, and to issue and comprehend directives [1, 91]. The incentive for acquiring grammar in humans is that it makes these kinds of interactions possible, and these interactions help us achieve our non-linguistic goals.

The artificial learning paradigm that comes closest to reproducing this aspect of the human learning environment is multiagent reinforcement learning [53]. In this framework, multiple artificial agents are given a cooperative goal, such as to solve a reference game, which requires developing a mode of communication. However, these emergent modes of communication generally do not resemble human language [52], and efforts to better align them to existing natural languages (for instance by initializing agents with language models trained on English) have had mixed results [55].

1.6 THE LEARNER

In this section we consider the final condition on a strong proof of learnability: a learner that does not exceed the inductive bias of humans.

1.6.1 What Is a Lower Bound on Inductive Bias?

Achieving a tight lower bound on human innate inductive bias is constrained by a number of factors. First, doing so requires knowing enough about humans' innate biases—including domain general biases—to avoid exceeding them. Second, defining how one bias can be "weaker" than another is not just a practical issue, but a challenging theoretical one which will take considerable work to formalize, making it difficult to directly compare the relative strengths of human and model biases. Third, we are limited by the available set of learners, and developing effective new artificial learners is a large and mature field of research in its own right. Acknowledging these practical limitations, what would an ideal model learner look like?

One possible misconception is a model learner must be an unbiased *tabula rasa* in order to prove some innate bias unnecessary for language acquisition. First, this would be an impossible standard to meet, since all learners have some inductive bias. A inductive bias is just a prior over the hypothesis space, and thus a necessary property of any learner that can converge [73]. Second, we know of no claims that humans are totally unbiased learners. Many do argue that *language-specific* biases are not necessary to explain language acquisition [45, 89, 16, 14]. They suggest that we may instead have innate *domain general* biases which aid us in language acquisition. For an existence proof of this claim, the model learner only needs to lack language specific bias, and can possess domain-general bias as long it is no stronger than a human's.

Some points here require a bit of refinement. First, what does it mean for one inductive bias to stronger than another? Defining a strength ordering over inductive biases in general is not a trivial matter, and we are only aware of some definitions that apply to special classes of learners [76]. As a rigorous solution to this problem is beyond the scope of this work, we can suggest a heuristic approach to compare inductive biases. An inductive bias B is a learner's prior probability distribution over the hypothesis space [36]. Intuitively, B is more biased than another inductive bias A if and only if B concentrates more probability mass on a smaller subset of the hypothesis space. In other words, given a list of hypotheses ordered by probability for each bias, the probability mass of all the hypotheses up to rank n will always be greater for B than for A.

Second, what does it mean for a bias to be language-specific? An example of the subtlety of this issue is hierarchical bias. Chomsky famously argues that humans have a bias towards forming a generalizations based on syntactic structures about grammatical operations like subject-auxiliary inversion, when linear generalizations would adequately describe most of the data [9]. However, it is possible to question how language-specific even this bias is, since non-linguistic aspects of human cognition such as music also make use of hierarchical structures [57]. More recently, Chomsky has claimed that the primary innate endowment that enables language learning is unbounded Merge, or the ability to form recursive concepts [11]. Merge in this view emerged prior to language as we know it: It would have evolved mainly to facilitate abstract thought, with language later co-opting this operation. While Chomsky suggests that Merge in this incarnation is implicated in the language of thought, whether it can be claimed to be truly language-specific seems to be a matter more of terminology than of actual substantive debate. One possible conclusion is that whether a learner's inductive bias is language-specific is a matter of degree, but we leave a formal exploration of this to future work.

1.6.2 Achieving a Lower Bound on Human Inductive Bias in Practice

Practically, our ability to choose appropriate model learners is constrained by the available models. In recent years, our understanding of these models' inductive biases has grown substantially thanks to much empirical work.

1.6.2.1 Available Models

Most research in contemporary natural language processing makes use of a small number of neural architectures. Recurrent neural networks (RNNs) [24], and more specifically LSTMs [38] and GRUs [15], grew massively in popularity around 2014 and 2015. Transformers [97] became dominant in NLP applications starting in 2018 with the advent of BERT [21].

RNNs and Transformers are both generally deep, in the sense that they consist of multiple layers, where the outputs of one layer serve as the input to the next. The most notable difference between RNNs and Transformers is in how they accept inputs. An RNN processes the input sequentially, updating a state representation incrementally after reading each word. Gated RNNs like LSTMs and GRUs follow

this same principle, while designing additional machinery to manage this update process. Transformers, which have proven far more effective at large scales, process each token of a string entirely in parallel using an attention mechanism, operating iteratively over a fixed set of layers, rather than over the length of the sequence. At each layer, the representation for each input word depends on a weighted combination of all representations from the previous layer.

1.6.3 The Inductive biases of Neural Network Architectures

Do any of these models represent a lower bound on humans' innate inductive bias? Strictly speaking, the answer is almost certainly "no". It would be extremely surprising if their inductive biases were weaker in all respects, as these models have become widely used due to their empirical success at NLP applications rather than a strict adherence to not exceeding humans' inductive bias.

So what are the inductive biases of these models, and are they well-suited to carrying out informative deprivation experiments? A growing body of work helps to address these question by evaluating neural networks for a variety of human-like inductive biases.

Numerous studies have found that ANNs lack a variety of human-like inductive biases prior self-supervised training. One striking example is that humans, but not ANNs, show a strong *compositionality bias*. A key property of language is that words and phrases in language make stable compositional contributions to the semantics of larger constituents [74, 25]. One consequence of this is that humans can understand the compositional semantic contribution of a newly learned word in any appropriate context [49]. However, ANNs have shown a general inability to make compositional generalizations like this [48, 43, 42].

ANNs also generally lack a bias towards adopting hierarchical generalizations. McCoy, Frank, & Linzen test several varieties of RNNs without any pretraining using the Poverty of the Stimulus method on an ambiguous subject auxiliary inversion task, and find that none converge on a systematic hierarchical generalization [72]. Subsequently, Petty & Frank have shown a similar result for Transformers [81].

The fact that ANNs lack these human-like biases might make them more appropriate model learners for deprivation experiments for two reasons. First, it means they do not have any special innate advantage over humans in these respects. Second, if the goal of the study is to establish, for example, whether an innate structural bias is necessary for learning some target, then an off-the-shelf ANN is already a relatively appropriate test subject without any special modification to remove the bias in question. However, much stronger evidence about their inductive biases is needed for strong existence proofs.

One practical question is whether there is an advantage to using RNNs or Transformers as model learners. RNNs have a strong locality bias [22, 88] which Transformers lack. This is a consequence of the models' architectures. RNNs have the notion of linear order built in, since they get information about the rest of the sequence only from the previous token's output. Transformers on the other hand only receive information about linear order through a set of dedicated *positional embeddings* added to

the input. As a result, Transformers must learn the semantics of positional embeddings including notions like locality from scratch.

On the other hand, the differences between the biases of LSTMs and Transformers may be weaker than one might expect when it comes to grammar learning. For example, [100] compute the correlation between the accuracy scores of pairs of LMs on BLiMP. Among a population of models including an n-gram model, an LSTM, and two Transformers, they found that most strongly correlated models were the LSTM and one of the Transformers.

1.7 CONCLUSION

Thanks to recent advances in machine learning, we are closer than ever before to being able to construct an ethically-sound experiment that can allow us to more precisely limit the necessary conditions for the human-learnability of grammar. We have argued that a deprivation experiment with an ideal design can give this degree of certainty. From a practical standpoint, however, we are far from being able to construct the ideal experiment, especially when it comes to having certainty that the inductive biases of our model learners are appropriately weak. While this means work on artificial learners in the near future is unlikely to yield any incontrovertible proof about human learnability, we do not find this to be cause for despair. A deprivation experiment that does not meet the stringent conditions of an ideal experiment can still contribute converging *evidence* about human learnability, even if it cannot provide proof. And the evidence becomes *stronger* as we come closer to constructing ideal learning environments and learners for this design.

At present, there are many actionable opportunities for constructing experiments that come closer to this ideal. The most obvious is restricting the quantity and kind of text data used to train model learners. That said, there are also many remaining ways to safely enrich the environments of model learners through the introduction of multimodal input and interaction with other agents.

On the other hand, the problem of constructing effective learners with the right inductive bias remains difficult. One serious limitation is our ability to compare inductive bias of humans and models. Until we can reliably quantify these properties of a learner, we cannot have the necessary degree of confidence in our choice of a model learner to obtain an existence proof about learnability. That said, most studies into the inductive biases of current model learners show that they *lack* biases thought to be important to human learning. Though we may have to wait for advances in models and learning theory for the final answers, experiments with current models still provide new evidence about what advantages humans do and do not need to acquire language.

ACKNOWLEDGMENTS

We are grateful to Ailis Cournane, Najoung Kim, Will Merrill, and Grusha Prasad for comments on this draft, and to audiences at the Machine Learning for Language group at NYU and Computational and Pyscholinguistics Lab group at NYU and

Johns Hopkins. We also thank Roger Levy and Tal Linzen for comments on a much earlier version of this work.

This project has benefited from financial support to SB by Eric and Wendy Schmidt (made by recommendation of the Schmidt Futures program), Samsung Research (under the project *Improving Deep Learning using Latent Structure*), Apple, and Intuit. This material is based upon work supported by the National Science Foundation under Grant No. 1850208. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Bibliography

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

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

[3] Marco Baroni. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *arXiv:2106.08694 [cs]*, June 2021. arXiv: 2106.08694.

[4] 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.

[5] 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.

[6] 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.

[7] 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. arXiv: 1504.00325.

[8] 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.

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

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

[11] Noam Chomsky. Biolinguistic Explorations: Design, Development, Evolution. *International Journal of Philosophical Studies*, 15(1):1–21, January 2007.

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

[13] 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.

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

[15] 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. arXiv: 1412.3555.

[16] Alexander Clark and Shalom Lappin. *Linguistic nativism and the poverty of the stimulus*. John Wiley & Sons, 2010.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Gabe Dupre. (What) Can Deep Learning Contribute to Theoretical Linguistics? *Minds and Machines*, September 2021.

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

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

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

[27] 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.

[28] 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.

[29] Jane Gillette, Henry Gleitman, Lila Gleitman, and Anne Lederer. Human simulations of vocabulary learning. page 42, 1999.

[30] 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.

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

[32] 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.

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

[34] 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.

[35] 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. Publisher: American Psychological Association.

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

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

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

[39] 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.

[40] 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.

[41] 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.

[42] Daniel Keysers, Nathanael Schárli, 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. arXiv: 1912.09713.

[43] 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.

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

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

[46] Vid Kocijan, Thomas Lukasiewicz, Ernest Davis, Gary Marcus, and Leora Morgenstern. A Review of Winograd Schema Challenge Datasets and Approaches. *arXiv:2004.13831 [cs]*, April 2020. arXiv: 2004.13831.

[47] 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.

[48] Brenden 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.

[49] 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.

[50] 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. arXiv: 2102.01192.

[51] 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.

[52] Angeliki Lazaridou and Marco Baroni. Emergent Multi-Agent Communication in the Deep Learning Era. *arXiv:2006.02419 [cs]*, July 2020. arXiv: 2006.02419.

[53] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-Agent Cooperation and the Emergence of (Natural) Language. In *International Conference on Learning Representations*, March 2017. arXiv: 1612.07182.

[54] 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.

[55] 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.

[56] 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.

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

[58] Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, March 2008.

[59] 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.

[60] 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, 2003. Publisher: Elsevier.

[61] 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.

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

[63] 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.

[64] 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.

[65] 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*, 2019.

[66] Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. Predicting Inductive Biases of Fine-tuned Models. In *International Conference on Learning Representations*, 2021.

[67] 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.

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

[69] Christopher D. Manning. Computational Linguistics and Deep Learning. *Computational Linguistics*, 41(4):701–707, December 2015.

[70] 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.

[71] 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.

[72] 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.

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

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

[75] 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. arXiv: 2011.11588.

[76] Christian W. Omlin and Sean Snyders. Inductive bias strength in knowledge-based neural networks: application to magnetic resonance spectroscopy of breast tissues. *Artificial Intelligence in Medicine*, 28(2):121–140, June 2003.

[77] 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.

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

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

[80] 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.

[81] Jackson Petty and Robert Frank. Transformers Generalize Linearly. *arXiv:2109.12036 [cs]*, September 2021. arXiv: 2109.12036.

[82] 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.

[83] Laura Pérez-Mayos, Miguel Ballesteros, and Leo Wanner. How much pretraining data do language models need to learn syntax? September 2021.

[84] 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.

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

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

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

[92] 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.

[93] 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.

[94] 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.

[95] Annmarie Van Dooren, Anouk Dieuleveut, Ail\'{i}s Cournane, and Valentine Hacquard. Figuring out root and epistemic uses for modals: The role of the input. *Journal of Semantics*.

[96] 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.

[97] 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.

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

[99] 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.

[100] 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*, 8:377–392, 2020. _eprint: https://doi.org/10.1162/tacl_a_00321.

[101] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Corpus of Linguistic Acceptability, 2018.

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

[103] 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 2020. Association for Computational Linguistics.

[104] 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.

[105] 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.

[106] 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.

[107] Charles D Yang. *Knowledge and learning in natural language*. PhD Thesis, Massachusetts Institute of Technology, 2000.

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

[109] 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.