Collocational bootstrapping: A hypothesis about the learning of subject-verb agreement in humans and neural networks

Claire Hobbs
Cognitive Science Program
Yale University
claire.hobbs@yale.edu

R. Thomas McCoy
Dept. of Linguistics & Wu Tsai Institute
Yale University
tom.mccoy@yale.edu

Abstract

In what ways might statistical signals in linguistic input assist with the acquisition of syntax? Here we hypothesize a mechanism called collocational bootstrapping, in which regularities in word co-occurrence patterns can provide cues to syntactic dependencies. We investigate whether this mechanism can support the acquisition of English subject-verb agreement. First, we simulate language acquisition by training neural networks on synthetic datasets that vary in how predictable their subject-verb pairings are. We find that there is a range of variability levels at which these statistical learners robustly learn subject-verb agreement. We then analyze the variability of subject-verb pairings in child-directed language, and we find that the variability in such data falls within the range that supported robust generalization in our computational simulations. Taken together, these results suggest that collocational bootstrapping is a viable learning strategy for the type of input that children receive.

1 Introduction

The sentences that we encounter do not come annotated with explicit syntax trees. How, then, do children acquire a language's syntax? Some proposals postulate innate predispositions that might guide children toward particular structural analyses (e.g., Chomsky, 1965). Other proposals point to ways in which non-syntactic aspects of the data, such as semantics (Wexler and Culicover, 1980) or prosody (Morgan and Demuth, 1996), might provide helpful cues from which the learner can "bootstrap" syntactic structure. For instance, prosodic boundaries tend to coincide with syntactic boundaries such that observable prosody might point toward accurate analyses of unobservable syntax.

Recent advances in artificial intelligence have given prominence to one particular type of data-driven cue: statistical properties of linguistic strings. Neural network models trained to capture the statistical properties of corpora perform well on tests that target syntactic phenomena such as filler-gap dependencies (Wilcox et al., 2024), negative polarity items (Jumelet and Hupkes, 2018), and subject-auxiliary inversion (Mueller et al., 2022). These systems do not have explicit syntactic predispositions built into them, so their strong syntactic abilities suggest that naturalistic text possesses statistical cues from which much of syntax can be inferred. This evidence *that* statistical properties can pave the way to syntax raises the question of *how* they might do so.

In this work, we propose **collocational bootstrapping** as a hypothesis for one specific way in which statistical cues could contribute to syntactic acquisition. Under this hypothesis, syntactic structure can be inferred from trends regarding which words frequently co-occur. As a case study, we consider English subject-verb agreement, the phenomenon in which a verb must have the same grammatical number as its subject (e.g., *the dogs bark* is grammatical, but *the dogs barks* is not). One challenge for acquiring subject-verb agreement is that there are (at least) two potential rules that could explain most examples of this phenomenon:

(1) AGREE-SUBJECT: A verb should agree with its subject.

(2) AGREE-RECENT: A verb should agree with the closest preceding noun.

We can tell that AGREE-SUBJECT is the correct rule by considering sentences where these rules make different predictions; e.g., when choosing a verb for the sentence *the dogs in the park [bark/barks]*, AGREE-SUBJECT would correctly choose *bark* while AGREE-RECENT would incorrectly choose *barks*. However, for most naturally-occurring sentences, a verb's subject is also the most recent noun, meaning that it may be challenging for learners to identify which rule is correct.

The proposed mechanism of collocational bootstrapping (described in more detail in Section 3) provides one way to select between these two rules even in the absence of direct disambiguating examples such as *the dogs in the park bark*. Under the collocational bootstrapping hypothesis, learners leverage information about word co-occurrence as a window into syntactic dependencies. E.g., given *the dog on the couch barks*, a learner could infer that there is more likely to be a dependency between *dog* and *barks* than between *couch* and *barks* because *dog* is more likely to co-occur with *barks* than *couch* is. After inferring many such potential dependencies, the learner then abstracts away from the specific words that are involved to recognize the abstract syntactic configurations that are truly at the heart of the dependency.

To investigate whether collocational bootstrapping is a viable strategy, we trained multiple neural network language models on synthetically-generated datasets. Because collocational bootstrapping depends on associations between subjects and verbs, we varied the extent to which a subject could be predicted from its verb. Specifically, subjects were sampled from Zipfian distributions (Zipf, 1949) where the probability of a verb's r^{th} most frequent subject is proportional to $1/r^{\alpha}$; varying the parameter α modulates how predictable the subject is given the verb. Critically, the models' training sets were constrained to be fully ambiguous between AGREE-SUBJECT and AGREE-RECENT (e.g., using sentences like *the dog in the park barks*), but the systems were then evaluated on sentences that disambiguated these rules.

We find that subject-verb co-occurrence statistics have a substantial effect on how well the models learn subject-verb agreement; there are some statistical settings (namely, when $\alpha \approx 1.4$, yielding moderate variability) where the models successfully learn AGREE-SUBJECT and others where they do not (namely, when α is very low—producing highly variable data—or very high—producing highly predictable data). The fact that certain statistical configurations support effective generalization supports the collocational bootstrapping hypothesis. Given that collocational bootstrapping is only effective in certain statistical settings, we next perform a corpus analysis of a dataset of child-directed language to see whether children's input has the properties that supported success in our simulations. We find preliminary evidence that child-directed language indeed has the requisite properties.

Overall, our neural-network experiments provide a proof of concept showing that collocational bootstrapping can guide a learner to accurate syntactic analyses, and our corpus analysis suggests that children's input has the statistical properties that make collocational bootstrapping effective. This work is a step toward understanding how quantitative aspects of a learner's input can support the learning of abstract, qualitative syntactic phenomena.[1]

2 Background and Related Work

Bootstrapping in language acquisition: Several mechanisms have been proposed by which learners might infer aspects of syntax from non-syntactic information such as prosody (Morgan and Demuth, 1996) or meaning (Wexler and Culicover, 1980; Pinker, 1984; Abend et al., 2017; Yedetore and Kim, 2024). The most relevant prior proposal is distributional bootstrapping, in which syntactic categories can be inferred from distributional properties—e.g., words occurring in similar contexts likely belong to the same part of speech (Maratsos and Chalkley, 1980; Finch and Chater, 1992; Mintz, 2003). Like distributional bootstrapping, collocational bootstrapping leverages distributional properties of words, but it is a strategy for acquiring relationships between words rather than word categories. Another proposal that is potentially related to collocational bootstrapping is semantic bootstrapping (Wexler and Culicover, 1980); see Section 6 for discussion. The various types of bootstrapping are not mutually exclusive—children might use many or all of them.

Subject-verb agreement in neural networks: A substantial body of work has investigated whether neural networks can learn English subject-verb agreement (Elman, 1991; Linzen et al., 2016). Such networks have been found to be capable of robustly learning subject-verb agreement from naturalistic text (Kuncoro et al., 2018; Gulordava et al., 2018; Goldberg, 2019; Wei et al., 2021). In this work, we train networks on controlled synthetic data to analyze what statistical properties of corpora might be supporting such learning. Our approach shares with Wei et al. (2021) the strategy of training neural networks on corpora that vary in controlled ways, but we investigate a different factor (namely the predictability of the subject given the verb, as opposed to word frequency).

[1] Our code is available on GitHub: https://github.com/ClaireHobbs/collocational-bootstrapping.

Distributional cues to syntax: Both the acquisition literature and the computational literature have discussed what properties of a learner's input might support the acquisition of syntax. Properties discussed include the presence of sentences that might directly disambiguate competing hypotheses (Pullum and Scholz, 2002; Mulligan et al., 2021), the presence of one phenomenon that might be helpful for acquiring different phenomena (Pearl and Mis, 2016; Patil et al., 2024; Misra and Mahowald, 2024; Yang et al., 2026a), the semantic features of a word's arguments (Misra and Kim, 2024), the statistical properties of function words (Yang et al., 2026b), the frequencies of particular words (Wei et al., 2021; Leong and Linzen, 2026), the diversity and complexity of observed syntactic structures (Qin et al., 2025), and the frequencies of syntactic configurations (Wonnacott et al., 2008; Yang, 2016). We instead study the distributional feature of variability in word co-occurrence statistics.

The role of variability in learning: Across domains of cognition, the tradeoff between predictability and variability is a central tension for learning (Raviv et al., 2022): predictable input can support faster learning, while variable input supports more abstract generalizations. Our work applies this idea to the learning of English subject-verb agreement by investigating variability in subject-verb pairings. The most relevant prior papers are Gómez (2002) and Onnis et al. (2004). In both, human participants were shown strings of the form *aXb*, where the first and third elements had an agreement dependency (akin to subject-verb agreement), and the *X* elements were arbitrary. These papers found that people can learn such patterns more readily when the set of possible *X* elements is larger, showing that variability in intervening elements can support learning of nonadjacent syntactic dependencies.

Our work differs from these papers in that we investigate variability in subject-verb pairings, rather than variability in the intervening material. Further, we test artificial neural networks rather than humans. Finally, while these prior papers achieved greater variability by increasing the number of possible *X* entities, we held the relevant sets constant and varied only the frequencies of their elements. This choice means that our conditions differ only quantitatively—there are no qualitative differences regarding which pairings are present (except in one case, the $\alpha \to \infty$ case).

3 The Collocational Bootstrapping Hypothesis

Words are not distributed uniformly in natural language use. Most relevantly for this paper, a given verb is much more likely to have some subjects than others due to the meanings that people are likely to express (e.g., *the dog barked* is much more likely than *the potato barked*). We hypothesize that co-occurrence properties are likely to be especially systematic for words that share a syntactic dependency, such that learners could leverage co-occurrence information to help learn aspects of syntax, such as subject-verb agreement.

As discussed above, there is a tension between predictability and variability. If a given verb always had the same subject, it would be easy for the learner to recognize which noun the verb is paired with such that AGREE-SUBJECT can be selected over AGREE-RECENT. However, the learner in this setting might simply memorize the few subject-verb pairings it has seen, thereby failing to generalize to novel pairings. At the other extreme, if subjects are sampled uniformly, this high degree of variability should support generalization to novel subject-verb pairings, but it would not provide any systematic co-occurrence information that would help point to which noun is the verb's subject.

Given this tension, the key question underlying our first experiment is whether there exists a level of variability that supports correct generalization of subject-verb agreement—a level that is predictable enough to make subject-verb associations apparent yet variable enough to support generalization to novel subject-verb pairs. To get at this question, we use simulations with neural networks trained on simple, synthetic grammars. Using synthetic grammars enables us to fully control and understand which cues are available to the learners so that we can isolate the statistical factors we have highlighted, in the same spirit as other connectionist work that similarly analyzes how neural networks generalize in simple, controlled settings (Elman, 1990, 1991; Frank et al., 2013; McCoy et al., 2020).

4 Experiment 1: Neural Networks

Neural networks allow us to simulate language acquisition in a statistical learner that lacks explicit predispositions for specific syntactic structures. To study the effect of certain statistical properties on learnability, we can train a neural language model on synthetic datasets in which we vary these prop-

Template	Example (Singular)	Example (Plural)
Det N V	the singer sings	the singers sing
Det N PP V	the singer by the dancer sings	the singers by the dancers sing
PP Det N V	by the dancer the singer sings	by the dancers the singers sing
PP Det N PP V	by the swimmer the singer by the dancer sings	by the swimmers the singers by the dancers sing

Table 1: Templates used in training set generation. Det = determiner, N = noun, PP = prepositional phrase, V = verb.

erties in controlled ways. In our case, to modulate the level of variability in subject-verb pairings, we sample pairings from Zipfian distributions (defined by the equation below) that vary the parameter α; note that α is a free parameter while K is a normalizing constant whose value is fully determined by the need for the set of $f(r)$ values to sum to 1:

$$f(r) = \frac{K}{r^{\alpha}} \tag{1}$$

Our α values ranged from 0 to 3, with lower α values producing highly variable pairings and higher α values producing predictable pairings, and we included an $\alpha \to \infty$ scenario in which each subject was seen paired with only one verb. After training, we evaluated the model's ability to generalize subject-verb agreement beyond its training data.

This highly simplified setup creates a proof-of-concept test to determine whether there exist situations in which collocational bootstrapping would be an effective strategy for a statistical learner. Specifically, there is no guarantee that collocational bootstrapping can ever succeed because it may be that all training sets are either too predictable to support abstract generalization or too variable to provide a clear statistical signal (see Section 3). By modulating α, we investigate multiple levels of variability to see if there exist levels that resolve this tension between predictability and variability, such that conditions exist under which collocational bootstrapping can succeed.

4.1 Data

We created synthetic datasets containing 12,000 unique grammatically-correct sentences for every α value tested. Each sentence contained a subject (made of a determiner and a noun) and an intransitive verb, and it could optionally include a prepositional phrase before and/or after the subject. This setup produced four sentence templates (Table 1). Within each sentence, all nouns had the same number: Either all were singular, or all were plural. This meant that the training set was always ambiguous between AGREE-SUBJECT and AGREE-RECENT as well as many other potential rules (e.g., AGREE-FIRST, in which a verb agrees with the first noun in the sentence). We enforced this ambiguity to isolate the hypothesized effect of collocational bootstrapping: Are co-occurrence patterns sufficient to disambiguate candidate agreement rules even in the absence of sentences that would directly disambiguate these rules? See Section 6 for discussion of future directions that relax this strict ambiguity.

Our vocabulary included 40 nouns and 40 present-tense verbs (each of which could be singular or plural, creating a total of 80 nouns and 80 verbs). For reasons described at the end of this section, each noun in the vocabulary is derived from one of the verbs, creating noun/verb pairs such as *orator*/*orates*. We chose *the* as the only determiner, and we used *by* and *near* as our prepositions.

To explain how sentences were sampled, we first give each noun stem and verb stem a numerical index from 0 to 39. To generate a sentence, first the verb was sampled uniformly from among the 80 options; call its index i. A subject was then sampled from a truncated Zipfian distribution with parameter α that assigns the following unnormalized probability to each of the 40 nouns that can agree with verb i, denoting a given noun's index as j: $1/((j-i \mod 40)+1)^{\alpha}$ if $j-i \mod 40 < 30$, else 0. That is, for each verb, there are 10 nouns that are withheld from appearing as the verb's subject so that we can evaluate how the models generalize to novel subject-verb pairs. If the sentence contained prepositional phrases, the prepositional object nouns were sampled uniformly from among the nouns with indices i to $i + 29 \mod 40$ that have the same number as the subject. One effect of this setup is that the same 10 nouns that were withheld from appearing as the verb's subject were also withheld from appearing as prepositional objects for purposes of evaluation.

By controlling α, we can adjust the level of variability in the training data with respect to which subjects appeared with which verbs. When α = 0, the subjects were distributed uniformly, creating

Condition	Example Minimal Pair
SEEN, MATCH	Near the fisher the bridger near the miner [listens / listen]
UNSEEN, MATCH	Near the crafter the twirler near the singer [lassos / lasso]
SEEN, MISMATCH	By the protectors the trader by the teachers [trades / trade]
UNSEEN, MISMATCH	By the climbers the driver by the writers [hunts / hunt]

Table 2: Example minimal pairs for a moderately variable condition (α = 1.5). Underlining indicates that the noun has not been seen in training data paired with this verb.

a dataset with the highest level of variability. At the other extreme, when $\alpha \to \infty$, each verb (e.g., *orate*) had only one noun that ever appeared as its subject—specifically, the noun that was morphologically related to it (e.g., *orator*). For in-between values of α, the nouns available for verb pairings were spread across a truncated Zipfian distribution as defined above, with the most likely subject being the one that is morphologically related to the verb, and other nouns having probabilities that decrease following Equation 1. We used Zipfian distributions because many linguistic units have been empirically shown to follow them, including in child-directed language (Lavi-Rotbain and Arnon, 2023); see Section 5.3 for evidence that subject-verb pairings follow Zipfian distributions in child-directed language. Figure 1 shows the distribution of nouns at each selected α value, and sample training sentences can be found in Appendix A.

Note that our models do not have access to the spellings of words; they are presented with words as atomic tokens. Thus, they cannot leverage the morphological cues of noun inflection, verb inflection, and the relatedness of certain nouns and verbs (e.g., *painter* and *paint*)—these morphological properties are included in the dataset to make the sentences easier for humans to reason about, but these properties play no role in the models' learning. Additionally, our dataset makes many simplifying assumptions compared to natural language use; see Section 6 for discussion.

4.2 Models

We trained and evaluated 2-layer decoder-only Transformer language models (Vaswani et al., 2017) in the style of GPT-2 (Radford et al., 2019), adapted from the nanoGPT implementation (Karpathy, 2023), which enables lightweight, research-oriented versions of GPT-2 to be trained from scratch. Our models used two transformer layers, each with four attention heads, an embedding size of 256, and approximately 1.6 million parameters. All code was developed using PyTorch.

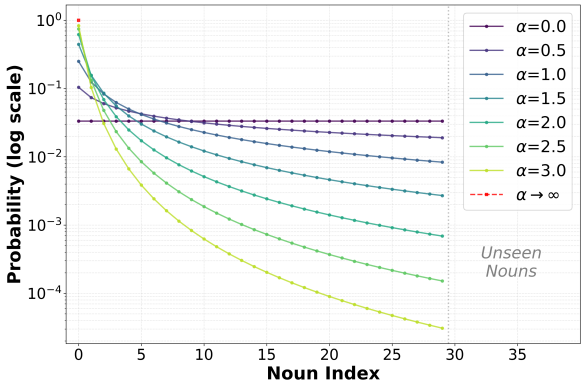

Figure 1: Noun probability distributions across α values (log scale). Lower α values produce flatter distributions with more uniform noun usage, while higher α values concentrate probability on fewer nouns. The distributions were truncated at the dotted line to leave some nouns unseen as the subjects of particular verbs.

4.3 Training

For each α value from 0.0 to 3.0 inclusive in increments of 0.1, as well as the case where $\alpha \to \infty$, we did 10 training runs with different random weight initializations. For each run, we generated a new set of 12,000 unique sentences and used a split of 80% train / 10% validation / 10% test. We used AdamW (Kingma and Ba, 2015; Loshchilov and Hutter, 2019) with a learning rate of 0.0006 which remained fixed throughout training. The batch size was 32, and each training run used 300 batches per epoch for 4 epochs (1,200 total iterations). The validation loss was computed every 300 steps, and the model version with the best validation loss was saved. Training and validation losses tracked each other closely (see Figure 6 in the Appendix), indicating that overfitting was not a concern.

4.4 Evaluation and Results

To evaluate model performance, we generated four sets of 1,000 minimal pair sentences (Marvin and Linzen, 2018), each targeting a different testing condition. In each pair, the first sentence was grammatical, and the second was ungrammatical due to

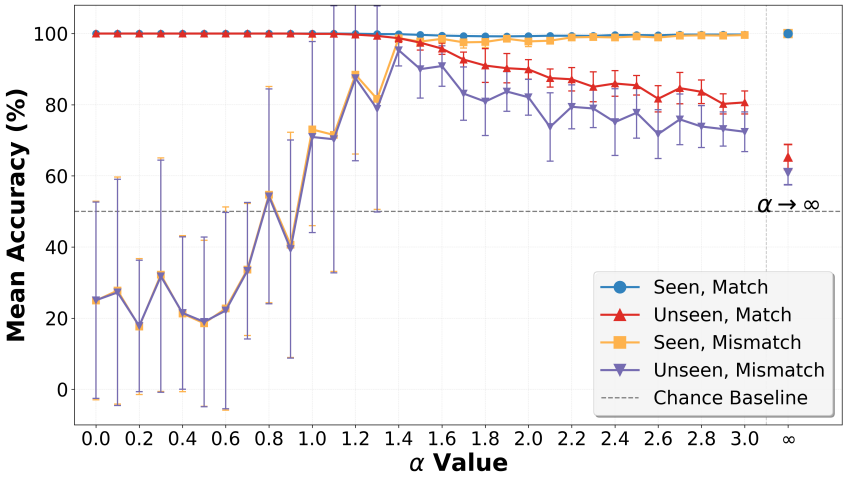

Figure 2: Model accuracy vs. Zipfian parameter α across four evaluation conditions. There is an optimal point where $\alpha = 1.4$ at which models perform robustly in all test conditions. Error bars show one standard deviation.

the verb not matching the subject's number. Sample minimal pairs are in Table 2.

We assessed each model's preferences by calculating the log probability it assigned to each sentence in a pair. For each pair, we considered the model to be correct if it assigned a higher log probability to the grammatical sentence than to the ungrammatical one, and we then computed the overall accuracy across the 1,000 pairs in each set.

The four sets varied in difficulty according to whether the subject-verb pairings and prepositional objects had appeared in the model's training data (SEEN) or not (UNSEEN), and whether the grammatical number of prepositional object nouns matched that of the subject (MATCH) or not (MISMATCH). The number mismatches served as attractors to assess whether the model had learned AGREE-SUBJECT (which would identify the correct noun for the verb to agree with) or an incorrect strategy such as AGREE-RECENT or AGREE-FIRST (both of which would select the incorrect verb inflection). All minimal pairs shared a uniform syntactic structure, [PP Det N PP V], presenting the model with three competing nouns as possible agreement targets for the verb. Below, we define these four conditions in detail and present results for each.

SEEN, MATCH: The sentences in this condition used subject-verb pairings the model encountered during training, with prepositional objects matching the subject's number. This condition presented the lowest difficulty for the models, serving primarily to verify that the models had successfully learned the patterns present in the training data. Across all α values, the models achieved 100% or near 100% accuracy as shown by the blue line in Figure 2.

UNSEEN, MATCH: Here, we introduce a source of lexical difficulty. For a given verb, the subject and prepositional objects in the test sentences were ones that had never appeared in the same sentence as that verb during training. As in the previous condition, the prepositional objects matched the subject's number. Success in this condition requires the model to generalize across words of the same number—that is, to use distributional commonalities to form the classes of *singular nouns* and *plural nouns*, and to recognize that the same verb form applies to any member of that class. This type of generalization should be easiest when α is low, meaning that all the singular nouns have similar distributions and are therefore easier for the model to group together into a cohesive class, and similarly for the plurals. As expected, accuracy was high for low α values but began to decline when $\alpha \approx 1.4$ (see the red line in Figure 2). At high α values, the models appear to learn strong associations between specific subjects and verbs, preventing them from generalizing well to unseen ones.

SEEN, MISMATCH: This condition presents a different type of difficulty: conflicting cues about number agreement. If the model has incorrectly learned that the verb should agree either with the closest noun or with the first noun in the sentence—both of which would succeed for all sentences in the training set—it will now fail when tested with

sentences in which the prepositional objects have a different grammatical number than the subject. High α values in the training data create greater predictability in subject-verb pairings, which we hypothesize would help models select AGREE-SUBJECT over other competing rules by making it easier to recognize which syntactic positions host the noun that the verb shares a syntactic dependency with. As shown by the yellow line in Figure 2, results confirmed that the models perform poorly at low α values but with high accuracy (approaching 100%) as α increases.

UNSEEN, MISMATCH: Our final condition combines both sources of difficulty: novel subject-verb pairings and prepositional objects that have a different number from the subject (note that the prepositional object-verb pairings are also novel, as in the UNSEEN, MATCH condition). As above, we predict that low α values will prevent the model from generalizing because it will struggle to identify the correct agreement target among mismatching competitors, and that high α values will also cause the model to perform poorly, as it will struggle to generalize to unseen noun/verb pairings. What happens between the low and high α values is harder to predict. The critical question is whether there exists a "sweet spot" between extremes, where the model can handle both the UNSEEN and MISMATCH aspects of this condition. We find that there is indeed such a sweet spot (Figure 2, purple line): The model showed poor performance at low and high α values, but there is a peak with near-perfect accuracy at intermediate values.

4.5 Discussion

The results show there is an ideal level of variability in subject-verb pairings in the training data that helps the model generalize robustly. Too much variation hinders the model from inferring the correct syntactic structure. Too much predictability prevents the model from forming an abstract rule, such that it generalizes poorly to novel subject-verb pairings. Between these extremes, when $\alpha \approx 1.4$, there is an optimal level of variability that supports robust generalization. This pattern demonstrates two key points. First, the fact that model performance varies with the level of variability indicates that these neural networks indeed use co-occurrence statistics to inform the learning of subject-verb agreement in the ways expected under the collocational bootstrapping hypothesis. Second, the existence of an optimal level at which we get robust generalization shows that, under the right conditions, collocational bootstrapping can be a viable learning strategy.

5 Experiment 2: Analysis of CHILDES

In the previous experiment, we observed that when $\alpha \approx 1.4$, the synthetic training data contained a level of variability that optimizes generalization. We now investigate whether a similar statistical signal is present in real-world data such that children could potentially leverage this signal to assist in learning subject-verb agreement. Toward this end, we consider the frequency of subject-verb pairings in a corpus of child-directed language.

5.1 Data

CHILDES, or the **Chi**ld **L**anguage **D**ata **E**xchange **S**ystem, is an open repository of transcripts and other supporting media containing conversations between children and their caretakers, which have been compiled and donated by researchers (MacWhinney, 2000). For this experiment, we used data from CHILDES participants tagged as English-language speakers. Because our goal is to understand the linguistic input that a child might receive (and not the utterances that the child produces), we filtered the speakers to include only those speaking to children but not the children themselves.

We extracted all utterances spoken by an adult to children with ages from 0 to 96 months, resulting in a set of 4,739,189 utterances; see Appendix E for more details about data filtering and cleaning.

We parsed the utterances using the spaCy dependency parser (Honnibal et al., 2020). We extracted all pairings of a subject noun and the corresponding verb (the pairings characterized by the dependency type *nsubj*), using lemmas for both subjects and verbs. This produced 2,802,071 subject-verb pairs.

5.2 Zipfian Analysis

We analyzed the subjects of the 100 most frequent verbs. We restricted ourselves to frequent verbs so that there would be sufficient data to achieve quantitatively meaningful results; all verbs in this set appeared at least 2,396 times. For each verb, we created a list of the subjects that co-occur with that verb, ranked by the number of times that the subject-verb pairing appeared. Next, we converted these subject counts to proportions and calculated the average proportion of the subjects at each rank across all verbs. That is, for each rank r

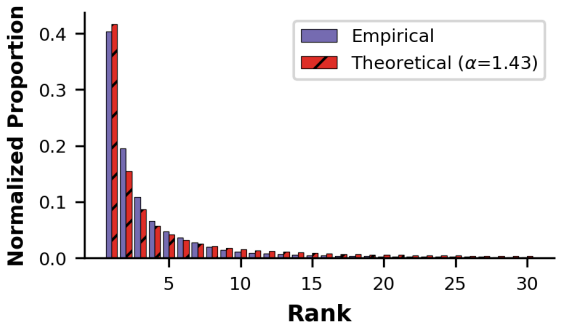

Figure 3: The empirical distribution of subject-verb pairings in CHILDES (averaged across verbs in accordance with Equation 2), along with the frequencies predicted by a Zipfian distribution with parameter $\alpha = 1.43$ (α was chosen by finding the best fit to the data).

we computed $f_{\text{empirical}}(r)$—the average frequency of a verb's r^{th} most common subject—as follows, where verb_k is the kth most common verb, and $\text{subj}_{r,k}$ is the noun that occurs as the rth most common subject for verb_k:

$$f_{\text{empirical}}(r) = \frac{1}{100} \sum_{k=1}^{100} \frac{\text{count}(\text{subj}_{r,k}, \text{verb}_k)}{\text{count}(\text{verb}_k)} \tag{2}$$

This formula gives the empirical frequencies of verb-subject pairings, which we then sought to fit to the theoretical predictions of Zipf's Law:

$$f_{\text{theoretical}}(r, \alpha) = \frac{K}{r^{\alpha}} \tag{3}$$

Zipf's Law has one free parameter α (note that K is a normalizing constant, so it is fully determined by α), so fitting $f_{\text{theoretical}}$ to $f_{\text{empirical}}$ amounted to finding the value of α that best fit the observed data. To do so, we tried all values of α ranging from 0 to 3.0 in increments of 0.01. For each α value, we computed the mean squared error (MSE) between $f_{\text{empirical}}$ and $f_{\text{theoretical}}$, defined as:

$$\text{MSE}(\alpha) = \frac{1}{R} \sum_{r=1}^{R} \left(f_{\text{empirical}}(r) - f_{\text{theoretical}}(r, \alpha) \right)^2 \tag{4}$$

where R is the number of ranks over which we computed the error. We then selected the α value that minimized $\text{MSE}(\alpha)$.

5.3 Results

We found that the best-fitting value of α was α=1.43. See Figure 3 for a comparison between $f_{\text{empirical}}$ and $f_{\text{theoretical}}$ with this α value. In addition to the dataset-wide fitting described above, we also broke down the analysis by the age of the child being spoken to in order to see whether the best-fitting α value varied by the age of the target child. As shown in Figure 4, the Zipfian parameter α generally decreases as the target child's age increases. Sample utterances by age group are in Appendix C.

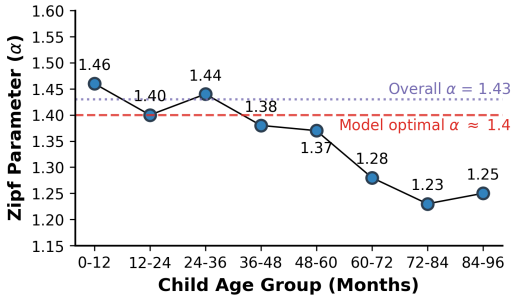

Figure 4: The fitted Zipf parameter α decreases with child age. The dashed red line indicates the optimal α found in neural network simulations; the dotted purple line indicates the overall corpus α.

Strikingly, both the α value calculated for all utterances (α=1.43) and the range of α values found for each age group (α=1.46 to 1.23) are close to the value of α where our model generalized best, $\alpha \approx 1.4$. This finding suggests that naturalistic English input has a level of subject-verb variability that facilitates the acquisition of agreement.

6 Discussion

We have used neural network language models to show that it is possible for statistical learners to robustly generalize English subject-verb agreement by using collocational bootstrapping. This bootstrapping strategy only succeeds under certain statistical conditions (when the α parameter in Zipf's law is about 1.4); we have further found preliminary evidence that child-directed speech has the right properties for this strategy to be viable.

Making inferences about child language acquisition: Due to the many differences between our synthetic text and natural child-directed language, we do not intend to draw strong conclusions about the similarity between the model-optimal α value ($\approx$ 1.4) and the empirical α found in CHILDES (1.43). It is worth noting that the type of simulations we conducted, whether done with fully synthetic data or data closer to child-directed language, can only provide evidence about which learning strategies could be effective, not whether children actually use those strategies during acquisition.

Toward greater naturalness: Our synthetic data sets were highly simplified, differing from naturalistic language in important ways. First, our existing data sets likely over-represent the presence of prepositional phrases before the verb. Second, our training sets were fully ambiguous between AGREE-SUBJECT and AGREE-RECENT whereas naturalistic data contain some disambiguating examples—though naturalistic data can also contain agreement attraction errors (Bock and Miller, 1991) that point toward AGREE-RECENT rather than AGREE-SUBJECT. Third, naturalistic data involve a much larger vocabulary than what we used here. Fourth, naturalistic English sentences often use verbs (e.g., past-tense verbs) that are not explicitly inflected for number.

Beyond these differences in the word sequences encountered by our models vs. human children, our models also differ from human learners in only having access to text, whereas children receive multimodal input that might support types of bootstrapping not available to our models. Children have access to prosody, which might provide syntactic cues through prosodic bootstrapping, and can also draw on real-world context, which might provide meaning that can serve as a cue to syntax, as suggested under semantic bootstrapping. Future work could explore the effects of modifying the training set in ways that overcome these qualitative gaps.

The difficulty of agreement acquisition: Our analysis of CHILDES found that its statistical properties make it well-suited for collocational bootstrapping. However, prior work has found that both children (Nozari and Omaki, 2022) and neural networks trained on child-directed language (Huebner et al., 2021; Padovani et al., 2025) make agreement attraction errors, meaning that they have not learned subject-verb agreement as robustly as might be expected from our analysis. A likely explanation for the discrepancy is that some of the other factors mentioned in the previous paragraph could counteract the favorable α value that we have observed in ways that add further difficulty to the acquisition task. A goal to ultimately work toward is investigating which types of input data and learning strategies can reproduce both the successes and failures of subject-verb agreement in humans.

Statistical co-occurrence or semantic relatedness? Semantic bootstrapping leverages the meaning of words as a cue to syntax. Since semantically related words often occur near each other, there may be overlap between semantic and collocational cues. Indeed, past computational work has found a relationship between a word's meaning and its statistical distribution in a corpus (Landauer and Dumais, 1997; Mikolov et al., 2013). Future work could tease apart the respective roles of semantics and statistics as cues to syntactic dependencies.

Semantic bootstrapping is typically framed as a mechanism for inferring syntactic categories such as parts of speech. Collocational bootstrapping is instead a strategy for acquiring word-word dependencies, under which learners can bootstrap from one type of word-word relatedness (co-occurrence) to another (syntactic dependencies). Since semantics and distribution overlap, this same broad strategy could instead use semantic relatedness rather than distributional co-occurrence as a cue to syntactic dependencies, providing a way to extend semantic bootstrapping to the learning of dependencies.

Extending collocational bootstrapping: Another direction for future work is extending collocational bootstrapping by analyzing whether it is an effective strategy for learning other syntactic dependencies beyond the one studied here (subject-verb dependencies). A natural first step would be to investigate other agreement phenomena in English and other languages, such as noun-anaphor number agreement and adjective-noun gender agreement.

7 Conclusion

We have proposed collocational bootstrapping as a potential mechanism by which word co-occurrence statistics can support the learning of syntax. We have tested our hypothesis using neural language models, training and evaluating them on synthetic data with varying levels of variability in subject-verb pairings. We have found that there is an optimal level of variability, specifically a Zipfian distribution with $\alpha \approx 1.4$, that maximizes the model's ability to generalize. Too little variability prevents the model from generalizing to novel noun-verb pairs, and too much variability prevents it from abstracting syntactic rules. The α value at which there is a sweet spot for optimal generalization is consistent with the level of variability observed in child-directed speech (α=1.43), suggesting that the statistical structure of natural language could guide learners in correctly acquiring syntax. These results provide one illustration of how statistical properties of linguistic data can facilitate the learning of abstract syntactic phenomena.

Limitations

Our neural network experiments involve simplified, synthetic training data that differ from children's input in qualitative ways, and we have only analyzed the effect of one statistical cue on one linguistic phenomenon; see Section 6 for discussion.

Acknowledgments

We extend our thanks to the anonymous reviewers and the feedback they provided on this paper, and to Jason Hobbs for his technical insight and support. We used Claude Code for assistance with coding, and we checked all AI-generated code. We also used Grammarly, and Claude Opus 4.6 and Sonnet 4.5, for feedback on style and grammar, but all ideas in the paper were ours. Any errors are our own.

References

Omri Abend, Tom Kwiatkowski, Nathaniel J. Smith, Sharon Goldwater, and Mark Steedman. 2017. Bootstrapping language acquisition. *Cognition*, 164:116–143.

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

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

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Jeffrey L. Elman. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7(2):195–225.

Steven Finch and Nick Chater. 1992. Bootstrapping syntactic categories. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 14.

Robert Frank, Donald Mathis, and William Badecker. 2013. The acquisition of anaphora by simple recurrent networks. *Language Acquisition*, 20(3):181–227.

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

Rebecca L. Gómez. 2002. Variability and detection of invariant structure. *Psychological Science*, 13(5):431–436.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1195–1205.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Philip A. Huebner, Elior Sulem, Cynthia Fisher, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? On the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics.

Andrej Karpathy. 2023. nanoGPT. GitHub. Computer software.

Vera Kempe, Patricia J. Brooks, and Steven Gillis. 2024. Four decades of open language science: The CHILDES project. *Language Teaching Research Quarterly*, 44:15–30.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. 2018. LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211.

Ori Lavi-Rotbain and Inbal Arnon. 2023. Zipfian distributions in child-directed speech. *Open Mind: Discoveries in Cognitive Science*, 7:1–30.

Cara Su-Yi Leong and Tal Linzen. 2026. Manipulating language models' training data to study syntactic constraint learning: The case of English passivization. *Journal of Memory and Language*, 149:104751.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum, Mahwah, NJ.

Michael Maratsos and Mary Anne Chalkley. 1980. The internal language of children's syntax: The ontogenesis and representation of syntactic categories. *Children's Language*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. 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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

Toben H. Mintz. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91–117.

Kanishka Misra and Najoung Kim. 2024. Generating novel experimental hypotheses from language models: A case study on cross-dative generalization. *arXiv preprint arXiv:2408.05086*.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929.

James L. Morgan and Katherine Demuth. 1996. *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Psychology Press.

Aaron Mueller, Robert Frank, Tal Linzen, Luheng Wang, and Sebastian Schuster. 2022. Coloring the blank slate: Pre-training imparts a hierarchical inductive bias to sequence-to-sequence models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1352–1368.

Karl Mulligan, Robert Frank, and Tal Linzen. 2021. Structure here, bias there: Hierarchical generalization by jointly learning syntactic transformations. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 125–135.

Nazbanou Nozari and Akira Omaki. 2022. Revisiting agreement: Do children and adults compute subject-verb agreement differently? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Luca Onnis, Padraic Monaghan, Morten H. Christiansen, and Nick Chater. 2004. Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in nonadjacent dependencies. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26.

Francesca Padovani, Jaap Jumelet, Yevgen Matusevych, and Arianna Bisazza. 2025. Child-directed language does not consistently boost syntax learning in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, page 19746–19767. Association for Computational Linguistics.

Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, and Shane Steinert-Threlkeld. 2024. Filtered corpus training (FiCT) shows that language models can generalize from indirect evidence. *Transactions of the Association for Computational Linguistics*, 12:1597–1615.

Lisa S. Pearl and Benjamin Mis. 2016. The role of indirect positive evidence in syntactic acquisition: A look at anaphoric one. *Language*, 92(1):1–30.

Steven Pinker. 1984. *Language Learnability and Language Development*. Harvard University Press.

Geoffrey K. Pullum and Barbara C. Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 19(1-2):9–50.

Tian Qin, Naomi Saphra, and David Alvarez-Melis. 2025. Data drives unstable hierarchical generalization in LMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11722–11740, Suzhou, China. Association for Computational Linguistics.

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

Limor Raviv, Gary Lupyan, and Shawn C. Green. 2022. How variability shapes learning and generalization. *Trends in Cognitive Sciences*, 26(6):462–483.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948.

Kenneth Wexler and Peter W. Culicover. 1980. *Formal Principles of Language Acquisition*. MIT Press.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2024. Using computational models to test syntactic learnability. *Linguistic Inquiry*, 55(4):805–848.

Elizabeth Wonnacott, Elissa L. Newport, and Michael K. Tanenhaus. 2008. Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, 56(3):165–209.

Charles Yang. 2016. *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. MIT press.

Xiulin Yang, Arianna Bisazza, Nathan Schneider, and Ethan Gotlieb Wilcox. 2026a. A unified assessment of the poverty of the stimulus argument for neural language models. *Preprint*, arXiv:2602.09992.

Xiulin Yang, Heidi Getz, and Ethan Gotlieb Wilcox. 2026b. From linear input to hierarchical structure: Function words as statistical cues for language learning. *Preprint*, arXiv:2601.21191.

Aditya Yedetore and Najoung Kim. 2024. Semantic training signals promote hierarchical syntactic generalization in transformers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4059–4073, Miami, Florida, USA. Association for Computational Linguistics.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, Cambridge, MA.

A Sample training sentences

Below are examples of sentences used in training for a few levels of variability.

Maximally Variable ($\alpha = 0$):

- the driver leads
- by the solver the challenger trades
- the dancers near the writers embezzle
- by the twirlers the painters near the singers navigate

Moderately Variable ($\alpha = 1.5$):

- the hunters listen
- by the builder the twirler collapses
- by the swimmers the bridgers bridge
- near the miner the jumper near the painter jumps

No Variability ($\alpha \to \infty$):

- the twirler twirls
- by the charmers the miners mine
- the builders near the protectors build
- near the swimmers the lassoers by the bakers lasso

B Speaker role utterance counts

See Figure 5 for utterance counts by each type of speaker in the CHILDES data that we analyzed.

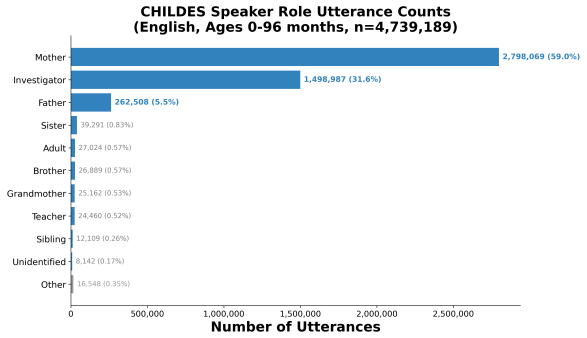

Figure 5: Distribution of utterances by speaker role in the English subset of CHILDES (ages 0-96 months).

C Examples of child-directed language at varying child ages

Below are examples of child-directed language spoken to children of varying ages.

Age 0-12 months ($\alpha = 1.46$)

1. "you put the block on"
2. "what else do we see in here"
3. "oh be be very gentle with baby right"

Age 12-24 months ($\alpha = 1.40$)

1. "what's that"
2. "yeah that's where we were"
3. "you don't like MacDonald's and I don't like MacDonald's"

Age 24-36 months ($\alpha = 1.44$)

1. "how do you know this is a duck"
2. "this is velcro"
3. "let's sit here on mama's mama's knee"

Age 36-48 months ($\alpha = 1.38$)

1. "you get milk from it"
2. "look at these"
3. "want mommy to read"

Age 48-60 months ($\alpha = 1.37$)

1. "i think we found the wheels or your mom did"
2. "just like we see up there remember"
3. "there's something wrong with her teeth aren't there"

Age 60-72 months ($\alpha = 1.28$)

1. "well I know but you know what I think this chair is"
2. "so you want listen come here I'm going to tell you"
3. "I don't think I would like those"

Age 70-84 months ($\alpha = 1.23$)

1. “I never heard of that one before”
2. “dad’s gonna dads can do it a lot”
3. “there’s how many bears on one wheel’

Age 84-96 months ($\alpha = 1.25$)

1. “I got you a pencil”
2. “alright well if you don’t put it on then the letter’s no good”
3. “uh what about a movie though”

D Training loss

See Figure 6 for the loss trajectories of the models we trained.

E Data Cleaning

We downloaded 5,147,586 utterances from participants categorized as English-language speakers, restricted to 25 target speaker roles: Adult, Caretaker, Father, Friend, Grandfather, Grandmother, Investigator, Mother, Narrator, Playmate, Relative, Sibling, Sister, Brother, Teacher, Unidentified, Visitor, Teenager, Participant, Girl, Male, Student, Environment, Doctor, Target Adult. Next, we removed rows with null text content, converted utterances to text strings, removed rows lacking a target child age, and removed rows where the target child age was greater than 96 months. After cleaning, 4,739,189 utterances remained. Of these, 59.0% were spoken by mothers and 31.6% by investigators, together comprising nearly 90% of all utterances as shown in Figure 5 in Appendix B. This proportion reflects the high concentration of speech from caregivers in the corpus (Kempe et al., 2024). During the subject-verb extraction step, subject and verb lemmas were converted to lower case, and the resulting verb counts were restricted to ASCII English forms before selecting the top 100 verbs for analysis.

F Analysis of subject-verb pairings by age

See Table 3 for statistics of subject-verb pairings in child-directed language broken down by the age of the child being spoken to.

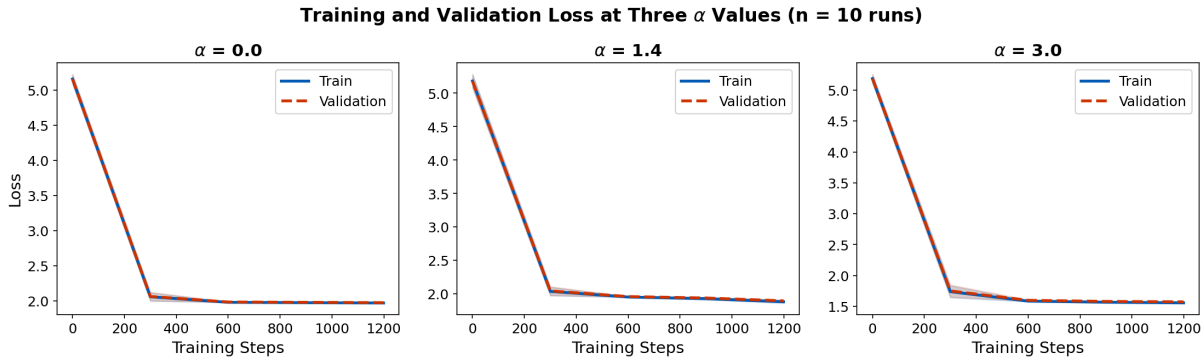

Figure 6: Training and validation loss at three α values. Loss curves track closely across all conditions, indicating no overfitting.

Age Group	Utterances	S-V Pairs	Unique Verbs	Unique Subjects	α	MSE
0–12mo	182,023	113,607	1,180	1,516	1.46	9.78e-07
12–24mo	671,559	350,859	3,580	6,135	1.40	2.56e-07
24–36mo	1,900,684	1,163,974	6,775	13,577	1.44	2.15e-07
36–48mo	923,858	553,675	4,940	9,817	1.38	3.90e-07
48–60mo	621,805	399,355	4,262	8,022	1.37	5.26e-07
60–72mo	237,449	121,409	2,661	4,805	1.28	2.33e-07
72–84mo	111,117	52,385	1,778	3,092	1.23	7.79e-07
84–96mo	90,694	46,807	1,467	2,491	1.25	4.87e-07

Table 3: Age-stratified analysis of subject-verb pairings in CHILDES (0–96 months). The Zipf parameter α decreases from 1.46 in the youngest age group to 1.25 in the oldest.