A systematic investigation of learnability from single child linguistic input

Yulu Qin[1] **(yq810@nyu.edu)**
Wentao Wang[1] **(ww2135@nyu.edu)**
Brenden M. Lake[1,2] **(brenden@nyu.edu)**
[1]Center for Data Science, [2]Department of Psychology, New York University

Abstract

Language models (LMs) have demonstrated remarkable proficiency in generating linguistically coherent text, sparking discussions about their relevance to understanding human language learnability. However, a significant gap exists between the training data for these models and the linguistic input a child receives. LMs are typically trained on data that is orders of magnitude larger and fundamentally different from child-directed speech (Warstadt & Bowman, 2022; Warstadt et al., 2023; Frank, 2023a). Addressing this discrepancy, our research focuses on training LMs on subsets of a single child's linguistic input. Previously, Wang, Vong, Kim, and Lake (2023) found that LMs trained in this setting can form syntactic and semantic word clusters and develop sensitivity to certain linguistic phenomena, but they only considered LSTMs and simpler neural networks trained from just one single-child dataset. Here, to examine the robustness of learnability from single-child input, we systematically train six different model architectures on five datasets (3 single-child and 2 baselines). We find that the models trained on single-child datasets showed consistent results that matched with previous work, underscoring the robustness of forming meaningful syntactic and semantic representations from a subset of a child's linguistic input.

Keywords: learnability; single-child; distributional learning; robustness; language models

Introduction

Young children are remarkably efficient language learners, yet the mechanisms behind language acquisition remain a scientific puzzle. Meanwhile, important advances in language models (LMs) for natural language processing provide us with new, powerful computational tools to investigate fundamental questions regarding language acquisition and its relationship with human cognition (Warstadt & Bowman, 2022; Frank, 2023b). Trained on trillions of written words, contemporary Transformer-based Large Language Models (LLMs) can produce coherent text with a proficiency that far exceeds the predictions of experts in the field from a decade ago (Chang & Bergen, 2023), raising important questions about the degree to which strong inductive biases and language-specific mechanisms are needed to acquire language beyond more general distributional learning mechanisms (Landauer, Foltz, & Laham, 1998; Elman, 1990). To improve the relevance of language models as cognitive models of human language acquisition, previous efforts trained models on aggregated linguistic input across multiple children (Warstadt et al., 2023; Huebner, Sulem, Cynthia, & Roth, 2021). As in several works (Wang et al., 2023; Vong, Wang, Orhan, & Lake, 2024; Abend, Kwiatkowski, Smith, Goldwater, & Steedman, 2017; Waterfall, Sandbank, Onnis, & Edelman, 2010), we train models on subsets of the linguistic input that just a single child was exposed to. Children must learn language from only their own input—they cannot share and aggregate input with others—and thus this is the setting we focus on here.

Here, we use a recent article by Wang et al. (2023) as a launchpad for our new learnability studies based on a single child's input. Wang et al. (2023) applied two neural language models, Continuous Bag-of-Words (CBOW; Mikolov, Chen, Corrado, & Dean, 2013) and Long Short-Term Memory (LSTM; Hochreiter & Schmidhuber, 1997), to the SAYCam-S dataset, a longitudinal collection of transcribed linguistic inputs to a single child aged 6 to 25 months (Sullivan, Mei, Perfors, Wojcik, & Frank, 2021). Wang et al. (2023)'s study revealed that these models successfully recovered lexical classes that reflect key syntactic and semantic distinctions, including nouns, verbs, animals, body parts, etc., from the process of learning to predict the next word in transcribed child-directed utterances. Additionally, they employed the Zorro test suite to evaluate the models' grammatical knowledge through acceptability judgments (Huebner et al., 2021). However, these promising findings are based on two model architectures trained on only one single child's data, thus limiting the generalizability of their results. Our research builds upon this groundwork by investigating the robustness of Wang et al. (2023)'s learnability results from one child's input across different settings, including multiple datasets and different model architectures, to see which combinations of datasets and architectures can produce successful learners.

Specifically, in this study, we examined 6 model architectures (3 model classes and 2 sizes each) trained on 5 datasets: 3 datasets representing input to individual children and 2 others representing meaningful baselines for comparison. Each combination of architecture and dataset was analyzed through linguistic acceptability tests, visualizations of word embeddings, and cloze tests. Across each of these settings, we find that the results are robust and similar to Wang et al. (2023)'s.

Methods

Datasets

We explored 5 datasets, three that capture child-directed speech at the level of a single child, one aggregating child-directed speech from multiple children, and one with an equivalent amount of text from the web.

SAYCam-S, Sarah and Ellie. These are three different single-child datasets in our experiments. SAYCam-S is the single child dataset used in Wang et al. (2023). The other two child-directed datasets are two sets of transcribed speech from CHILDES (MacWhinney, 2000), each directed to one individual child: Sarah (age ranging from 2;3 to 5;1) from the

Table 1: **Dataset Statistics.** SAYCam-S, Sarah, and Ellie are three single-child datasets. Note that all datasets except CHILDES have a similar number of training tokens.

		SAYCam-S	**Sarah**	**Ellie**	Wikipedia	CHILDES
Training	Number of utterances	26,322	32,965	38,140	10,504	1,151,816
	Mean (SD) utterance length	8.06 (5.46)	6.71 (3.32)	6.29 (3.14)	24.81 (14.60)	7.09 (4.19)
	Number of tokens	212,064	221,211	239,807	260,580	8,163,820
	Out-of-vocabulary rate	1.85%	1.26%	1.74%	9.69%	0.26%
	Vocabulary size	2350	2333	2780	8833	15,762
Validation	Number of utterances	1462	1786	2269	588	64,254
	Mean (SD) utterance length	7.95 (5.46)	6.79 (3.50)	6.03 (3.00)	25.50 (14.63)	7.16 (4.09)
	Number of tokens	11,621	12,119	13,676	14,995	459,787
	Out-of-vocabulary rate	2.21%	2.24%	3.58%	12.04%	0.50%

Brown corpus (Brown, 1973) and Ellie (age ranging from 0;9 to 5) from the Sakali corpus (Beaupoil-Hourdel, 2015). These two datasets, respectively sourced from the North American English and the British English sections of the CHILDES database, capture longitudinal recordings in naturalistic contexts. As shown in Table 1, these three datasets present similar statistics in terms of vocabulary size, length of utterances and number of tokens.

Wikipedia. As a comparison, we also have a randomly sampled Wikipedia dataset with a parallel amount of text tokens to Ellie, the child dataset that contains the most tokens. (After filtering sentences with fewer than 2 words, as discussed below in Data Preprocessing, the final token counts varied slightly.) Notably, with its longer average utterance length and more complex content, this Wikipedia set has fewer sentences but a larger vocabulary than the aforementioned child-directed datasets. Detailed statistics can be found in Table 1.

CHILDES. Finally, as a reference, we incorporated the North American Portion of the CHILDES corpus. It contains aggregated child-directed data with a nearly 6× larger vocabulary and approximately 30× more tokens than the single child datasets. See the detailed statistics in Table 1.

Data Preprocessing

Built on top of Yedetore, Linzen, Frank, and McCoy (2023)'s data preprocessing procedure, we excluded children's own utterances to replicate data as similar as possible to the sentences children receive and replaced tokens that appear fewer than 3 times with an `<unk>` token. We split approximately 90% of each dataset to training, 5% to validation, and 5% to testing. We also filter out sentences that contain fewer than 2 words during training and validation. Details of dataset statistics for training and validation can be seen in Table 1.

Model Architectures and Training

Wang et al. (2023) investigated n-gram models, CBOWs and LSTMs. Our evaluation expands to 6 different model architectures, including GPT-2-style and RoBERTa-style Transformers called BabyBERTa[1] (Radford et al., 2019; Liu et al., 2019; Huebner et al., 2021), in addition to LSTMs (Hochreiter & Schmidhuber, 1997). We test two model sizes of each model class. The comprehensive list of model architectures used is detailed in Table 2.

Table 2: **Model Architectures.** # of trainable parameters are based on the SAYCam-S dataset, with slight variation across datasets due to differences in vocabulary size.

Model	# of parameters
LSTM (1-layer)	3.3M
LSTM (2-layer)	5.4M
GPT-2 (2-layer)	7.8M
GPT-2 (8-layer)	26.7M
BabyBERTa (2-layer)	7.8M
BabyBERTa (8-layer)	26.8M

Training objectives. All models were trained from scratch. For LSTMs and GPT-2-based Transformers, the models aimed to predict the next token in a short utterance, using cross-entropy loss for training. For the BabyBERTa-based Transformer, the model was trained to predict randomly masked tokens, such that 15% of the tokens in each utterance were masked anew during each presentation.

Model configurations. We trained 2 architectures of large and small sizes for each model class, resulting in a total of 6 architectures. These include uni-directional LSTMs (1 layer and 2 layers), as well as GPT-2-based and BabyBERTa-based Transformers (2 layers and 8 layers), as listed in Table 2. Subsequently, we performed an extensive hyperparameter search. We tuned and identified the best hyperparameters based on validation perplexity for each of our five datasets. For the hyperparameter search, we standardized all model embedding and hidden sizes to 512 and all FFN intermediate sizes for Transformer-based models to 2048. We used `ReduceOnPlateau` learning rate scheduler in PyTorch, which reduces the learning rate by a factor of 10 after the validation loss plateaus for 2 consecutive epochs. We used early

[1]Prior research has shown that a scaled-down version of RoBERTa-base termed BabyBERTa, trained on child-directed data, achieves grammatical knowledge comparable to the full RoBERTa-base on the Zorro benchmark (Huebner et al., 2021). We applied their insights and will refer to our RoBERTa-based Transformer as a BabyBERTa-based Transformer in the following sections.

Table 3: **Validation Perplexity.**

Model	**SAYCam-S**	**Sarah**	**Ellie**	Wikipedia	CHILDES
LSTM (1-layer)	18.01	18.45	23.86	102.00	23.45
LSTM (2-layer)	18.47	18.40	23.59	98.70	23.74
GPT-2 (2-layer)	18.74	18.97	23.93	127.58	20.81
GPT-2 (8-layer)	18.42	18.46	23.94	130.54	20.15
BabyBERTa (2-layer)	10.41	10.96	16.24	74.38	10.39
BabyBERTa (8-layer)	**9.25**	10.67	14.94	65.10	10.35

Table 4: **Zorro Test Accuracies (%).**

Model	**SAYCam-S**	**Sarah**	**Ellie**	Wikipedia	CHILDES
LSTM (1-layer)	66.43	68.98	66.45	59.44	78.28
LSTM (2-layer)	69.18	68.25	64.59	61.64	81.49
GPT-2 (2-layer)	68.22	68.70	65.40	57.47	86.40
GPT-2 (8-layer)	65.76	**70.49**	66.45	61.88	87.83
BabyBERTa (2-layer)	69.57	70.23	66.28	59.02	84.63
BabyBERTa (8-layer)	65.45	66.42	64.46	59.54	81.65

stopping to select the checkpoint with the best validation loss. We tuned other hyper-parameters based on validation performance, including:

- **learning rate** $\in \{1\times10^{-4},\ 3\times10^{-4},\ 1\times10^{-3},\ 3\times10^{-3}\}$
- **batch size** $\in \{8, 16, 32\}$
- **weight decay** $\in \{0.01,\ 0.05,\ 0.1,\ 0.15,\ 0.24\}$
- **dropout rate** $\in \{0.01,\ 0.05,\ 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5\}$
- **number of attention heads** (for Transformer-based models) $\in \{8,\ 16,\ 32\}$

Performance for a particular configuration is averaged across 3 runs with different random seeds. As a measure of generalization, the validation perplexity score[2] is shown in Table 3.

Tokenizer

Simple word-level tokenizers were used to facilitate our analyses of the learned word embeddings (e.g., Fig. 3), constructed with Hugging Face Tokenizers for each dataset. Refer to Table 1 for the vocabulary size for each dataset.

Results

We analyze each trained model through linguistic acceptability tests for linguistic knowledge, visualizations of word embeddings for syntactic and semantic category structures, and cloze tests for noun-verb distinction within context. In each analysis, we find robust results similar to Wang et al. (2023) across all models with different configurations.

Linguistic Acceptability Tests

Following Wang et al. (2023), we tested models' sensitivity to linguistic knowledge such as subject-verb agreement on the Zorro test suite (Huebner et al., 2021). This test suite evaluates 13 grammatical phenomena on 23 tests, each containing 2000 minimal sentence pairs. To avoid out-of-vocabulary words, Wang et al. (2023) filtered out all minimal pairs containing tokens outside of their SAYCam-S vocabulary, left with 15 tests, each containing fewer than 700 pairs. In this work, we regenerated Zorro based on the original linguistic templates and the intersected vocabulary of our 5 datasets, resulting in a full 23 tests.[3]

Test accuracy. From Table 4, we can see average Zorro test accuracies over 3 different random seeds are consistent among 3 single-child datasets (Sarah, Ellie, and SAYCam-S), nearly all of which reached over 65% correct (chance is 50%). Among all single-child-directed datasets, the Sarah dataset trained models with the best Zorro accuracy in all model architectures except the LSTM (2-layer). Comparatively, across all 5 datasets studied, models trained on the Wikipedia dataset exhibit the lowest Zorro accuracy,[4] while those trained on the CHILDES dataset achieve the highest. Furthermore, for each specific linguistic test, models trained on single child datasets give consistent performances as seen in Figure 1. The first row illustrates four linguistic tests where most models trained on single-child datasets perform well, whereas the second row shows models perform poorly on subject-verb agreement.[5]

In particular, all models trained on child-directed datasets exhibit high performance on the "quantifiers–existential there" test and perform near chance levels on the "subject-verb agreement–across relative clause" test, which aligns to

[2]The perplexity is the exponentiation of the validation cross-entropy loss, defined as: $\text{perplexity} = \exp(\text{H}(X))$, $\text{H}(X) = -\frac{1}{N}\sum_{i=1}^{N}\log \text{P}(x_i)$, where H is the cross-entropy, and X is a random variable denoting a token. We used it as a more straightforward measure of model performance on next-word prediction tasks.

[3]The regenerated Zorro test suite can be found in `https://github.com/wwt17/Zorro`.

[4]As an example, models trained on the Wikipedia dataset perform the worst on the test for "argument structure dropped argument" (as shown in Figure 1, row 1, plot 3), where models are tested on sentences pair such as "the purple bear gave her./give her the purple bear." Since the Wikipedia training dataset does not contain sentences that start with the word "give", models yield a high perplexity score on this token and make incorrect judgments.

[5]A complete plot for model performances on all tests can be found in `https://github.com/yuluqinn/single-child-robustness`.

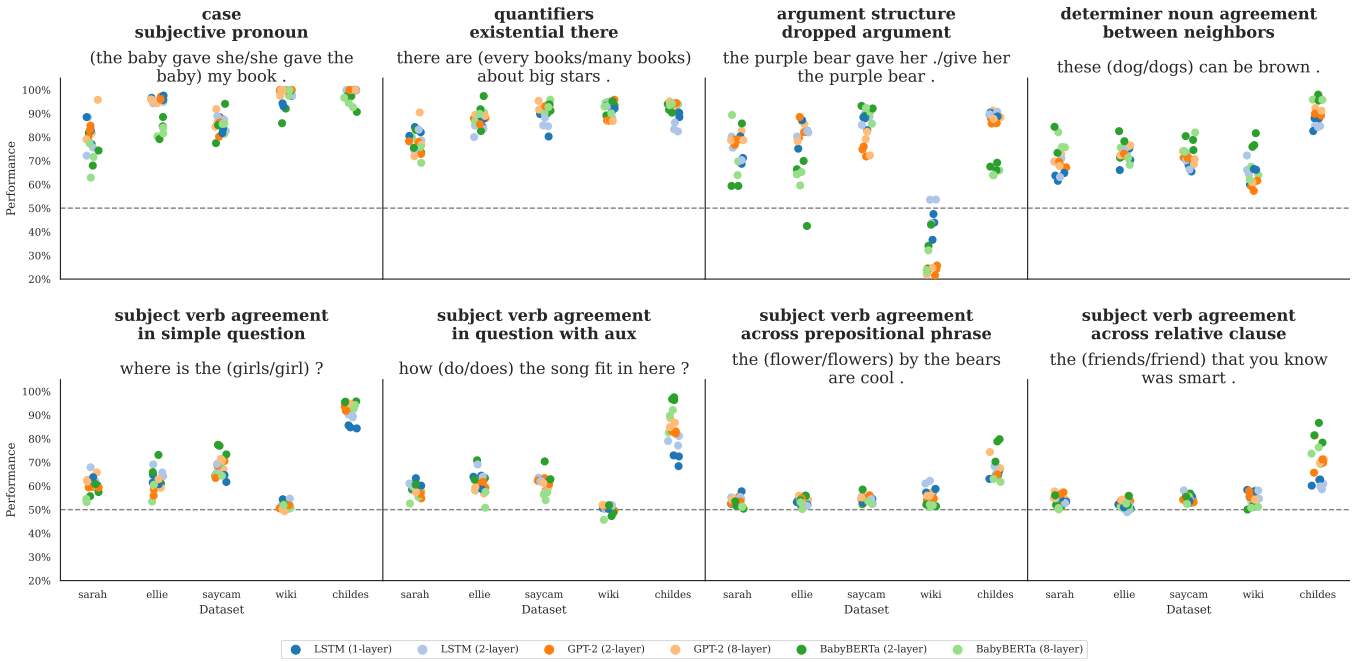


Figure 1: **Zorro test accuracies across different settings.** We tested 6 model architectures on 23 linguistic tests in Zorro. Each model architecture, trained with 3 seeds, yielded 18 accuracy data points per dataset. Our scatter plots show results for 8 selected tests, with the test name and an example sentence pair (unacceptable/acceptable) highlighted above each. For example, models evaluate which is more acceptable in the "case–subjective pronoun" test: "the baby gave she my book." or "she gave the baby my book." We found models trained on single-child datasets excel in specific tests but struggle in others, like subject-verb agreement. Four high-performing tests are shown in the first row, and four lower-performing tests, particularly for subject-verb agreement, are in the second row. Chance is the dotted line. Runs with 3 seeds show variability, similar to previous findings (Sellam et al., 2022; Yedetore et al., 2023).

Wang et al. (2023) conclusion from previous evaluations. As a comparison, models trained on CHILDES achieve higher test accuracy than models trained on other datasets, yet there is a noticeable variance in their accuracy as shown in the bottom right plot of Figure 1. This variability underscores the challenge of mastering the syntactic knowledge required for subject-verb agreement tests, despite the more enriched linguistic context CHILDES provides. More generally, the CHILDES corpus, which is much larger than other datasets, also yielded the best performance in many other tests.

Visualizations for Syntactic and Semantic Categories

In their study, Wang et al. (2023) followed a plan of analysis from Elman's pioneering work (Elman, 1989, 1990, 1991), demonstrating that CBOW and LSTM models when trained solely on the SAYCam-S dataset can form emergent clusters corresponding to syntactic categories such as nouns, transitive verbs, and intransitive verbs, and semantic categories such as food, animals and body parts. To analyze the cluster structures of word embeddings in their trained models, they visualized the embeddings by t-SNE (Van der Maaten & Hinton, 2008) and cluster dendrograms.

To test the robustness of Wang et al.'s findings, our study expands these visualizations to all models we mentioned above. As for syntactic distinctions, we found all models consistently exhibited clustering patterns in t-SNE plots and dendrograms across various datasets. We first analyze word embeddings of four syntactic categories (nouns, verbs, adjectives, and adverbs) using t-SNE, as illustrated in Figure 2. Focusing specifically on the three single-child datasets, we observe a distinct separation between nouns (marked in red) and verbs (marked in blue). Although some overlap exists, clusters of adjectives and adverbs are still discernible. Models trained on CHILDES and Wikipedia datasets displayed more distinct clustering, likely due to their broader vocabularies compared to single-child datasets.

As for semantic categorization, we use the same 8 child-directed semantic categories in Wang et al. (2023), which was derived from WordBank (Frank, Braginsky, Yurovsky, & Marchman, 2016). Due to differences in vocabulary, we cannot use the same set of words across all datasets. Therefore, for each dataset, we adapt the set of words in each category, enabling visualization of the six most frequent words per category. Figure 3 displays three models and reveals visually identifiable clusters such as body parts, clothing and animals.

Cloze Tests

In addition to examining emergent lexical classes in the representation space, we wanted to further test if models can properly identify the syntactic category of a missing word based on its surrounding context. Therefore, following Wang et al. (2023), we apply cloze tests (Taylor, 1953) to provide further evidence for syntactic category structures, specifically

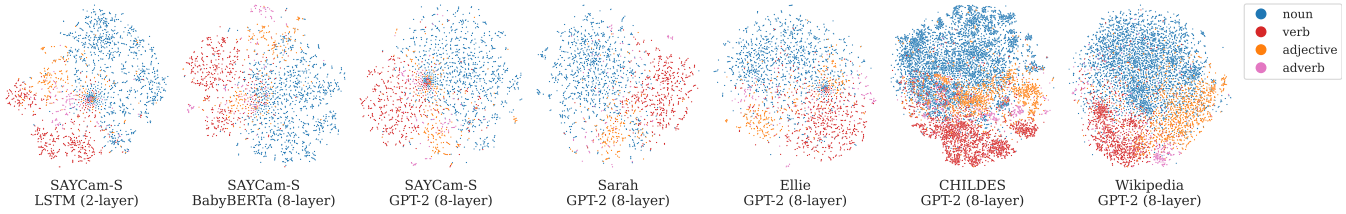
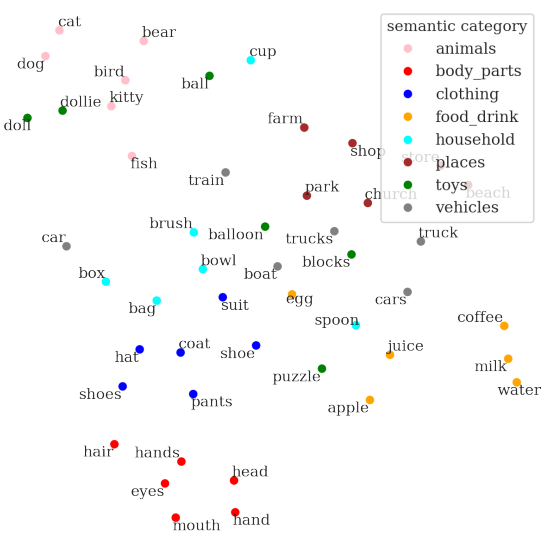
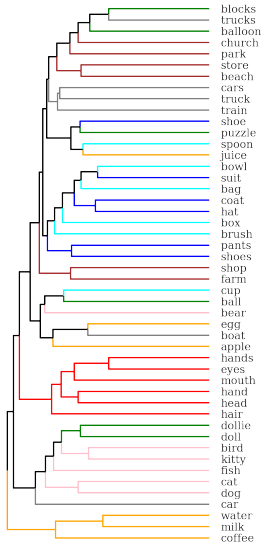


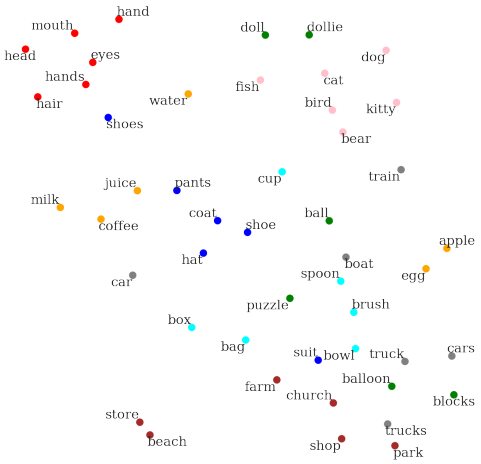
Figure 2: **Clustering different models' word embeddings for syntactic categories.** We ran t-SNE to visualize embeddings of all words in the vocabulary that are categorized into one of the four syntactic categories: noun, verb, adjective, and adverb. t-SNE uses $1 - \cos(u, v)$ as the distance metric. We show seven visualizations here from various training datasets and model architectures labeled below the plots. Nouns and verbs form two large salient clusters, while adjectives and adverbs are mostly clustered together.



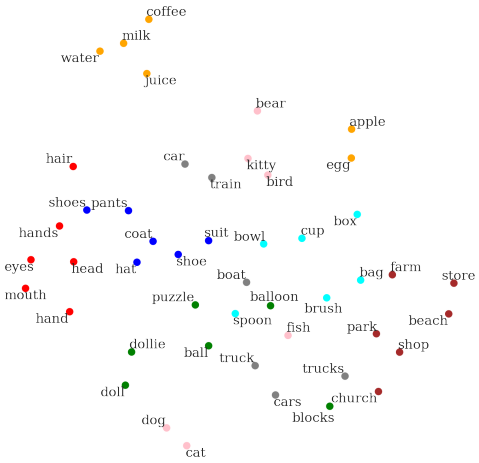
(a) GPT-2 (2-layers) t-SNE



(b) GPT-2 (2-layer) Dendrogram clustering



(c) BabyBERTa (2-layer) t-SNE



(d) LSTM (1-layer) t-SNE

Figure 3: **Clustering word embeddings for semantic categories.** Here we visualize word embeddings of three architectures trained on the Sarah dataset: (a, b) GPT-2 (2-layer), (c) BabyBERTa (2-layer), (d) LSTM (1-layer). Again, t-SNE and dendrogram plots use the cosine measure in Figure 2. We present the 6 most frequent words from 8 different categories. There are distinct clusters corresponding to semantic categories, including body parts, clothing, animals, and places.

the noun-verb distinction. We use clozes such as "we are going to __ here", where this cloze expects either a noun or a verb.[6] We follow the same process as Wang et al. (2023) to generate and evaluate the clozes for each dataset. Cloze

[6]Similar to the category distinction test in Kim and Smolensky (2021).

Table 5: **Cloze test statistics and accuracies (%) of differentiating noun vs. verb.** We build the cloze tests from the validation set for each dataset independently and evaluate the models correspondingly.

	SAYCam-S	**Sarah**	**Ellie**	Wikipedia	CHILDES
Number of clozes	2412	1763	1801	343	74266
Ratio of noun clozes	35.16%	34.66%	38.87%	69.97%	38.76%
LSTM (1-layer)	97.89	96.48	94.23	93.88	96.66
GPT-2 (2-layer)	**98.09**	95.92	94.39	93.88	97.23
GPT-2 (8-layer)	97.97	96.31	94.11	92.13	97.40
BabyBERTa (2-layer)	96.93	95.07	93.78	93.59	97.22
BabyBERTa (8-layer)	97.51	94.55	93.73	94.75	96.33

test statistics and accuracies are shown in Table 5. All of our models achieve over 90% accuracy, consistently demonstrating their ability to contextually differentiate nouns and verbs.

General Discussion

In order to study the robustness of Wang et al. (2023)'s learnability results from one child's linguistic input, we systematically trained 6 model architectures on 3 different single-child datasets. We found all trained models achieved consistent results in distinguishing syntactic and semantic categories of words, as well as sensitivity to several linguistic phenomena. We observed high performance on linguistic tests such as quantified existential "there" constructions, case of subjective pronouns, and dropped argument for ditransitive verb. But these models consistently failed on more complicated linguistic tests, such as subject-verb agreement across relative clause.

Unlike other work considering the importance of the domain of child-directed speech for learnability, this paper focuses specifically on the role of input to a single child. This approach offers a more realistic baseline than methods that train models on larger, aggregated data sources. With a similar goal, BabyLM challenge (Warstadt et al., 2023) explores learning under limited data conditions. However, even the smallest data track in the BabyLM challenge contains about 40 times more data (10M word tokens) than our single-child dataset. Similarly, in the study by Huebner et al. (2021), a RoBERTa-based Transformer was trained on 5M tokens from an age-ordered version of CHILDES (Huebner & Willits, 2020) and an equivalent amount from a Wikipedia dataset. Their analysis of the model's performance across various linguistic phenomena was conducted on Zorro. Intriguingly, we observed comparable patterns in our study, even though we used a much smaller dataset comprising single-child linguistic input and a corresponding Wikipedia dataset. Specifically, we found that models trained on the Wikipedia dataset struggled with tests such as dropped argument for ditransitive verb and local attractor in question with auxiliary verb, while the single-child datasets consistently outperformed in these areas. This closely mirrors the findings from Huebner et al. (2021)'s study using aggregated data sources and larger data quantity. Our results suggest that even limited data can be indicative of differences between datasets and, potentially, that child-directed speech may better equip models with the necessary linguistic abilities for certain tests.

The second key contribution of our study is an in-depth examination of the robustness of the findings by Wang et al. (2023), which were originally based on one single-child dataset: SAYCam-S. We expanded this investigation to include 3 single-child datasets with 2 baselines and 6 model architectures, significantly broadening the scope. Additionally, we enhanced the methodology for linguistic evaluation using the Zorro test suite (Huebner et al., 2021). Wang et al. previously limited their analysis to sentence pairs from Zorro that matched SAYCam-S's vocabulary, which resulted in a reduced test scope covering only 15 out of 23 tests and fewer than 700 sentence pairs per test. This limited size potentially weakened the validity of their conclusions. In contrast, we regenerated the Zorro test suite to align with the intersected vocabulary. Our models were then tested on comprehensive new 23 tests encompassing all 13 linguistic phenomena, with 2,000 sentence pairs in each test. This approach has yielded more robust and reliable results.

Our study demonstrates that models with different configurations can consistently learn to distinguish several syntactic and semantic categories and are sensitive to certain linguistic tests based solely on the linguistic input from a single child. However, we acknowledge several limitations. Firstly, while models demonstrate the ability to form syntactic and semantic clusters distinguishing lexical classes, it remains unclear how they acquire this representation and whether their understanding of these categories aligns with human cognition. Secondly, our evaluation methods, though insightful, are not exhaustive. The behavioral tests using Zorro are valuable for assessing responses to grammatical variations in sentences. However, it is important to note that Zorro has its limitations (Vázquez Martínez, Lea Heuser, Yang, & Kodner, 2023), and we still lack more systematic semantic evaluations. Lastly, our models are exclusively trained on transcribed speech. Wang et al. (2023) and Warstadt et al. (2023) suggest that integrating multiple modalities given realistic experience is a significant challenge in language learning, although there has been recent progress (Vong et al., 2024). We see multi-modal learning as a promising means of enhancing model data efficiency and realism by better capturing the learning problem faced by a young child.

Acknowledgments

We thank Wai Keen Vong, Cara Leong, Cindy Luo and Solim LeGris for helpful feedback on earlier drafts. This work was supported by NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science.

References

Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., & Steedman, M. (2017). Bootstrapping language acquisition. *Cognition*, *164*, 116–143.

Beaupoil-Hourdel, P. (2015). *Multimodal acquisition and expression of negation. analysis of a videotaped and longitudinal corpus of a french and an english mother-child dyad* (Unpublished doctoral dissertation). Ph. D. Dissertation, Sorbonne Nouvelle University, Paris.

Brown, R. (1973). *A first language: The early stages*. Harvard University Press.

Chang, T. A., & Bergen, B. K. (2023). Language model behavior: A comprehensive survey. *arXiv preprint arXiv:2303.11504*.

Elman, J. L. (1989). *Representation and structure in connectionist models*. Center for Research in Language, University of California, San Diego.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, *7*, 195–225.

Frank, M. C. (2023a). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, *27*(11), 990-992.

Frank, M. C. (2023b). Large language models as models of human cognition. *PsyArXiv*. Retrieved from `https://psyarxiv.com/wxt69`

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2016). Wordbank: an open repository for developmental vocabulary data*. *Journal of Child Language*, *44*, 677 - 694.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). Babyberta: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th conference on computational natural language learning* (pp. 624–646).

Huebner, P. A., & Willits, J. A. (2020). Order matters: Developmentally plausible acquisition of lexical categories. In *Cogsci.*

Kim, N., & Smolensky, P. (2021, February). Testing for grammatical category abstraction in neural language models. In *Proceedings of the society for computation in linguistics 2021* (pp. 467–470). Online: Association for Computational Linguistics.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259–284.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk* (3rd ed.). Lawrence Erlbaum Associates Publishers.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Sellam, T., Yadlowsky, S., Tenney, I., Wei, J., Saphra, N., D'Amour, A. N., ... Pavlick, E. (2022). The multiberts: Bert reproductions for robustness analysis. In *International conference on learning representations (iclr).*

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021). Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open mind*, *5*, 20–29.

Taylor, W. L. (1953). "Cloze Procedure": A New Tool for Measuring Readability. *Journalism & Mass Communication Quarterly*, *30*, 415 - 433.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Vázquez Martínez, H., Lea Heuser, A., Yang, C., & Kodner, J. (2023, December). Evaluating neural language models as cognitive models of language acquisition. In *Proceedings of the 1st genbench workshop on (benchmarking) generalisation in nlp* (pp. 48–64).

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*, 504–511.

Wang, W., Vong, W. K., Kim, N., & Lake, B. M. (2023). Finding structure in one child's linguistic experience. *Cognitive science*, *47*(6), e13305.

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. *Algebraic Structures in Natural Language*, 17–60.

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., ... others (2023). Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the babylm challenge at the 27th conference on computational natural language learning.*

Waterfall, H. R., Sandbank, B., Onnis, L., & Edelman, S. (2010). An empirical generative framework for computational modeling of language acquisition. *Journal of child language*, *37*(3), 671–703.

Yedetore, A., Linzen, T., Frank, R., & McCoy, R. T. (2023, July). How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed

speech. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 9370–9393).