2024. *Proc Ling Soc Amer* 9(1). 5693. https://doi.org/10.3765/plsa.v9i1.5693.

What language models can tell us about learning adjectives

Megan Gotowski & Forrest Davis*

Abstract. It has been argued that language models (LMs) inform our knowledge of language acquisition. While LMs are claimed to replicate aspects of grammatical knowledge, it remains unclear how this translates to acquisition directly. We ask if a language model trained specifically on child-directed speech (CDS) is able to capture grammatical knowledge of adjectives. Ultimately, our results reveal that what the model is "learning" is how adjectives are distributed in CDS, and not the grammatical properties of different adjective classes. While highlighting the ability of LMs to learn distributional information, these findings suggest that LMs alone cannot explain how children generalize beyond their input.

Keywords. word learning; adjectives; language models; corpus research

1. Introduction. Language is regarded as a unique aspect of human cognition. All typically-developing children acquire language without any explicit instruction, rapidly building a vocabulary and displaying a sensitivity to linguistic information from birth. Research in acquisition has focused on both *what* children know at various stages in development, as well as *how* children learn. There are different theories in the literature on the acquisition process, which primarily differ in terms of the assumptions about the "scaffolding" that children have access to along the way– some theories assume that children have innate knowledge of language, in the form of certain expectations that constrain the learning process (Chomsky 1965), other theories assume that learning does not involve any language-specific mechanisms, but may be captured by appealing to broader aspects of cognition, and conceptualize the acquisition process in terms of distributional learning and pattern recognition (see e.g., Saffran et al. 2006; Romberg & Saffron 2010; among others). Different theories also make different claims about the importance of the input on learning; while the input is always critical for learning, theories that advocate for distributional or statistical learning claim that the input is sufficient on its own.

Moreover, recent work in computational linguistics has also placed renewed attention on this debate, specifically the contribution of the input, and has argued that neural models of language offer new ways of understanding the language learning process (see Warstadt & Bowman 2023). Support for reasoning that language models are useful as models of acquisition stems from overlap between neural models and human grammaticality judgments (e.g., Warstadt et al. 2020), reading times (at least qualitatively, see van Schijndel and Linzen, 2021), and brain data (e.g., Schrimpf et al. 2021). However, there remains a crucial inferential gap: language models are trained on significantly more data than a human child experiences (cf. Linzen, 2020). While attempts have been made to address this issue (for example, Huebner et al. (2021) trained smaller models on CHILDES), it remains an open question as to whether a full adult-like grammar emerges from a more impoverished setting.

The ability of neural models to mimic human linguistic behaviors is often taken as evidence for non-nativist positions. In particular, they are argued to be extensions of statistical learning

* Special thanks to audiences at MIT for feedback on earlier versions of this research. Authors: Megan Gotowski, Pomona College (megan.gotowski@pomona.edu) & Forrest Davis, Colgate University (fdavis@colgate.edu).

(e.g., Chang & Bergen 2022; Contreras Kallens et al. 2023). Learning for a language model proceeds from the statistical regularities in the input, with no a priori commitment to the representational content of language. In linking models to acquisition, the claim is that grammatical knowledge follows directly from distributions observed in corpora. At least two, interrelated, components are necessary for this claim to hold: (i) linguistic data should contain the distributions corresponding to grammatical knowledge, and (ii) neural models should be able to recover these facts (see the discussion of superficialism in Davis (2022)).

Prior work has often assumed (i) – linguistic data has the necessary information – and instead, evaluated models for correct behavior. Some work within natural language processing has argued that other modalities may be necessary for robust human-like systems (e.g., Bender & Koller 2020; Bisk et al. 2020). However, the limitations of specific modalities are usually connected to types of knowledge, for example, color, taste, etc., rather than gaps in linguistic knowledge itself (i.e. whether linguistic contrasts or grammatical categories are learnable from linguistic data). In this study, we explore both (i) and (ii), asking both what (theoretically relevant) information is available in child-directed speech (CDS), and what a neural model learns when trained on CDS. Specifically, we address the issue of the contribution of language models as applied to acquisition by focusing on adjectives as a case study.

Research on language models has largely investigated particular rules (e.g., agreement, co-reference) that are robustly attested in corpora and/or are extremely regularized. However, we know that the accuracy of language models decreases as we move from frequently-attested and robust phenomena to more nuanced properties of the grammar (e.g., Davis 2022). Thus, in order to determine what they are actually learning, we argue that more "ambiguous" data should be considered – specifically, surface strings that could support different interpretations – given that the child has to learn a range of constructions and rules for which the availability of evidence varies in terms of its frequency or informativeness. Adjectives provide a challenge due to the fact that they do not constitute a homogeneous class; thus, despite the fact that adjectives surface in the input to children, the child still needs to distinguish between subtypes of adjectives that have differing syntactic and semantic properties.

In the paper, we first situate our work within the existing literature on word learning and language models in Section 2. We then turn to our study, detailing the underlying corpus data we use, the model we evaluate, and our particular methodology. After presenting our results in Section 3, we turn to a discussion of how language models address – or fail to address – acquisition. Ultimately, we find that neural models obtain a state more closely resembling the distributions in the data, rather than the one mimicking the target grammar. As such, we conclude by suggesting that neural models are productive tools for exposing the (complicated) relationship between input and grammar.

2. Background. As children are building their early vocabularies, they must map a word form to a meaning. Previous research has illustrated that children benefit from recruiting the syntax as they learn word meanings, in a process known as *syntactic bootstrapping* (Landau & Gleitman 1985; Gleitman 1990; Gleitman et al. 2005). One of the core assumptions of bootstrapping is that the syntax functions as a filter—enabling learners to narrow down the hypothesis space, ruling in certain meanings, and ruling others out. Most of the early research on bootstrapping has focused on verb learning and has demonstrated that children are sensitive to argument structure (see e.g., Naigles 1990; Papafragou et al. 2007; Yuan & Fisher 2009; among many others). For example, if a child encounters the novel verb *gorp* in (1), they may think that it means *kick* or *throw*, but not

sleep or *think*. This is because the syntactic environment – referred to as a frame – informs the learner that this is a verb that selects for an object. Conversely, this same verb in (2) could mean *sleep*, but not *kick* or *throw*. In (3), the change in frame again constrains the meaning, such that the verb could mean *think* but not *kick* or *sleep*, given the inclusion of the clausal complement.

(1) The dax gorps the blicket.
(2) The dax gorps.
(3) The dax gorps that the blicket pilks.

In other words, children subcategorize verbs by paying attention to the number of arguments and the type(s) of complement (see discussion in Lidz 2020). Beyond the verbal domain, research has indicated that distributional cues are also helpful for learning other grammatical categories, such as adjectives (see Syrett 2007; Syrett & Lidz 2010; Becker 2015; Gotowski 2022). Although not receiving as much attention in the literature, adjectives provide an interesting case study for the nature of bootstrapping because they are compatible with a range of frames—and unlike verbs, the environment in which an adjective is found may not be informative. In both (4a) and (4b), the novel adjective *daxy* is consistent with just about any adjective meaning, and adjectives are often found in these sorts of "neutral environments" (Gotowski 2022). Research on child-directed speech indicates that they are indeed often found in prenominal position (4b) (Davies et al. 2020).[1]

(4) a. This one is daxy.
    b. This is a daxy one.

Nevertheless, not all adjectives behave alike. In order to distinguish (sub)types of adjectives, the availability of frames is critical—whereas (4) might be neutral in terms of informativity, (5) is only compatible with a subset of adjectives, like *tough*, *easy*, or *fun*.

(5) It is daxy to do this.

Learners must be able to extract cues (such as expletive subjects and infinitival clauses) from the input in order to successfully pair an adjective with a meaning. While the *availability* of frames in the input is arguably necessary, it is still an open question as to whether or not the input alone is sufficient for learning theoretically-motivated categories. In order to address this question, we first consider different subtypes of adjectives and the syntactic frames associated with them, before discussing prior research on the application of language models in acquisition research.

2.1. TYPES OF ADJECTIVES IN THE TARGET GRAMMAR. Adjectives do not form a homogenous class of predicates – rather, they differ in terms of their syntactic and semantic profiles. Here we focus on *subjective* adjectives. These are a subtype of gradable predicates that participate in what is known as faultless disagreement, as in (6) below. In other words, it is possible for A to assert *p* and B to negate *p* without a contradiction arising (see Stephenson 2007; Anand 2009; Lasersohn 2009; Pearson 2013; among others). This is because these predicates express properties that are

[1] In conducting preliminary analyses of CHILDES, we found that prenominal adjectives are more frequent, occurring around 72% of the time.

evaluated with respect to a judge, which is often co-indexed with the speaker. Conversely, non-subjective adjectives do not participate in faultless disagreement (7).

(6) A: This game is fun!
B: This game is not fun!

(7) A: This chair is wooden.
B: #This chair is not wooden.

Additionally, these subjective adjectives may be further subcategorized into different classes. We follow the convention in e.g., Bylinina (2014) and Gotowski (2022) in referring to these classes in terms of a canonical member of that class, and summarize five distinct subtypes of interest: the TOUGH, SMART, PRETTY, TASTY and TALL-class. Here in this section, we briefly review the distributional signatures associated with each class.

Adjectives in the TOUGH-class (e.g., *tough, easy, difficult, fun*) denote properties of events, as opposed to individual-denoting DPs, and thus select for an infinitival clause (Lasnik & Fiengo 1974; Hartman 2012; Hicks 2003; among others). In other words, these adjectives denote a property of the lower clause, and not the matrix subject. While the infinitival clause may be overt or implicit (if recoverable from linguistic context; see discussion in Gluckman 2021), the subject is interpreted as the object (8a). Because there is no thematic relation between the TOUGH-class predicate and the subject position, an expletive *it* may also be inserted (8b), or a verbal subject (gerund) (8c), giving rise to the *tough*-alternation.

(8) a. This paper is hard (to write) ___.
b. It is hard (for me) to write this paper.
c. Writing this paper is hard.

In these constructions, the judge/experiencer may be expressed with the inclusion of a *for*-phrase (8b).

Adjectives in the SMART-class (e.g. *smart, nice, silly*) behave quite similarly to those in the TOUGH-class adjectives on the surface because they too exhibit a syntactic alternation between a subject DP (9a), an expletive (9b) and gerund (9c), and are compatible with infinitival clauses.

(9) a. Jane is smart [PRO to take this class].
b. It is smart (of Jane) to take this class.
c. Taking this class is smart.

However, the adjectives in this class are distinguished by their ability to participate in evaluative control (EC; see Stowell 1991; Kertz 2010), in which the subject DP controls PRO in the lower clause (9a), and (unlike the TOUGH-class) is not interpreted as the object of the lower clause. The adjective denotes a property of an individual. As with the TOUGH-class, it is possible to express an experiencer overtly, but often with the inclusion an *of*-phrase, as in (9b).

The PRETTY-class (e.g. *pretty, beautiful, handsome*) consists of what are at times referred to as "aesthetic adjectives" (Stojanovic 2007). These adjectives denote properties of individuals, not events, and yet they optionally surface with an infinitival clause. In such constructions (as in (10)), the subject is interpreted as the object of the lower clause. On the surface, this causes them to resemble TOUGH-class adjectives– crucially, however, expletives and verbal subjects are not possible and there is no alternation as in (8) or (9).

(10) This dress is pretty (to wear).

The adjectives in the TASTY-class (e.g. *tasty, spicy*) are those that are often referred to as predicates of personal taste (PPTs), and denote properties of individuals. From a distributional standpoint, these differ from the previous three classes as they are incompatible with infinitival clauses (11), and expletives (12a) and verbal subjects (12b) are likewise not possible. However, these adjectives optionally select for an overt judge phrase (13).

(11) The pasta tastes delicious (?? to eat).
(12) a. *It is delicious to eat this pasta.
b. *Eating this pasta is delicious.
(13) Carolina Reapers are spicy to/for me.

Lastly, adjectives in the TALL-class (e.g. *tall, big, old*) differ from the previous classes in several ways. First, although these adjectives are subjective, they do not combine with a judge *for*-phrase. However, they combine with different type of (surface-similar) *for*-phrase, denoting the comparison class (14) that restricts how the adjective is evaluated.

(14) John is tall for a 3rd grader.
(15) John is tall (*for me).

As with the TASTY-class these denote properties of individuals alone and are not compatible with infinitival clauses, unless modified with the excessive *too* (16b) to form a standard.

(16) a. John is 6 feet tall (*to ride the roller coaster).
b. John is too tall to ride the roller coaster.

In sum, all of these classes have distinct distributional patterns associated with them (for more information on the profiles of these adjectives and the ability of learners to recruit such cues, please see Gotowski (2022) and Gotowski & Syrett (under review)). The environments in which adjectives are found should signal subcategory membership.

2.2. NEURAL MODELS AND ACQUISITION. As this work emphasizes the properties of input data, it is worth highlighting how language models are trained and what aspects of context they utilize. At their core, all language models learn mappings between strings and probabilities. They do so by learning to predict words conditioned on a context.[2] For a model like BERT (Devlin et al. 2019), the context for prediction is the entire sentence or paragraph. A subset of the words in the sentence (or paragraph) are sampled for the model to predict (termed masking) using the surrounding context as a signal to facilitate the prediction. Concretely, consider the string "the cat is hungry". During training, the model might have to predict the word "cat" conditioned by the context "the" to the left and "is hungry" to the right. Through the course of training, the model builds representations which facilitate aligning the model's prediction with the distribution observed in the training data.

Despite the simplicity of the learning objective – predict words in a context – and the form of the data, transformer language models have been claimed to learn a range of linguistic

[2] Models make predictions about tokens which can be words or sub-words.

phenomena. Points of overlap include subject-verb agreement (e.g., Warstadt et al. 2020), syntactic islands (e.g., Wilcox et al. 2022), negative-polarity items (e.g., Warstadt et al. 2019), and aspects of control constructions (e.g., Stengel-Eskin and Van Durme 2022). Despite these apparent successes, there are noted limitations. For example, neural models struggle with infrequent constructions (e.g., Wei et al. 2021) and in resolving conflicting linguistic processes (e.g., Davis 2022).

The majority of studies linking neural models to (human) linguistic knowledge emphasize the adult grammar. Nonetheless, increased overlap between model and human behavior has fueled interest in understanding neural models as models of acquisition (e.g., Warstadt and Bowman, 2022). There are immediate difficulties in comparing neural model learning and child acquisition. Presently, models are trained on vast amounts of text data – BERT, for example, was exposed to ~33 times the data a human experiences in maturation (Linzen 2020). Moreover, it appears that the dominant modeling approach requires this much data in order for a model to learn (at least some) linguistic generalizations (Warstadt et al. 2020b). To draw more meaningful comparisons between computational models and human language learning, something must be done to address the gap.

Two approaches for comparing language acquisition and neural models have crystallized in recent work. One focuses on data, and the other learning trajectories. With respect to data, the aim is to shift model training from vast amounts of internet text to both smaller quantities (e.g., the emphasis on 100 million words in Linzen (2020); Hosseini et al. (2022)) and to a different distribution (mainly child-directed speech; e.g., Pannitto & Herbelot 2020; Huebner et al. 2021; Yedetore et al. 2023). Smaller amounts of naturalistic data are meant to clarify whether models obtain similar levels of knowledge under more human-like conditions. In the present study, we evaluated a model trained under these conditions, BabyBERTa (Huebner et al. 2021). BabyBERTa was trained on data from CHILDES that they estimate as the same amount of language as a 6-year-old child is exposed to. Adjusting for this more impoverished setting, Huebner et al. find that BabyBERTa performs comparably to much larger models.

With respect to learning trajectories, the emphasis is on understanding whether models and children learn phenomena in the same order, despite differences in amount of exposure to language (e.g., Chang & Bergen 2022; Evanson et al. 2023). There appears to be some degree of overlap in order of acquisition (e.g., Lavechin et al. 2023), though there are systematic gaps stemming from the text-only data of the model (Chang & Bergen 2022). That is, models lag in their acquisition of words relating to the physical or social world (e.g., taste, love). Within the context of this study, we set aside the development of knowledge in neural models and instead focus on what form the final state of the model takes. In contrast to much existing work, we also focus on what would be *reasonable* for a model to learn by carefully investigating what information can be ideally extracted from the training data (here, child-directed speech).

To this end, we might imagine that the learning process unfolds in one of two ways: (i) learners could have expectations for each of these classes. They might have knowledge of what to look for and what to filter. Learners would be biased in terms of syntactic and semantic information. Alternatively, (ii) learners may not have any expectations, and classes are extracted purely from the input. There is no built-in filter, but simply a calculation across the distribution. Learners would be biased in terms of frequencies.

3. Corpus study. In this study, we analyze the distribution of the five classes of subjective adjectives in the input to the child, relying on multiple CHILDES corpora, and compare what we find in the input to the output of a model trained on the same set of CHILDES data.[3]

3.1. CORPORA. We analyze 42 corpora from CHILDES.[4] The corpora included here importantly feature conversations between parents or other caregivers and children. We chose CHILDES because we wanted to investigate the actual input that children are exposed to, as opposed to other sources of data containing either language from printed texts (e.g., newspapers, Wikipedia) or conversations directed at adults. Additionally using CHILDES creates a smaller sample size, which we argue more realistically depicts the input to the child learner.

3.2. METHODS. We first extracted all utterances with adjectives produced by mothers (found in the *MOT tiers) for each corpus listed above. We scraped all of the data from the corpora using the TalkBank browser, and downloaded the results to Excel.[5] We recorded the adjective and the utterance it was found in. From here, we extracted five adjectives from each of the five adjective classes. We selected the top 3 most frequent adjectives, and the 2 least frequent adjectives (with at least 10 hits across all of the corpora): *good, bad, hard, evil, impossible* (TOUGH); *funny, spooky, yummy, soft, basic* (TASTY); *big, little, long, chilly, rich* (TALL); *pretty, beautiful, lovely, posh, splendid* (PRETTY); *nice, silly, mean, foolish, cruel* (SMART). We chose to balance the sample in this way to account for any potential frequency effects; it is possible that certain adjectives are simply preferred by a model because they are found more often than others. The result of this initial procedure is a CHILDES sample with around 100,000 tokens of 25 adjective types. We used this sample for our analysis. The data was analyzed in two subsequent stages, which we will refer to as Annotated Sampling and Model Analysis. For the annotated sample, we selected a subset of the sample, featuring 10 of the 46 corpora to be coded by hand. We followed the coding schema in Gotowski (2022), in which we code for surface-level distributional cues around the target adjective. We focus on the environment to the left and right of the adjective, and code for the following subset of environments in (17), previewed in the previous section: infinitival complement clauses (a), gerunds (b), experiencer *to*-phrases (c), *of*-phrases, and (d) *for*-phrases. Note that the function of the phrases in (d-e) differs, but we code based on distribution alone, given that on the surface these phrases are ambiguous. We return to this issue when discussing our results.

(17) a. It is **hard** to open.
b. Writing is **easy**. / It is nice seeing that.
c. This show is **funny** to me.
d. It is **nice** of us to bring a gift.
e. It is **good** for us to be early. / John is **short** for a basketball player.

This sample provides us with a snapshot of the distribution, and enabled us to compare the model results. In the next phase of the study, however, the model is trained on all corpora.

3.3. SAMPLE (HUMAN CODED). Here we report the distribution of adjectives in the sample within each class, through the lens of theoretically-motivated environments, to address the availability

[3] Code for the following sections can be found on GitHub: https://github.com/forrestdavis/Adjectives.
[4] Details about the specific corpora consulted can be found here: https://github.com/forrestdavis/Adjectives/blob/main/CHILDES-Corpora-References.md.
[5] We used https://naclo.cs.umass.edu/childes-search/ made available by Andrew Wang at UMass Amherst and the childesr package: https://github.com/langcog/childesr.

of these environments with adjective classes to evaluate the richness of the input to the child. We consider the distributional profiles of these adjectives as the "idealized" input for the learner– that is, a case where every adjective is found in a compatible informative frame, and the likelihood of that adjective being in that frame is equally distributed based on the compatibility (acceptability) with the frame of interest. We are able to compare the actual input from CDS to this idealized learning situation.

The results of the annotated sampling are found in Table 1, with grey shading to indicate where we would expect to find a higher percentage of adjectives with the frame of interest. Note that here we report the distribution of each frame between the adjective classes in our sample (i.e. when frame X surfaces, how often is it found with class Y), and not the percentage of adjectives within these classes found in that frame. Overall, we find that general patterns that we expect to find based on the literature are indeed reflected in CDS. For example, gerunds are possible with TOUGH-class and SMART-class adjectives, but not the other adjective classes, and this is what we find here. However, note that the distributions are skewed in particular ways that favor certain adjective classes, regardless of acceptability. For example, although three of the five adjective classes are actually compatible with infinitival complements, these are found with a TOUGH-class adjective 66% of the time, and rarely ever with PRETTY-class, suggesting that this environment actually *favors* TOUGH-adjectives.

	Infinitival	Gerunds	For XP	To XP	Of XP
TOUGH	66%	64%	44%	27%	14%
SMART	20%	32%	6%	72%	84%
TALL	12%	4%	48%	1%	2%
TASTY	1%	0%	2%	0%	0%
PRETTY	1%	0%	0%	0%	0%

Table 1. Results of the human coded sample with grey shading indicating where we expect to find a higher percentage of adjectives with the relevant frame

This annotated sampling analysis also reveals interesting gaps. The fact that certain adjective classes are more often found in particular environments, while others seem to be "underrepresented" in the sample (largely failing to surface in these environments), poses a potential challenge for learning. Given that children successfully learn the meanings of words like e.g. *pretty,* we want to be careful to say that children are learning from negative evidence to distinguish e.g. *tough* from *tall*, given that *pretty* is rarely found in informative frames, and yet there is no reason to believe that children are delayed in acquiring PRETTY-class adjectives (see Blackwell 2005).

From the annotated sample, we are able to conclude that the input to the child contains both (informative) distributional cues and (uninformative) noise, and that the distribution is skewed in favor of particular adjective classes with certain frames—at times even creating an "illusion of ungrammaticality," i.e. a situation in which an adjective that is compatible with a frame is simply rarely found with it. However, perhaps the input alone is still enough to support learning. Here is where we turn to the second part of this study: the model analysis. Recall that the model is trained on the distribution alone. If the model is learning how to subcategorize from the input available, we might expect it to perform based on frame compatibility. Conversely, if the model is simply "learning" frequency-based patterns, we expect the model will replicate the same distributional patterns in the CHILDES sample.

3.4. MODELING. We evaluated BabyBERTa, a masked language trained on CHILDES (Huebner et al., 2021).[6] Recall, despite the much smaller and simpler training data, BabyBERTa achieved similar accuracy to larger models on a grammar test suite (see Table 1 in Huebner et al. (2021)) suggesting that the model captures (aspects) of the adult grammar. In this work, we evaluated its knowledge of adjectives and their syntactic environments.

In order to evaluate the model on a broader set of sentences, we automatically coded utterances (totaling around 100,000 tokens) for adjectives and their syntactic environments. This was done by extracting information from the part of speech labels tier in CHILDES. Hand-written rules grouped part-of-speech labels into phrases (e.g., "det:num n" becomes a noun phrase). Additionally, for/to/of/with phrases were determined by the form of the preposition tagged in CHILDES (excluding "to" as infinitivals). Finally, existing tags marked infinitivals (e.g., inf tag) and gerunds (e.g., n: gerund). Syntactic environments for an adjective were determined by the presence of a desired environment (e.g., "for XP") directly to the right of the adjective.[7] The automatic coding was able to recapitulate the labels determined via hand-coding above. A full breakdown of the automatically labeled data is reported in Table 2.[8]

	Infinitival	Gerunds	For XP	To XP	Of XP
TOUGH	52%	18%	30%	24%	9%
SMART	16%	6%	4%	44%	9%
TALL	28%	68%	60%	25%	79%
TASTY	3%	7%	4%	7%	2%
PRETTY	0%	1%	2%	0%	1%

Table 2. Results of the automatically coded sample

We gathered the likelihood assigned by the language model of each adjective in the annotated data using the masking strategy advocated in Kauf & Ivanova (2023). Our aim is to evaluate whether BabyBERTa has learned how to subcategorize adjectives. Note that many syntactic environments are compatible with more than one adjective class, so we evaluate the ability of BabyBERTa to associate syntactic environments with any of the grammatical adjective classes for a given frame. For example, the utterance in (18) was annotated as containing an infinitival complement.

(18) see it's [redacted] to read when there are so many toys

To investigate whether BabyBERTa associates adjective classes consistently and grammatically, we gathered the probabilities of each of the adjectives, such as e.g., *bad, yummy, big, splendid, and nice*, in the place of the adjective *hard* (which has been masked). If BabyBERTa has learned the correct grammatical contrasts, we predict that *big* and *yummy* (exponents of the TALL and TASTY classes, respectively) should be assigned a lower probability than, *bad, splendid*, and *nice* (exponents of the TOUGH, PRETTY, and SMART classes, respectively). That is, TOUGH, PRETTY, and

[6] We accessed the model version 'phueb/BabyBERTa-1' via HuggingFace: https://huggingface.co/phueb/BabyBERTa-1

[7] We considered looser inclusion criteria for associating an adjective with an environment (including, anywhere in the same sentence) to potentially better approximate a truly naïve model. Qualitatively similar distributions of environments and adjectives were obtained in these looser cases, so we focus here on the stricter set for evaluating models.

[8] The code for automatically labeling CHILDES utterances and gathering model preferences can be found here: https://anonymous.4open.science/r/Adjectives-D11E/README.md

SMART classes are grammatically licensed by this syntactic environment, while TALL and TASTY are not. Empirically, for this sentence and those adjectives, we observed that BabyBERTa exhibited the following ranked order: *nice* (smart) > *big* (tall) > *bad* (tough) > *yummy* (tasty) > *splendid* (pretty). While partially aligned with the desired behavior (preferring SMART), there are notable differences with a fully grammatical system (e.g., TALL and TASTY are more likely than PRETTY).

	Infinitival	Gerunds	For XP	To XP	Of XP
TOUGH	61%	25%	48%	36%	6%
SMART	14%	5%	2%	42%	18%
TALL	25%	65%	49%	22%	76%
TASTY	0.2%	5%	1%	0.2%	0%
PRETTY	0.1%	1%	0.2%	0%	0%

Table 3. BabyBERTa Preferences by Syntactic Environment with red shading indicating where the model incorrectly predicts a frame to be compatible

In Table 3, the distribution of adjective classes predicted by BabyBERTa conditioned on each syntactic environment is provided. Broadly, we find that the BabyBERTa predictions are aligned with the distribution observed in the corpus. For example, infinitival clauses are favored with TOUGH-class adjectives, "for XP" environments are favored with both TOUGH and TALL-class adjectives, and the gerund environment favors TOUGH and TALL-class adjectives. However, the overlap between BabyBERTa's predictions and the corpora distributions yields behavioral mismatches with the adult grammar. In particular, recall that PRETTY and TASTY-class adjectives are not well represented in the corpora. BabyBERTa does not overcome this sparsity, and subsequently deemed these adjective classes unlikely regardless of the syntactic environment.

BabyBERTa does not fully align with the contours of the distribution in training. Notably, it has acquired some ungrammatical preferences that appear unsupported in our annotated subset of CHILDES. In aggregate, model preferences aligned with grammatical knowledge only around 44% of the time. Critically, frequency does not seem to be the main driver of the depressed accuracy. If we restrict the results to only predictions of infrequent adjectives (where we might expect lower accuracy), we found that the model preferences adhered to grammaticality 46% of the time. Zooming in on particular adjective classes, TALL-class adjectives have a strikingly ungrammatical profile. Recall, they are licensed in the "for XP" syntactic environment. These facts are supported by our hand-coded data, where we found the vast majority of the distribution of TALL-class adjectives concentrated with "for XP" constructions. Nonetheless, BabyBERTa assigns moderate probability to such adjectives across all syntactic environments. Worse yet, the model favors TALL-class adjectives over all other adjective classes in two ungrammatical environments: gerunds and "of XP". Finally, we noted in Section 2.3.1 a challenge to learners restricted to strings when acquiring TALL-class adjectives. Namely, infinitival clauses can occur to the right of such adjectives when under *too* (e.g., *the boy is too tall to ride the coaster*), but not in other contexts. It appears that BabyBERTa has not overcome this challenge and, instead, acquired a preference for TALL-class adjectives without enforcing this long-distance dependency.

Altogether, the results suggest that the model behaved in ways that deviate from the ultimate target grammar. The differences appear to follow from generalizations that are either too tightly tethered to the distributional facts of the training data or (in more limited cases) unrelated to the corpora as we investigated them. Put another way, the model did not address the gaps we

observed in child-directed speech, and even acquired new ones. While encouraging evidence for the ability of such models to acquire distributional knowledge, it leaves open questions of what other information or mechanism may be needed to align model behavior with an idealized distribution as to what the grammar allows for and what it prohibits.

4. Conclusions and discussion. In this paper, we examined whether language models are able to serve as models of language acquisition, as has been claimed in recent work. We approach this topic by focusing on a case study, namely learning adjectives and their corresponding syntactic environments. Our study focused on two dimensions: (i) the properties of the input that children and models are exposed to, and (ii) an assessment of a language model's behavior. Importantly, in our interpretation of the results we distinguish distribution (i.e. frequency) in the input from the syntactic behavior of different adjective classes. That is, we separate the adult-like target grammar from the facts of primary linguistic data. We argue that this three-fold understanding of the data, the grammar, and the learner model (e.g., language model) is important for both clarifying what model performance is indicating, and linking results from natural language processing to the study of acquisition.

Although our results suggest that the language model does succeed at discovering patterns, a critical concern, however, is the fact that the model does not seem to be "learning" adjective categories, but simply learning distributional frequencies, largely replicating what is in CHILDES. While this is on the one hand not surprising that the model would reflect the data it was trained on, it does illuminate a potential complication of any attempt to claim that models are truly replicating the acquisition process. Children are not simply beholden to mere input frequencies. We know from previous research that certain constructions are rarely found in the input, and children still learn them with ease (see e.g., Lidz et al. 2003). Thus, the successful learning by language models of distributional properties presents a puzzle: what information is necessary for acquiring the target grammar? Dominant approaches in natural language processing (implicitly) assume that this process is facilitated by more language data. Our corpus results suggest that this will not do, data appear to systematically differ from the target grammar. Although studies on word learning with adults suggests that learners are able to recruit the environments that we have examined– despite the relative rarity of informative frames in the input, participants nevertheless succeed at recruiting syntactic information (Gotowski 2022) – it remains to be determined how children "close the gap" and figure out how to subcategorize adjectives that are often not found in informative frames.

This seems especially important in addressing how children learn e.g., the PRETTY and TASTY-class adjectives, which are rarely found with a frame in the human-coded sample, and are also never preferred by the model. One possibility is that there are additional cues in the input that are not accounted for here (other frames that are found more often with these adjectives, in terms of distribution), or that there are cues that are not syntactic in nature that would help learners converge on a meaning. For instance, the fact that PRETTY-class adjectives are known as "aesthetic" adjectives that describe visual properties, and TASTY-class adjectives often describe sensory experiences (e.g., taste, smell, touch) might hint at the contribution of lexical semantics. Existing work within natural language processing has emphasized that text-only data lack a variety of sources of data, including sensory and social experiences (e.g., the "learning environment" in Warstadt & Bowman (2022), or the limitations of text only data discussed in Bender & Koller (2020); Bisk et al. (2020)). It is not clear how the model could pick up on these sorts of properties without new techniques or sources of data.

While our results suggest that language models as they are currently trained fall short of meaningfully capturing the language acquisition process, language models are nonetheless useful tools. Their success at embodying distributional or statistical learning (e.g., Chang & Bergen 2022; Contreras Kallens et al. 2023) allows one to critically evaluate what information is readily available in the acquisition process. In finding divergences between such models and human behavior, we find evidence for a divergence between input data and the adult grammar. Insights about data ultimately both strengthen our understanding of the successes and limitations of current language models and also highlight instances where humans leverage additional resources to bridge the gap between their experience with language and their knowledge of grammar.

References

Anand, Pranav. 2009. Kinds of taste. Manuscript. UC Santa Cruz.

Becker, Misha. 2015. Animacy and the acquisition of tough adjectives. *Language Acquisition* 22(1). 68–103. https://doi.org/10.1080/10489223.2014.928298.

Bender, Emily & Andrew Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 58. 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463.

Bisk, Yonathan, Ari Holtzman, … & Joseph Turian. 2020. Experience grounds language. In Bonnie Webber, Trevor Cohn, Yulan He & Yang Liu (eds.), *Proceedings of EMNLP*, 8718–8735. Online: ACL Anthology. https://doi.org/10.18653/v1/2020.emnlp-main.703.

Blackwell, Aleka. 2005. Acquiring the English adjective lexicon: Relationships with input properties and adjectival semantic typology. *Journal of Child Language* 32(3). 535–562. https://doi.org/10.1017/S0305000905006938.

Bylinina, Lisa. 2014. *The grammar of standards: Judge-dependence, purpose-relativity, and comparison classes in degree constructions*. Amsterdam: Netherlands Graduate School of Linguistics dissertation.

Caselli, Maria Cristina, Elizabeth Bates, Paola Casadio, Judi Fenson, Larry Fenson, Lisa Sanderl & Judy Weir. 1995. A cross-linguistic study of early lexical development. *Cognitive Development* 10(2). 159–199. https://doi.org/10.1016/0885-2014(95)90008-X.

Chang, Tyler & Benjamin Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics* 10. 1–16. https://doi.org/gpcvx9.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Contreras Kallens, Pablo, Ross Dean Kristensen-McLachlan & Morten Christiansen. 2023. Large language models demonstrate the potential of statistical learning in language. *Cognitive Science* 47(3). e13256. https://doi.org/10.1111/cogs.13256.

Davies, Catherine, Jamie Lingwood & Sudha Arunachalam. 2020. Adjective forms and functions in British English child-directed speech. *Journal of Child Language* 47(1). 159–185. https://doi.org/10.1017/S0305000919000242.

Davis, Forrest. 2022. *On the limitations of data: Mismatches between neural models of lan- guage and humans*. Ithaca, NY: Cornell University dissertation.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran & Thamar Solorio (eds.), *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and short papers)*, 4171–4186. Minneapolis: ACL. https://doi.org/10.18653/v1/N19-1423.

Evanson, Linnea, Yair Lakretz, & Jean Remí King. 2023. Language acquisition: Do children and language models follow similar learning stages? In Anna Rogers, JordanBoyd-Graber &

Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics*: *ACL 2023*, 12205–12218. Online: ACL Anthology. https://doi.org/ms6n.
Gleitman, Lila. 1990. The structural sources of verb meanings. *Language Acquisition* 1(1). 3–55. https://doi.org/10.1207/s15327817la0101_2.
Gleitman, Lila, Kimberly Cassidy, Rebecca Nappa, Anna Papafragou & John C. Trueswell. 2005. Hard words. *Language Learning and Development* 1(1). 23–64. https://doi.org/cxgchm.
Gluckman, John. 2021. The meaning of the *tough*-construction. *Natural Language Semantics* 29(3). 453–499. https://doi.org/10.1007/s11050-021-09181-3.
Gotowski, Megan. 2022. *Syntactic bootstrapping in the adjectival domain: Learning subjective adjectives.* New Brunswick, NJ: Rutgers University dissertation. https://rucore.libraries.rutgers.edu/rutgers-lib/67916/.
Gotowski, Megan & Kristen Syrett. under review. Using surface structure to acquire subjective adjective meanings. *Language Acquisition*.
Hartman, Jeremy. 2012. (Non-)intervention in A-movement. *Linguistic Variation* 11(2). 121–148. https://doi.org/10.1075/lv.11.2.01har.
Hicks, Glyn. 2003. *"So easy to look at, so hard to define":* Tough *movement in the Minimalist framework.* York, UK: University of York dissertation.
Hosseini, Eghbal, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky & Evelina Fedorenko. 2022. Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. Preprint in bioRxiv. https://doi.org/10.1101/2022.10.04.510681.
Lasnik, Howard, & Robert Fiengo. 1974. Complement Object Deletion. *Linguistic Inquiry* 5(4). 535–572.
Huebner, Phillip, Elior Sulem, Fisher Cynthia & Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In Arianna Bisazza & Omri Abend (eds.), *Proceedings of the 25th Conference on Computational Natural Language Learning*, 624–646. Online: ACL Anthology. https://doi.org/10.18653/v1/2021.conll-1.49
Kauf, Carina & Anna Ivanova. 2023. A better way to do masked language model scoring. In Anna Rogers, Jordan Boyd-Graber & Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short papers),* 925–935. Toronto: ACL. https://doi.org/10.18653/v1/2023.acl-short.80.
Kertz, Laura. 2010. The argument structure of evaluative adjectives: A case of pseudo-raising. In Norbert Hornstein & Maria Polinsky (eds.), *Movement theory of control,* 269–298. Amsterdam: John Benjamins. https://doi.org/10.1075/la.154.10ker.
Landau, Barbara & Lila Gleitman. 1985. *Language and experience*. Cambridge: Harvard University Press.
Lasersohn, Peter. 2009. Relative truth, speaker commitment, and control of implicit arguments. *Synthese* 166(2). 359–374. https://doi.org/10.1007/s11229-007-9280-8.
Lavechin, Marvin, Maureen de Seyssel, … & Emmanuel Dupoux. 2022. Can statistical learning bootstrap early language acquisition? A modeling investigation. PsyArXiv preprint. https://doi.org/10.31234/osf.io/rx94d.
Lidz, Jeffery. 2020. Learning, memory and syntactic bootstrapping: A Meditation. *Topics in Cognitive Science* 12. 78–90. https://doi.org/10.1111/tops.12411.
Lidz, Jeffery, Henry Gleitman & Lila Gleitman. 2003. Understanding how input matters: Verb learning and the footprint of universal grammar. *Cognition* 87(3). 151–178. https://doi.org/b726h4.
Linzen, Tal. 2020. How can we accelerate progress towards human-like linguistic generalization? In Dan Jurafsky, Joyce Chai, Natalie Schluter & Joel Tetreault (eds.), *Proceedings of*

the 58th Annual Meeting of the Association for Computational Linguistics, 5210–5217. Online: ACL Anthology. https://doi.org/10.18653/v1/2020.acl-main.465.
Naigles, Letitia. 1990. Children use syntax to learn verb meanings. *Journal of Child Language* 17(2). 357–374. https://doi.org/10.1017/S0305000900013817.
Pannitto, Ludovica, & Aurélie Herbelot. 2020. Recurrent babbling: Evaluating the acquisition of grammar from limited input data. In Raquel Fernández & Tal Linzen (eds.), *Proceedings of the 24th Conference on Computational Natural Language Learning,* 165–176. Online: ACL Anthology. https://doi.org/10.18653/v1/2020.conll-1.13.
Papafragou, Anna, Kimberly Cassidy, & Lila Gleitman. 2007. When we think about thinking: The acquisition of belief verbs. *Cognition* 105(1). 125–165. https://doi.org/bxrg58.
Pearson, Hazel. 2013. A judge-free semantics for predicates of personal taste. *Journal of Semantics* 30(1). 103–154. https://doi.org/10.1093/jos/ffs001.
Romberg, Alexa R, & Saffran, Jenny R. 2010. Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science* 1(6). 906–914. https://doi.org/10.1002/wcs.78.
Schrimpf, Martin, Idan Blank, … & Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences (PNAS)* 118(45). e2105646118. https://doi.org/10.1073/pnas.2105646118.
Stengel-Eskin, Elias & Benjamin Van Durme. 2022. The curious case of control. In Yoav Goldberg, Zornitsa Kozareva & Yue Zhang *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11065–11076. Abu Dhabi: ACL Anthology. https://doi.org/10.18653/v1/2022.emnlp-main.760.
Stephenson, Tamina. 2007. Judge dependence, epistemic modals, and predicates of personal taste. *Linguistics and Philosophy* 30. 487–525. https://doi.org/10.1007/s10988-008-9023-4.
Stojanovic, Isidora. 2007. Talking about taste: Disagreement, implicit arguments, and relative truth. *Linguistics and Philosophy* 30(6). 691–706. https://doi.org/bpsn2t.
Stowell, Tim. 1991. The alignment of arguments in adjective phrases. In Susan Rothstein (ed.), *Perspectives on phrase structure: Heads and licensing,* 05–135. London: Brill.
Syrett, Kristen. 2007. *Learning about the structure of scales: Adverbial modification and the acquisition of the semantics of gradable adjectives*. Evanston, IL: Northwestern dissertation. https://arch.library.northwestern.edu/concern/generic_works/ms35t867c.
Syrett, Kristen, & Jeffery Lidz. 2010. 30-month-olds use the distribution and meaning of adverbs to interpret novel adjectives. *Language Learning and Development* 6(4). 258–282.
van Schijndel, Marten, & Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science* 45(6). e12988. https://doi.org/10.1111/cogs.12988.
Warstadt, Alex, Yu Cao, … & Samuel R. Bowman. 2019. Investigating BERT's knowledge of Language: Five analysis methods with NPIs. In Kentaro Inui, Jing Jiang, Vincent Ng & Xiaojun Wan (eds.), *Proceedings of the 9th EMNLP-IJCNLP*, 2877–2887. Hong Kong: ACL. https://doi.org/10.18653/v1/D19-1286.
Warstadt, Alex & Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In Shalom Lappin & Jean-Philippe Bernardy (eds.), *Algebraic structures in natural language*, 17–59. Boca Raton: CLC Press. https://doi.org/10.1201/9781003205388.
Warstadt, Alex, Alicia Parrish, … & Samuel R. Bowman. 2020a. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics* 8. 377–392. https://doi.org/10.1162/tacl_a_00321.

Warstadt, Alex, Yian Zhang, Xiaocheng Li, Haokun Liu, & Samuel R. Bowman. 2020. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In Bonnie Webber, Trevor Cohn, Yulan He & Yang Liu (eds.), *Proceedings of EMNLP*, 217–235. Online: ACL Anthology. https://doi.org/10.18653/v1/2020.emnlp-main.16.

Wei, Jason, Dan Garrette, Tal Linzen & Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In Marie-Francine Moens et al. (eds.), *Proceedings of ENLNP*, 932–948. Online: ACL Anthology. https://doi.org/10.18653/v1/2021.emnlp-main.72.

Wilcox, Ethan, Richard Futrell, & Roger Levy. 2022. Using computational models to test syntactic learnability. *Linguistic Inquiry* Online. 1–44. https://doi.org/10.1162/ling_a_00491.

Yedetore, Aditya, Tal Linzen, Robert Frank, & R. Thomas McCoy. 2023. How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. In Anna Rogers, Jordan Boyd-Graber & Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long papers),* 9370–9393. Toronto: ACL. https://doi.org/10.18653/v1/2023.acl-long.521.

Yuan, Sylvia, & Cynthia Fisher. 2009. "Really? She blicked the baby?" Two-year-olds learn combinatorial facts about verbs by listening. *Psychological Science* 20(5). 619–626. https://doi.org/10.1111/j.1467-9280.2009.02341.x.