On the Learnability of Syntax from Raw Speech with Autoregressive Predictive Coding

Shunsuke Kando, Yusuke Miyao
Department of Computer Science, The University of Tokyo,
{skando,yusuke}@is.s.u-tokyo.ac.jp

Abstract

Children are known to generalize syntactic knowledge at ages when their linguistic input is predominantly raw speech rather than text. This raises the question of whether syntactic generalization can emerge directly from acoustic input. We address this question using Autoregressive Predictive Coding (APC), a simple prediction-based self-supervised speech model. To approximate the input available to human learners while enabling controlled comparison, we train models on both child-directed speech and audiobook speech. We evaluate the models on a minimal-pair benchmark targeting elementary syntactic phenomena, designed to be acquisition-friendly. Our results show that APC partially generalizes word-order regularities when trained to predict near-future frames. However, the model fails to generalize agreement phenomena, suggesting that predictive learning from acoustic signals alone is insufficient. Furthermore, we observe distinct learning dynamics across word-order phenomena, suggesting that some improvements may be driven by shallow statistical regularities rather than genuine syntactic generalization.

1 Introduction

Children acquire a wide range of linguistic knowledge from speech input. Starting with sensitivity to prosody at birth (Mehler et al., 1988), they are reported to acquire typical vowel categories by 6 months (Kuhl et al., 1992), common nouns by 9 months (Bergelson and Swingley, 2012), and clausal units by 10 months (Hirsh-Pasek et al., 1987). Higher-level syntactic knowledge, such as agreement and word order, is typically acquired around the age of 3–4 (Kenney and Wolfe, 1972; Akhtar, 1999). Although such knowledge takes longer to develop than lower-level linguistic abilities, children's input remains predominantly raw speech rather than text throughout this period. A central question in developmental linguistics is how children can generalize a wide range of linguistic knowledge from noisy and limited speech input. While behavioral experiments provide insights into what children know about language, they do not directly reveal how such knowledge is acquired (Rowland et al., 2025). Moreover, such studies are often constrained by individual variability and limited scalability, making it difficult to obtain systematic and reproducible evidence.

In contrast to behavioral experiments, computational modeling offers a complementary approach for investigating the mechanisms of language acquisition in a controlled and scalable manner (Räsänen, 2026). While classical computational approaches have tested the language learning hypothesis primarily with artificial data or transcribed speech (Elman, 1990; Aslin et al., 1996; de Marcken, 1996), recent advances in neural network models enable simulation from realistic input, including raw speech (Dupoux, 2018; Räsänen, 2026). Previous research has examined phonetic/lexical learning (Lavechin et al., 2025; Khorrami et al., 2023), structural prosodic knowledge (De Seyssel et al., 2023), and word or syllable segmentation (Algayres et al., 2022; Pasad et al., 2024; Cho et al., 2025; Baade et al., 2025), using raw audio or audiovisual input. However, most work that examines higher-level linguistic knowledge involves optimization of multiple components (e.g., representation learning and language modeling (Lavechin et al., 2025)), making it difficult to isolate which components contribute to the observed performance.

In this paper, we examine the learnability of syntax with a predictive coding model. We employ Autoregressive Predictive Coding (APC; Chung et al., 2019), a simple yet effective self-supervised learning method. To evaluate syntactic knowledge, we use BabySLM (Lavechin et al., 2023), a minimal-pair benchmark designed to capture elementary syntactic phenomena. In the original pa-

Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL), pages 77–82

per, Lavechin et al. (2023) trains both representation learning and language modeling, where the latter models sequential patterns in the learned representations. In contrast, our approach isolates representation learning by removing language modeling over discrete symbolic units, allowing us to directly assess whether predictive coding alone can induce syntactic generalization. We trained models on both audiobook and child-directed speech audio.

Our results show that certain word-order phenomena can be learned with APC, although the specific patterns that are acquired depend on the training data. In contrast, agreement phenomena remain at chance level, suggesting that predictive coding alone is insufficient for capturing more complex syntactic dependencies. Beyond final performance, we also analyze the learning trajectory of syntactic phenomena over the course of training. We observe that the accuracy on Adj–noun order rises above chance at the beginning of training (far before 1 epoch), which might indicate the model's reliance on shallow local regularities rather than genuine syntactic generalization. Furthermore, different word-order phenomena exhibit distinct learning dynamics, indicating that not all apparent improvements reflect the same type of generalization.

The experimental codebase is made publicly available[1].

2 Autoregressive Predictive Coding

Autoregressive Predictive Coding (APC; Chung et al., 2019) is an unsupervised representation learning method based on predictive coding. Given a sequence of log Mel spectrograms $\boldsymbol{x} = (x_1, x_2, \ldots, x_T)$, the model is optimized to predict a frame n step ahead of the current one. Prediction vectors $\boldsymbol{y} = (y_1, y_2, \ldots, y_T)$ are produced by a recurrent neural network (RNN). The objective is defined as the L1 loss between the predicted and target frames:

$$\mathcal{L} = \sum_{i=1}^{T-n} |x_{i+n} - y_i|,$$

where n denotes the number of frames ahead to predict. The hyperparameter n encourages the model to capture global structures beyond immediate frame-level continuity.

Despite its simplicity and unsupervised nature, APC has been shown to be effective across a wide range of speech tasks, including phone classification and discrimination (Blandón and Räsänen, 2020), speaker verification (Chung et al., 2019), and automatic speech recognition (Yang et al., 2022).

A major variant of APC is Contrastive Predictive Coding (CPC) (van den Oord et al., 2019), which differs in two key aspects: (1) it employs a convolutional neural network for feature extraction instead of Mel spectrograms, and (2) it is trained with a contrastive loss instead of an L1 loss. In this work, we focus on APC due to its architectural simplicity and interpretability as a predictive model, which makes it suitable for analyzing the learnability of syntactic structure. We note that CPC has also been widely studied in the context of language acquisition modeling, and extending our analysis to CPC remains an important direction for future work.

3 Experimental Setup

To examine whether developmentally plausible speech audio can induce syntactic knowledge, we trained models on a child-directed speech (CDS) dataset. In addition, we train separate models on audiobook speech as a contrasting condition to assess the role of input characteristics. Syntactic knowledge is evaluated using a minimal-pair benchmark.

3.1 Dataset

To construct the CDS dataset, we extract English subsets[2] from CHILDES database (Macwhinney, 2000). We trim the spoken part of the audio using time alignments provided in the CHAT transcriptions. Since our focus is on child-directed speech audio, we exclude utterances produced by children or speakers with unknown roles. The resulting CDS dataset has a total duration of 995 hours. As a contrasting condition, we use LibriSpeech (Panayotov et al., 2015) as an audiobook dataset, with a comparable total duration of 960 hours. The total number of utterances is 2M for CDS and 281K for audiobook, indicating that CDS consists of shorter utterances on average. The dataset is split into training set and validation set at the ratio of 99:1.

3.2 Model Setup

We use an APC model consisting of three unidirectional LSTM layers with 512 hidden units each, followed by a linear layer for frame prediction. We observe that the number of training steps required

[1] https://github.com/gifdog97/babyslm_apc

[2] Eng-AAE, Eng-NA, and Eng-UK.

Table 1: Example pairs of syntactic tasks in BabySLM.

Phenomenon	Pair example
Adjective–noun order	✓ The good mom. ✗ The mom good.
Noun–verb order	✓ The dragon says. ✗ The says dragon.
Anaphor–gender agreement	✓ The dad cuts himself. ✗ The dad cuts herself.
Anaphor–number agreement	✓ The boys told themselves. ✗ The boys told himself.
Determiner–noun agreement	✓ Each good sister. ✗ Many good sister.
Noun–verb agreement	✓ The prince needs the princess. ✗ The prince need the princess.

for convergence differs between CDS and audiobook data, likely due to differences in the number and length of utterances. Hence, we train for 5 epochs on CDS and 25 epochs on the audiobook dataset. We used an AdamW optimizer with a batch size of 256 and an initial learning rate of 10^{-3}. We varied the step size n from 1 to 16.

3.3 Evaluation

To evaluate syntactic knowledge, we use BabySLM (Lavechin et al., 2023), a minimal-pair benchmark targeting elementary syntactic phenomena (Table 1). In this setup, each pair consists of a grammatical and an ungrammatical sentence. The model is evaluated based on whether it assigns a higher score to the grammatical audio. In the original BabySLM setup, language models are trained on top of learned speech representations, and scores are computed using negative log-likelihood. In contrast, we directly use the prediction error of APC as a scoring function. Specifically, the score of an input audio sequence $\boldsymbol{x}$ is defined as:

$$score(\boldsymbol{x}) = -\frac{1}{T-n}\sum_{i=1}^{T-n}|x_{i+n} - y_i|,$$

where higher scores indicate better predictions.

4 Results

4.1 The effect of dataset and n

Table 2 shows the average accuracy across syntactic phenomena. Overall, most accuracies remain close to chance level, regardless of the dataset and the prediction step n. This confirms the inherent difficulty of the syntactic tasks, which has also been

Table 2: Overall accuracy on BabySLM syntactic test. Chance rate is 0.5.

n	CDS		AudioBook	
	dev	test	dev	test
1	0.52	0.505	0.51	0.502
2	0.533	0.513	0.527	0.51
3	0.528	0.501	0.535	0.502
4	0.518	0.5	0.537	0.509
5	0.491	0.479	0.548	0.504
6	0.498	0.477	0.523	0.495
7	0.494	0.472	0.507	0.491
8	0.464	0.469	0.515	0.504
9	0.467	0.464	0.494	0.494
10	0.468	0.461	0.495	0.502
11	0.459	0.46	0.494	0.51
12	0.445	0.444	0.488	0.49
13	0.456	0.458	0.472	0.494
14	0.45	0.457	0.462	0.482
15	0.469	0.475	0.469	0.485
16	0.459	0.455	0.497	0.493

observed in prior work on speech-based language models (Lavechin et al., 2023). Figure 1 shows accuracy broken down by syntactic phenomenon on the development set. These results show that certain word-order phenomena can be partially learned with APC. In particular, Adj–noun order is captured when models are trained on child-directed speech with near-feature prediction ($2 \leq n \leq 7$); Noun–verb order is captured under a narrower range ($3 \leq n \leq 5$) when trained on audiobook data. In contrast, models trained with $n = 1$ fail to capture word-order regularities, indicating that immediate frame-level prediction is insufficient and that integrating information over a slightly longer temporal context is necessary.

On the other hand, all agreement phenomena remain at near-chance levels across all values of n and training data. This suggests that predictive learning from local acoustic signals alone is insufficient to capture agreement, which involves dependencies between more abstract linguistic units. This may indicate the potential importance of modeling over discrete or symbolic units, as in language models.

Interestingly, performance on Noun–verb order falls significantly below chance at larger values of n, suggesting a preference for the opposite order.

4.2 Analysis of Learning Trajectory

To better understand the nature of syntactic learning of APC, we analyze the learning trajectory of each syntactic phenomenon. We track the accuracy for the $n = 3$ setting throughout training, evaluating every 100 steps from 0 to 6,300. Figure 2

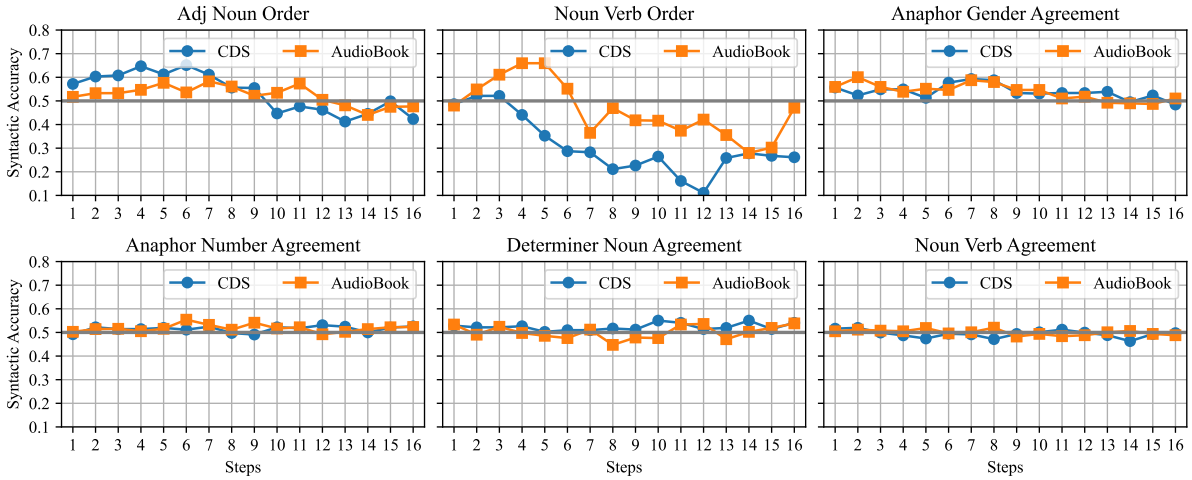

Figure 1: Accuracy of each phenomenon. X-axis represents n, the number of frame steps ahead to predict.

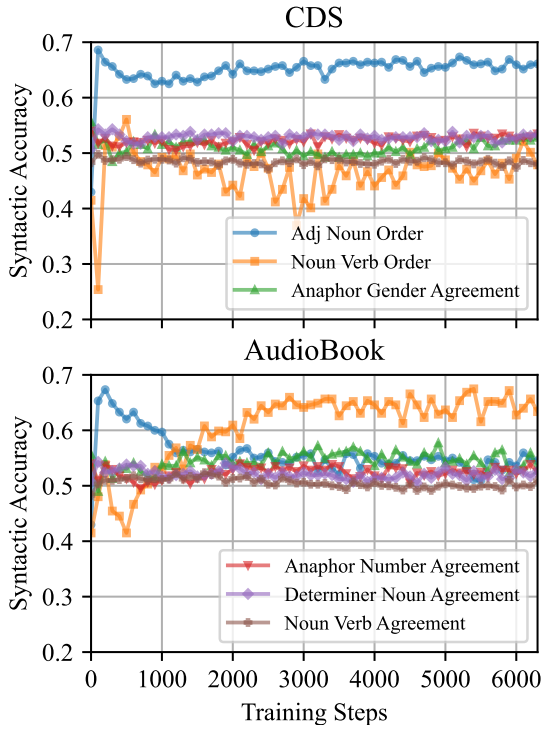

Figure 2: The trajectory of accuracy with $n = 3$.

shows the result. We observe that the accuracy of agreement phenomena remains at near-chance levels throughout training, further supporting the difficulty of learning it with predictive coding. In contrast, word-order phenomena exhibit distinct learning dynamics. For both datasets, the accuracy of Adj–noun order increases rapidly at the first 100 steps, well before completing a single epoch[3]. One possible explanation for this behavior is that the model relies on shallow local regularities (e.g., local co-occurrence patterns) rather than genuine syntactic knowledge. Since APC is optimized for local acoustic prediction, it may exploit short-range patterns that correlate with the correct answer. Such cues may not reflect the word-order structure that the benchmark is intended to probe. Further investigation is required to clarify whether the model genuinely acquires syntactic generalization.

We also observe the swap of accuracy between Adj–noun order and Noun–verb order when models are trained on audiobook data. This contrast suggests that the two phenomena may rely on qualitatively different cues. The rapid rise and subsequent degradation in Adj–noun order accuracy are consistent with the model relying on shallow local regularities that do not generalize beyond early training. In contrast, the gradual improvement observed in Noun–verb order may reflect the acquisition of more robust and generalizable patterns. In this sense, Noun–verb order might provide a more reliable indicator of genuine generalization than Adj–noun order.

5 Conclusion

In this paper, we investigated the learnability of syntax from raw speech using Autoregressive Predictive Coding (APC). While APC captures certain word-order phenomena, it fails to generalize agreement, suggesting that predictive coding over local acoustic signals is insufficient for modeling complex syntactic dependencies. Analysis of learning trajectories further reveals that rapid improvements may stem from shallow statistical regularities rather than genuine generalization. These results highlight the need for additional mechanisms, such

[3]1 epoch equals 7,600 steps for CDS and 1,000 steps for audiobook.

as representations over more abstract or symbolic units, for acquiring syntactic knowledge.

Limitations

First, we focused on a single predictive coding model (APC), whilce CPC (van den Oord et al., 2019) has also been widely studied in speech representation learning or language acquisition modeling literature. Future work should examine CPC and related models to determine whether our findings generalize beyond APC. Second, we used only BabySLM benchmark for evaluation. To better assess fine-grained syntactic knowledge of the models, future work should complement this benchmark with additional evaluation methods, such as probing (He et al., 2025) or canonical correlation analysis (Pasad et al., 2024). Third, the cause of the distinct learning trajectories observed for word-order phenomena remains unclear. In particular, our hypothesis that Adj–noun order may be solvable without syntactic generalization requires more careful verification. Finally, the training dataset is not fully developmentally plausible. Children are reported to be exposed not only to CDS, but also to overheard speech (Thompson, 2018) and media audio (Gowenlock et al., 2024), both of which constitute a non-negligible portion of their linguistic input. In addition, the audio quality of CHILDES recordings is not always ideal, particularly for older corpora. To better facilitate the reverse engineering of language acquisition, further work is needed to construct more ecologically valid models of children's language inputs.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 26KJ0792 and JST ACT-X Grant Number JPMJAX24C9.

References

Nameera Akhtar. 1999. Acquiring basic word order: Evidence for data-driven learning of syntactic structure. *Journal of child language*, 26:339–56.

Robin Algayres, Tristan Ricoul, Julien Karadayi, Hugo Laurençon, Salah Zaiem, Abdelrahman Mohamed, Benoît Sagot, and Emmanuel Dupoux. 2022. DP-Parse: Finding Word Boundaries from Raw Speech with an Instance Lexicon. *Transactions of the Association for Computational Linguistics*, 10:1051–1065.

Richard N. Aslin, Julide Z. Woodward, Nicholas P. LaMendola, and Thomas G. Bever. 1996. Models of word segmentation in fluent maternal speech to infants. In *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*, pages 117–134. Lawrence Erlbaum Associates, Inc.

Alan Baade, Puyuan Peng, and David Harwath. 2025. SyllableLM: Learning Coarse Semantic Units for Speech Language Models. In *ICLR 2025*. OpenReview.net.

Elika Bergelson and Daniel Swingley. 2012. At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258.

María Andrea Cruz Blandón and Okko Räsänen. 2020. Analysis of Predictive Coding Models for Phonemic Representation Learning in Small Datasets. In *workshop on Self-supervision in Audio and Speech at ICML 2020*.

Cheol Jun Cho, Nicholas Lee, Akshat Gupta, Dhruv Agarwal, Ethan Chen, Alan W. Black, and Gopala K. Anumanchipalli. 2025. Sylber: Syllabic Embedding Representation of Speech from Raw Audio. In *ICLR 2025*. OpenReview.net.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. 2019. An Unsupervised Autoregressive Model for Speech Representation Learning. In *INTERSPEECH 2019*, pages 146–150.

Carl de Marcken. 1996. *Unsupervised Language Acquisition*. phdthesis, MIT.

Maureen De Seyssel, Marvin Lavechin, Hadrien Titeux, Arthur Thomas, Gwendal Virlet, Andrea Santos Revilla, Guillaume Wisniewski, Bogdan Ludusan, and Emmanuel Dupoux. 2023. ProsAudit, a prosodic benchmark for self-supervised speech models. In *INTERSPEECH 2023*, pages 2963–2967. ISCA.

Emmanuel Dupoux. 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59.

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

Anna Elizabeth Gowenlock, Courtenay Norbury, and Jennifer M. Rodd. 2024. Exposure to language in video and its impact on linguistic development in children aged 3–11: A scoping review. *Journal of Cognition*.

Linyang He, Qiaolin Wang, Xilin Jiang, and Nima Mesgarani. 2025. Layer-wise Minimal Pair Probing Reveals Contextual Grammatical-Conceptual Hierarchy in Speech Representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35338–35353. Association for Computational Linguistics.

Kathy Hirsh-Pasek, Deborah G. Kemler Nelson, Peter W. Jusczyk, Kimberly Wright Cassidy, Benjamin Druss, and Lori Kennedy. 1987. Clauses are perceptual units for young infants. *Cognition*, 26(3):269–286.

Terrence J. Kenney and Jean Wolfe. 1972. The acquisition of agreement in English. *Journal of Verbal Learning and Verbal Behavior*, 11(6):698–705.

Khazar Khorrami, María Andrea Cruz Blandón, and Okko Räsänen. 2023. Computational Insights to Acquisition of Phonemes, Words, and Word Meanings in Early Language: Sequential or Parallel Acquisition? *Proceedings of the Annual Meeting of the Cognitive Science Society*, 45(45).

P. K. Kuhl, K. A. Williams, F. Lacerda, K. N. Stevens, and B. Lindblom. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044):606–608.

Marvin Lavechin, Maureen de Seyssel, Hadrien Titeux, Guillaume Wisniewski, Hervé Bredin, Alejandrina Cristia, and Emmanuel Dupoux. 2025. Simulating Early Phonetic and Word Learning Without Linguistic Categories. *Developmental Science*, 28(2):e13606.

Marvin Lavechin, Yaya Sy, Hadrien Titeux, María Andrea Cruz Blandón, Okko Räsänen, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. 2023. BabySLM: Language-acquisition-friendly benchmark of self-supervised spoken language models. In *INTERSPEECH 2023*, pages 4588–4592. ISCA.

Brian Macwhinney. 2000. The CHILDES Project: Tools for Analyzing Talk (third edition): Volume I: Transcription format and programs, Volume II: The database. *Computational Linguistics*, 26:657–657.

Jacques Mehler, Peter Jusczyk, Ghislaine Lambertz, Nilofar Halsted, Josiane Bertoncini, and Claudine Amiel-Tison. 1988. A precursor of language acquisition in young infants. *Cognition*, 29(2):143–178.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. 2024. What Do Self-Supervised Speech Models Know About Words? *Transactions of the Association for Computational Linguistics*, 12:372–391.

Caroline F. Rowland, Gert Westermann, Anna L. Theakston, Julian M. Pine, Padraic Monaghan, and Elena V. M. Lieven. 2025. Constructing language: A framework for explaining acquisition. *Trends in Cognitive Sciences*.

Okko Räsänen. 2026. Computational modeling of early language learning from acoustic speech and audiovisual input without linguistic priors. *Preprint*, arXiv:2603.08359.

Abbie Thompson. 2018. Who's Talking to Whom and Does It Matter? The Impact of Multiple Speakers, Overheard Speech, and Child-Directed Speech on Infants' Language Development.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. *Preprint*, arXiv:1807.03748.

Gene-Ping Yang, Sung-Lin Yeh, Yu-An Chung, James Glass, and Hao Tang. 2022. Autoregressive Predictive Coding: A Comprehensive Study. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1380–1390.