From Infant to Toddler to Preschooler: Analyzing Language Acquisition in Language Models

Stanford CS224N Custom Project

Yash Shah
Department of Computer Science
Stanford University
ynshah@stanford.edu

Abstract

A key problem explored within the field of developmental cognition is language acquisition—infants acquire linguistic skills over a period of time, with the learning trajectory reflecting the idiosyncratic ways in which adults talk to them. Here, we model the first six years of language acquisition by training an artificial language model on child-directed transcripts of spoken speech presented in an *age-ordered* fashion. We show that developmentally-plausible trajectories of syntactic knowledge acquisition emerge with age. These trajectories are underscored by evolving attention patterns of children over different components of sentences during comprehension as well as the formation of increasingly meaningful textual representation clusters. Complex semantic understanding, however, seems to require cognitive development beyond the first six years of life. These results not only provide insights into language development that can benefit researchers and linguists but also help in identifying effective learning strategies for children.

Mentor: Chaofei Fan

1 Introduction

Developmental psychology, cognitive science, and psycholinguisitics research have long explored the roles of language acquisition and scaffolding in infants from a behavioral point of view (Kuhl, 2000; Sakai, 2005; Lust, 2006; Clark and Casillas, 2015; Guasti, 2017). A key component of language acquisition is grammatical development—syntactic information is learned gradually over different early stages in the child's lifespan, with the exact course of acquisition dependent on the child's mother's underlying grammar (Cameron-Faulkner et al., 2003). According to Gleason and Ratner (2022), young children (ages $\sim$2) often produce sentences that lack tense and person markers (for e.g., *elephant fall down*). As early as 2½ years, children begin combining sentences to express complex or compound propositions. By age 3 or 4, they start acquiring a rule-governed grammar system. This system leads them to incorrectly use irregular verbs (*teeth* $\rightarrow$ *teeths*) and tenses (*fall* $\rightarrow$ *falled*). Additionally, children start learning different types of sentences, such as negatives (for e.g., *no go movies*), wh-questions (without auxiliary verbs; for e.g., *where daddy go?*), and imperatives. There is a heavy reliance on different intonation patterns used by adults for language development. 4 year-olds could work with relative clauses (for e.g., *pick up the walrus that is tickling the zebra*). Yet, certain constructions are not fully controlled by children at the time they enter school, like anaphora.

From a point of view of computational cognition and neuroscience, prior works have optimized artificial neural networks of the ventral visual pathway on an auxiliary objective to understand brain function and development (Yamins et al., 2014). Within language, large Transformer-based language models (TLMs) like GPT, BERT, and RoBERTa have significantly advanced performance on several benchmarks of natural language understanding tasks, although when pre-trained on massive datasets (Radford et al., 2018, 2019; Brown et al., 2020; Devlin et al., 2018; Liu et al., 2019). While this

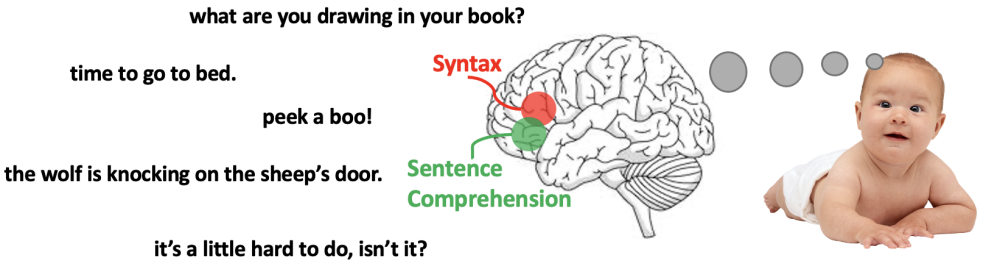

Figure 1: **Language acquisition in infants.** Infants pick up linguistic skills from the way adults, typically their parents, speak to them. This way of speaking differs significantly over the child's growth period, and the idiosyncrasies shape the trajectory of language acquisition. Sakai (2005) identified the left lateral pre-motor cortex as the "grammar center" and the left inferior frontal gyrus for sentence comprehension.

is an impressive feat, questions around language acquisition theory, requiring models to be trained on developmentally plausible input of the quality and quantity that children are exposed to, still prevail. Huebner et al. (2021); Vázquez Martínez et al. (2023); Yedetore et al. (2023) have previously evaluated acquisition of linguistic skills like grammar and hierarchical rules in language models using transcripts of child-directed speech. While informative, they fail to consider that linguistic traits often emerge gradually over several years.

In this paper, we are the first to ask how linguisitic skills are developed over the course of several years in young children (figure 1). We do so by evaluating grammatical knowledge acquisition in language models when trained on transcripts of child-directed speech in an *age ordered* fashion. Specifically, we address the following:

1. The vocabulary of a child is significantly affected by how adults (parents or caregivers) speak to them (Cameron-Faulkner et al., 2003), with, for e.g., the verbs first used by them being those most frequently used by their mothers (Naigles and Hoff-Ginsberg, 1998). If adults use simple sentences that are mostly rhetorical with lots of repetition for infants (for e.g., *Is that a baby? Is that a picture of a baby? Oh what a nice baby.*) and longer, more complex, sentences for preschoolers, how is grammar acquisition affected by this change? (section 4.1).
2. In addition to grammatical knowledge, to what extent is complex semantic knowledge learned in early years? (section 4.2).
3. What underlying mechanisms support such acquisition? Is there a qualitative change in the attention maps and textual embedding representations over time? (sections 4.3 and 4.4).

2 Related Work

Evaluating language models. There is a rich set of literature on language model evaluation—previous works have analyzed the capability of language models to perform natural language understanding in the form of sentiment analysis (Liang et al., 2023; Zeng et al., 2023), text classification (Liang et al., 2023), natural language inference (Lee et al., 2023), semantic understanding (Tao et al., 2023), reasoning (Xu et al., 2023; Wu et al., 2023; Gendron et al., 2024), performance on multilingual tasks (Ahuja et al., 2023), factuality (Honovich et al., 2022; Wang et al., 2023), robustness (Yang et al., 2023) and bias (Gehman et al., 2020), and several others. All these evaluation datasets, benchmarks, and analyzes are aimed at understanding language models for the sake of being able to deploy and work with them in everyday life. Language models evaluated are those that are typically huge in the number of parameters and are trained on developmentally unrealistic amounts of data from the internet.

Computational language acquisition. In contrast to a general evaluation of language models, research that has tried to model language acquisition have done so to further our understanding of

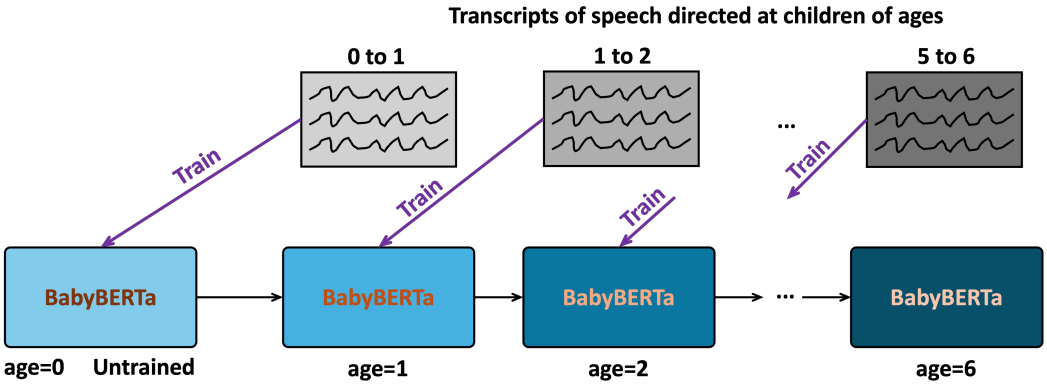

Figure 2: **Schematic of the acquisition mechanism.** We start with an untrained BabyBERTa network to represent an infant who has not yet started acquiring linguistic skills. We first train it on transcripts of speech directed at children of ages between 0 and 1. Next, we take this partially trained BabyBERTa and feed it transcripts for children between the ages of 1 and 2. We repeat this process.

psycholinguisitics and developmental cognition. Huebner et al. (2021) develop BabyBERTa—a scaled-down version of RoBERTa-base (Liu et al., 2019)—as well as a grammar test suite called Zorro to observe learning of strong grammatical knowledge even on small corpora of child-directed language. Vázquez Martínez et al. (2023) carry this forward by evaluating BabyBERTa not just on Zorro but a wide range of other linguisitic benchmarks. Yedetore et al. (2023) try to understand the use of hierarchical rules by language models when trained on child-directed speech. A limitation of these works is that they train language models on the entire training data of child-directed language in a random order all at once. We know that adults adapt the way they talk to an infant versus a preschooler. We thus distinguish ourselves by training BabyBERTa on child-directed language in an *age-ordered* fashion to analyze acquisition trajectories across linguistic skills.

3 Methods

We use BabyBERTa (Huebner et al., 2021) as an artificial model of how language acquisition works in the biological brain. It is trained using a masking removal policy. This means that if the objective is to reconstruct masked tokens when $p\%$ of the tokens are masked, the loss function is defined as

$$\mathcal{L} = -\sum_{i=1}^{n}\sum_{j=1}^{m_i} \log P(w_{i,j}|\tilde{x}_i),$$

where $w_{i,j}$ is the ground truth of the j^{th} masked token of the i^{th} sequence, $\tilde{x}_i$ is the masked context, n is the total number of sentences, and m_i is the number of masked tokens in the sentence. Masked tokens are unmasked for prediction 0% of the time. This allows BabyBERTa to attend to lexical context over the input for prediction.

To model acquisition, we use the following approach (figure 2): we first train BabyBERTa on transcripts of speech an infant would hear during its first year of life. We evaluate the model's capability of acquiring linguistic skills (syntactic and semantic) at this stage. We then train this model on transcripts of speech an infant would hear during its second year of life. This means that the model obtained at this stage would have seen data that children of both ages 1 and 2 would hear. The process continues.

3.1 Data

We train BabyBERTa on the Age-Ordered CHILDES Dataset (Huebner and Willits, 2021). Sentences in this dataset are attached to the age of the child to which they were spoken to by an adult. There are 3304 different transcripts in the whole corpus, and 4,960,141 words. Example sentences include *let's go*, *we're gonna get some clothes*, *this is a toy*, and *alright well we'll just play around*. If grouped into

years, there are 89 transcripts for ages 0-1, 854 for ages 1-2, 1385 for ages 2-3, 691 for ages 3-4, 253 for ages 4-5, 32 for ages 5-6. The standard pre-processing routine of lowercasing sentences and using a byte-level Byte-Pair Encoding (BPE) tokenizer are used. Because 5M words are relatively small for a dataset, we duplicate each input sequence 10 times and apply a novel random mask to each. This ensures that BabyBERTa receives $\sim$50M words, which becomes developmentally plausible.

3.2 Metric

To evaluate *syntactic* knowledge acquisition, we use Zorro (Huebner et al., 2021). This grammar test suite contains pairs of test sentences that isolate specific phenomena in syntax and morphology, such as island effects and determiner-noun agreement. For example, BabyBERTa must score higher on the grammatically correct sentence *this dollar must be sufficient* over the grammatically incorrect sentence *this dollars must be sufficient*. A preference score is computed by summing the cross-entropy errors at each position in the sentence, with a higher sum denoting a higher surprisal. A complete list of phenomena evaluated are shown in section A.1. Additionally, we evaluate *semantic* knowledge acquisition using tasks from the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018): sentiment analysis, paraphrasing, natural language inference, textual entailment, etc. Additional details on the tasks are presented in section A.2. During evaluation, we assume that the vocabulary of a model "aged" $\mathcal{X}$ is the same as the vocabulary of a model "aged" $\mathcal{Y}$ for all $\mathcal{X}, \mathcal{Y} \in [6]$; i.e., we use the same sentences in the benchmark tasks across trained models to make results comparable.

3.3 Implementation Details

For training BabyBERTa, we use an AdamW optimizer (Loshchilov and Hutter, 2019). The learning rate is initialized to 1e-4 and then decayed according to a linear schedule with a warmup period of 1k steps for age 1, 9k for age 2, 150k for age 3, 7k for age 4, 2k for age 5, and 0 for age 6. We use a batch size of 16. The data is split into a training and validation set based on a train split fraction of 0.9 for age 1, 0.95 for age 2, 3, 4, and 0.9 for age 5, 6. Hyper-parameter tuning is performed using the validation set perplexities. Experiments are run on an NVIDIA T4 GPU on Google Colab and take around 1 hour each for training.

For evaluation on the GLUE dataset, we use an Adam optimizer with the learning rate initialized to 2e-5 and decayed according to a linear schedule. The optimizer is trained for 10 epochs with 500 warmup steps and a batch size of 16. Experiments are run on an NVIDIA T4 GPU on Modal,[1] taking around 5 hours each.

4 Results

4.1 With age, developmentally-plausible trajectories of syntactic knowledge acquisition emerge for different phenomena

With our initial evaluation of grammar learning by BabyBERTa, we observe that there is an increase in syntactic knowledge with age, with most of the skills learned by the age of 3 (figure 3). Different syntactic phenomena are learned differently and with different speeds. Determiner-noun agreements, subject-verb agreements, filler-gaps, and existential quantifiers are learned gradually and mastered (accuracy $>$80%) by an age of 3. Phenomena such as swapped arguments, subjective pronouns, and local attractors are mastered from the first year of life. In contrast, phenomena like anaphora, irregular verbs, and ellipsis are either never learned or mastered. These results mostly align with developmentally-plausible learning trajectories. Low performance on irregular verbs demonstrates the plausible learning of a rule-based system that is observed in children as well (Gleason and Ratner, 2022). Contrastly, it is surprising that performance on superlative quantifiers is better than chance right from an age of 1, but goes down by an age of 5. When we compare our results to the results from Huebner et al. (2021) who trained BabyBERTa on the same dataset but in an *age-agnostic* fashion, we find that our model is able to achieve the same accuracy across all syntactic phenomena. Moreover, we often find some phenomena like determiner-noun agreements, irregular verbs, and island effects with coordinate structure constraints achieve that accuracy with even fewer amounts

[1] https://modal.com/

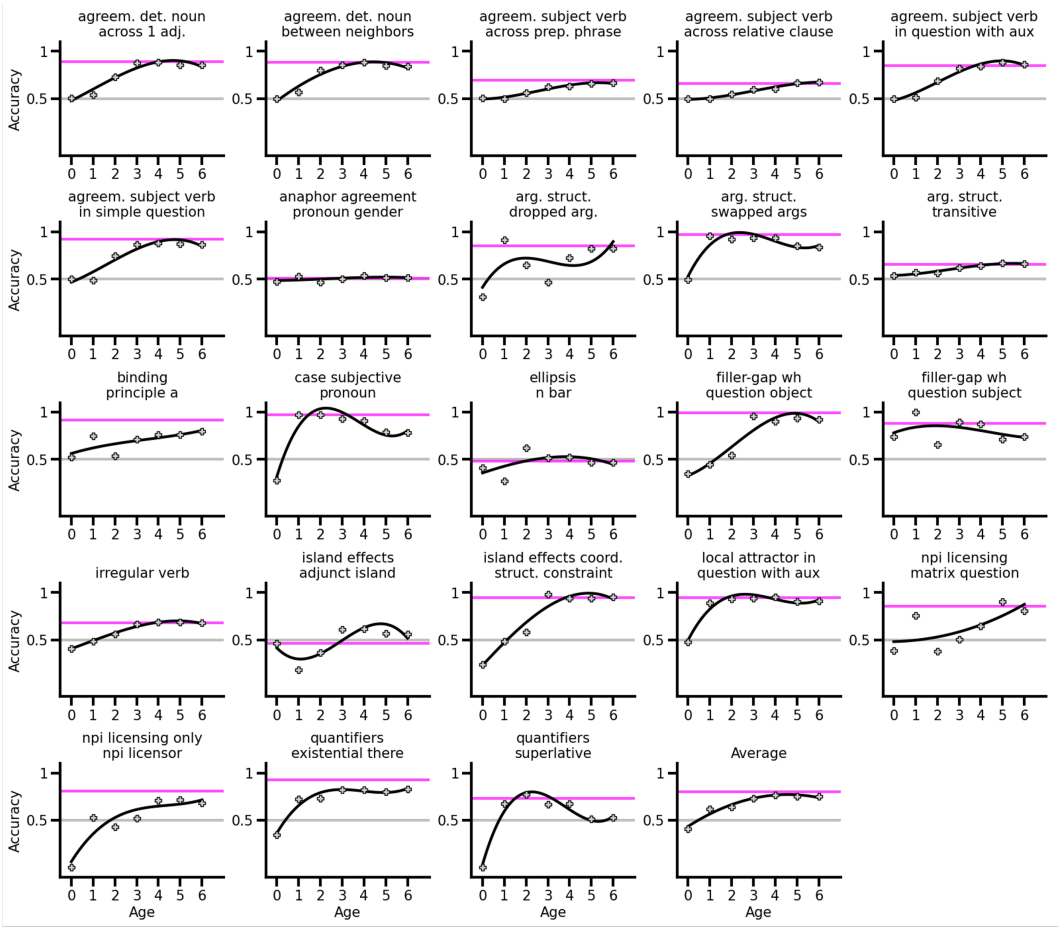

Figure 3: **Evaluation on Zorro.** Gray markers represent the top-1 test accuracy of BabyBERTa as a function of its "age". Black lines are degree 3 curves of best fit. Pink lines represent the accuracy that BabyBERTa achieves when trained on the entire training data in an *age-agnostic* manner (Huebner et al., 2021).

of data (earlier along the developmental trajectory) than Huebner et al. (2021), highlighting the remarkable capabilities of language acquisition with low amounts of verbal stimuli.

4.2 Complex semantic understanding requires cognitive development beyond the first few (=6) years of life

Next, we evaluate the semantic understanding of BabyBERTa across different "ages" through complex tasks like sentiment analysis, textual entailment, paraphrasing, etc. We find that even after a period of 6 years, semantic relationships for complex tasks are not learned (figure 4). Across different tasks from the GLUE dataset, the performance of BabyBERTa at an age of 1 is approximately the same as that at an age of 6. This is not surprising; these tasks not only contain sentences that are semantically complex and most likely not similar to what children would hear at their age (for e.g., *The city Tenochtitlan grew rapidly and was the center of the Aztec's great empire*), the vocabulary that they encompass expands beyond what children learn within their first 6 years of life. Good performance on these tasks will require more time as well as learning from speech or written text directed at adults (like Wikipedia).

4.3 The sentence components that children attend to during comprehension evolves with age

In an attempt to understand the mechanisms underlying early grammar acquisition, we visualize attention heatmaps of BabyBERTa at different developmental stages for five different sentences

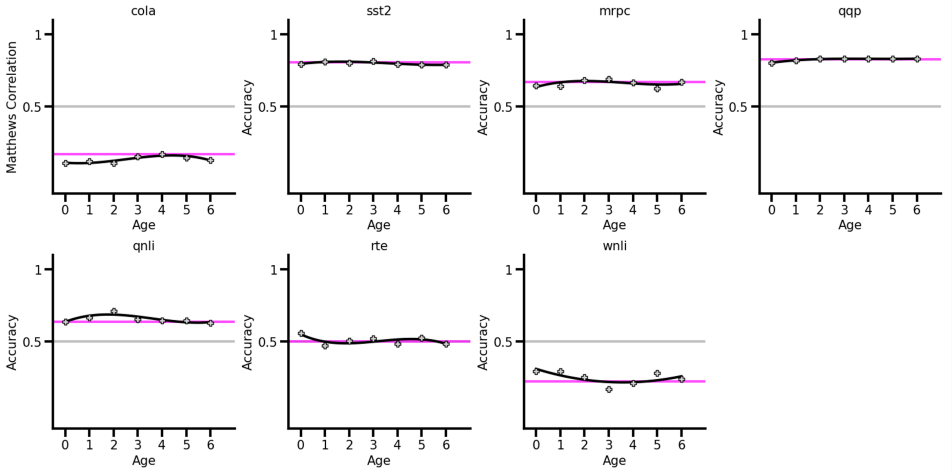

Figure 4: **Evaluation on GLUE.** Gray markers represent the top-1 (unbalanced) test accuracy of BabyBERTa as a function of its "age". Black lines are degree 3 curves of best fit. Pink lines represent the accuracy that BabyBERTa achieves when trained on the entire training data in an *age-agnostic* manner (Huebner et al., 2021).

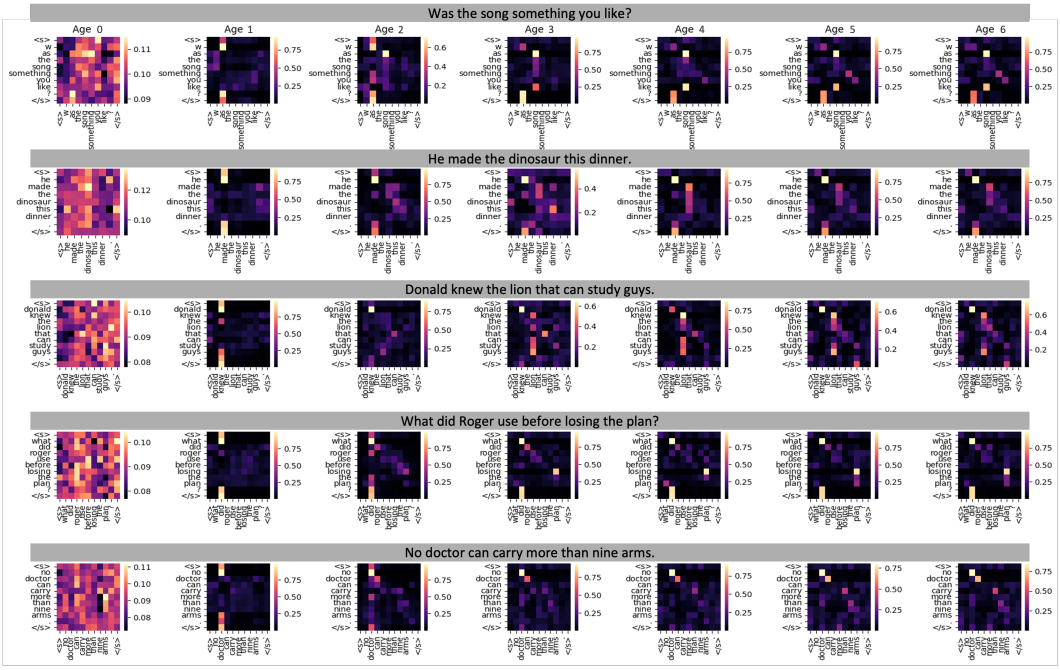

Figure 5: **Attention heatmaps.** For five different sentences, we visualize how BabyBERTa's attention to different parts of the sequences evolves with "age".

(figure 5). At an age of 1, BabyBERTa has the first word (mostly a subject) attend to the word immediately following (mostly the verb), as well as the last word (mostly the punctuation) attend to the start of the sentence. This suggests that the infant is able to infer long-range relationships between sentence components for comprehension. With age, more nuanced relationships such as those between determiners and nouns (for e.g., *this dinner*), verbs and objects (for e.g., *made ... dinosaur*), multiple nouns (for e.g., *lion, guys*), and relative clauses (for e.g., *lion ... study*) are identified. These results show that children, with age, acquire increasingly complex syntactic information that helps them attend to different sentence components for effective comprehension.

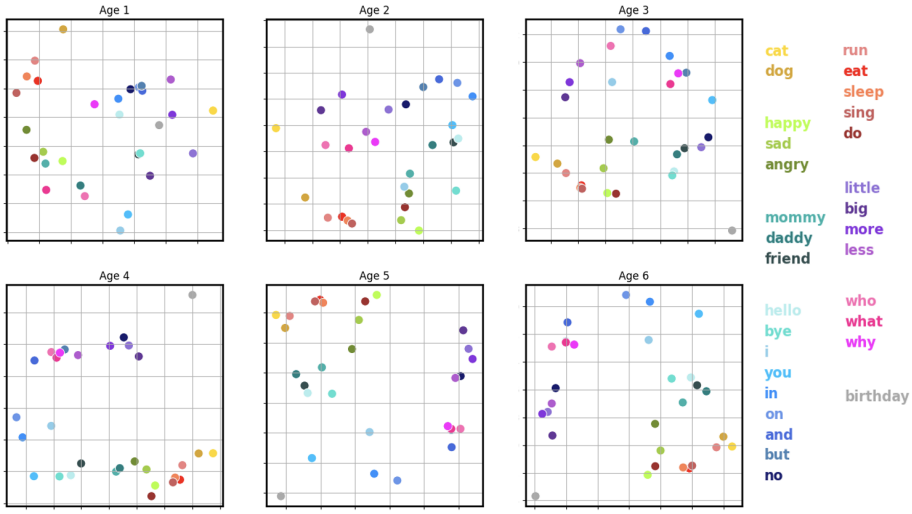

Figure 6: **Textual embedding representations.** Here we visualize how BabyBERTa adapts the way it clusters words and tokens together with "age".

4.4 Children form increasingly meaningful textual representation clusters as they grow, especially since an age of 4

In addition to visualizing attention maps, we also analyzed how textual representation clusters change with age (figure 6). Initially, while the clusters are not entirely distinct, there is some degree of clustering present among verbs (*run*, *eat*, etc.), quantifiers (*little*, *more*, etc.), and emotions (*happy*, *sad*, etc.). A major change from age 1 to 2 is the strong clustering of verbs. Wh- question words (*who*, *what*, etc.) also start coming closer. The noun *birthday* which was initially clustered along with quantifiers and prepositions is now clearly separated out. By the age of 3, animal categories come together. However, a clear clustering of all categories together is seen by the age of 4—a cluster of prepositions and conjunctions (*in*, *on*, *and*, etc.) become visible. And these clusters get further spaced out by the age of 5. The formation of well-defined distinct clusters for different syntactic groups by the age of 4 support the acquisition of most syntactic phenomena in figure 3 and are developmentally plausible.

An interesting observation is that the verb *do* is always clustered with emotions than with the other verbs. Similarly, *mommy*, *daddy*, and *friend* are always clustered with *hello* and *bye*. We think that this might be because of the contexts in which these words often co-appear, with for e.g., the verb *do* being followed by expressions of happiness or sadness based on what the child did.

5 Discussion

In this work, we modeled early language acquisition in children by analyzing both syntactic and semantic learning trajectories with age. We observed that syntactic learning trajectories are developmentally-plausible, underscored by evolving attention patterns of children over different components of sentences during comprehension as well as the formation of increasingly meaningful textual representation clusters. Complex semantic understanding, on the other hand, requires cognitive development beyond the first six years of life. These results not only provide insights into language development but also help in identifying effective learning strategies for children.

Some limitations of our work involve the following modeling assumptions that we make:

Innateness. An interesting yet important consideration that we avoid here is whether language acquisition is innate or not. While most linguists and neurobiologists do not believe in linguistic nativism, it is still important to ask ourselves, especially when trying to create a biologically-plausible model of language acquisition, how best to resolve arguments put forth by Chomsky and the presence

of a universal grammar (Chomsky, 2014), of language localization to Broca's and Wernicke's area Pinker (1994), or of language evolution (Botha, 2021).

Prosody. Here, we model language acquisition purely through speech transcripts (i.e., text). Children during early language acquisition receive stimuli mostly verbally. Prosody—the use of stress and intonations while talking—can encode substantial linguistic information that can play a crucial role in syntactic acquisition (Warstadt and Bowman, 2022). How would language models (or models of audition, more specifically) behave when having to detect word, phrase, and sentence boundaries themselves? Does providing processed input with pre-defined tokens to language models significantly affect how linguisitic skills are acquired?

Second language acquisition. Several studies have previously argued for the presence of a "critical period" beyond which learning of syntactic relationships becomes difficult. These hypotheses, however, talk about second language acquisition (Kuhl, 2004, 2010). The initial coding of native-language patterns due to neural commitments eventually interferes with the learning of new patterns in a foreign language. With extensive research in second language acquisition, can we try to simulate what happens developmentally? How do syntactic relationships vary across the languages? Is there a common system of semantic patterns formed that underlies proficiency in bilingual or multilingual children?

As future work, we might try to analyze the interaction of prosodic and multimodal stimuli in both native and second language acquisition, hoping to unravel more interesting observations and insights into developmental cognition.

6 Ethics Statement

- The primary challenge with a work of this nature is data privacy. Child-directed transcripts of speech might contain sensitive information about the family that, if fallen into an adversary's hands, could lead to bad outcomes in the form of extortion or social embarassment. Specifically, speech transcripts often encode idiosyncratic information about behaviors that a child exhibits, or the relationship that the family sustains. Thorough data cleaning and sensitive-information masking/removal efforts are required. This can be done by anonymizing names by pseudonyms like *Alpha* or *Delta*.
- In relation to the trained model itself, it is important to consider a downstream user's over-reliance on BabyBERTa's ability to acquire language. Since the data that BabyBERTa is trained on comes from humans, there might be inherent stereotypical and discriminatory biases that the model might pick up, being potentially visible in the way it constructs its semantic representation space. This is further amplified by the dataset being small and consisting of transcripts from a handful of families. Quantifying social biases might be one way in which this can be identified (Czarnowska et al., 2021).
- And finally, it is always important to keep in mind providing open-source access to the way we train BabyBERTa. Fei-Fei Li, Percy Liang, and others have been big advocates of open-sourcing research and models,[2] and for others to be able to see the architecture, training data, training method, as well as algorithmic nuances before deploying a model blindly to a downstream task becomes important. We open-source any and all experiments that we carry out for the benefit of future research. Furthermore, having more resources be invested in the development of small-scale models trained on small datasets to control the amount of environmental resources (water, electricity, etc.) we use would need to be advocated for, not just by governments and not-for-profit institutions but also by every one of us.

[2] https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Roy Bar-Haim, Ido Dagan, Iddo Greental, Idan Szpektor, and Moshe Friedman. 2007. Semantic inference at the lexical-syntactic level for textual entailment recognition. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 131–136, Prague. Association for Computational Linguistics.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. *TAC*, 7(8):1.

Rudolf P Botha. 2021. Unravelling the evolution of language. In *Unravelling the Evolution of Language*. Brill.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Thea Cameron-Faulkner, Elena Lieven, and Michael Tomasello. 2003. A construction based analysis of child directed speech. *Cognitive science*, 27(6):843–873.

Noam Chomsky. 2014. *Aspects of the Theory of Syntax*. 11. MIT press.

Eve V Clark and Marisa Casillas. 2015. First language acquisition. In *The Routledge handbook of linguistics*, pages 311–328. Routledge.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, page 177–190, Berlin, Heidelberg. Springer-Verlag.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Gael Gendron, Qiming Bao, Michael Witbrock, and Gillian Dobbie. 2024. Large language models are not strong abstract reasoners.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

Jean Berko Gleason and Nan Bernstein Ratner. 2022. *The development of language*. Plural Publishing.

Maria Teresa Guasti. 2017. *Language acquisition: The growth of grammar*. MIT press.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, 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, Online. Association for Computational Linguistics.

Philip A Huebner and Jon A Willits. 2021. Using lexical context to discover the noun category: Younger children have it easier. In *Psychology of learning and motivation*, volume 75, pages 279–331. Elsevier.

Patricia K Kuhl. 2000. A new view of language acquisition. *Proceedings of the National Academy of Sciences*, 97(22):11850–11857.

Patricia K Kuhl. 2004. Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, 5(11):831–843.

Patricia K Kuhl. 2010. Brain mechanisms in early language acquisition. *Neuron*, 67(5):713–727.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Barbara C Lust. 2006. *Child language: Acquisition and growth*. Cambridge University Press.

Letitia R Naigles and Erika Hoff-Ginsberg. 1998. Why are some verbs learned before other verbs? effects of input frequency and structure on children's early verb use. *Journal of child language*, 25(1):95–120.

Steven Pinker. 1994. How could a child use verb syntax to learn verb semantics? *Lingua*, 92:377–410.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Kuniyoshi L Sakai. 2005. Language acquisition and brain development. *Science*, 310(5749):815–819.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Zhengwei Tao, Zhi Jin, Xiaoying Bai, Haiyan Zhao, Yanlin Feng, Jia Li, and Wenpeng Hu. 2023. Eveval: A comprehensive evaluation of event semantics for large language models. *arXiv preprint arXiv:2305.15268*.

Héctor Vázquez Martínez, Annika Lea Heuser, Charles Yang, and Jordan Kodner. 2023. Evaluating neural language models as cognitive models of language acquisition. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 48–64, Singapore. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023. Evaluating open-QA evaluation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4144–4150. AAAI Press.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language*, pages 17–60. CRC Press.

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

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2023. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023. Are large language models really good logical reasoners? a comprehensive evaluation from deductive, inductive and abductive views. *arXiv preprint arXiv:2306.09841*.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2023. GLUE-X: Evaluating natural language understanding models from an out-of-distribution generalization perspective. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12731–12750, Toronto, Canada. Association for Computational Linguistics.

Aditya Yedetore, Tal Linzen, Robert Frank, and R. Thomas McCoy. 2023. 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)*, pages 9370–9393, Toronto, Canada. Association for Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

A Datasets

A.1 Zorro

The Zorro dataset evaluates language models on the syntactic phenomena presented in table 1.

A.2 GLUE

From the GLUE benchmark we use the following tasks:

Corpus of Linguistic Acceptability (CoLA): This task (Warstadt et al., 2019) asks an evaluatee to determine if an English sentence is grammatical. The metric used is Matthews correlation coefficient, with values $\in [-1, 1]$.

Stanford Sentiment Treebank (SST-2): This task (Socher et al., 2013) requires evaluatees to predict the sentiment (positive or negative) of sentences from movie reviews and human annotations.

Microsoft Research Paraphrase Corpus (MRPC): This (Dolan and Brockett, 2005) is a corpus of sentence pairs from online news sources, asking evaluatees to determine whether the sentences in the pair are semantically equivalent.

Quora Question Pairs (QQP): This task (Wang et al., 2017) consists of a collection of question pairs from Quora for determining whether a pair of questions are semantically equivalent.

Question Natural Language Inference (QNLI): This task (Rajpurkar et al., 2016) consists of a question-answering dataset with question-paragraph pairs, requiring to determine if the context sentence contains the answer to the question.

Recognizing Textual Entailment (RTE): This task combines data from RTE1 (Dagan et al., 2005), RTE2 (Bar-Haim et al., 2007), RTE3 (Giampiccolo et al., 2007), and RTE5 (Bentivogli et al., 2009) to pose whether the two sentences provided demonstrate entailment.

Winograd Natural Language Inference (WNLI): This (Levesque et al., 2012) is a reading comprehension task in which an evaluatee is presented with a sentence pair and must determine if the sentence with a pronoun substituted by a referent is entailed by the original sentence.

B Code Availability

We release all code under the MIT license at `https://github.com/ynshah3/LanguageAcquisition`.

Table 1: **Zorro evaluation phenomena:** example positive and negative sentences.

Phenomenon	Negative	Positive
Determiner-noun agreement across 1 adjective	Look at this purple things.	Look at this purple thing.
Determiner-noun agreement between neighbors	This colors must be white.	This color must be white.
Subject-verb agreement across prepositional phrase	The lies on the foot is flat.	The lie on the foot is flat.
Subject-verb agreement across relative clause	The books that I like is poor.	The book that I like is poor.
Subject-verb agreement in question with auxilliary	Where does the horses go?	Where does the horse go?
Subject-verb agreement in simple question	Where is the ways?	Where is the way?
Anaphora agreement	Will Mark want herself?	Will Mark want himself?
Dropped argument	The poor boat gives me.	Give me the poor boat.
Swapped argument	The wolf made her ball they.	They made the wolf her ball.
Transitive	Philip affected.	Philip thinks.
Binding: principle A	Ben thinks about himself called this fuel.	Ben thinks about himself calling this fuel.
Case: subjective pronouns	The wolf brought I my hill.	I brought the wolf my hill.
Ellipsis: N-bar	A sister visited one whale and Roger visited three little/	A sister visited one little whale and Roger visited three.
Filler-gap: why question object	Sarah knew what the general knew the doctor.	Sarah knew the general that the doctor knew.
Filler-gap: why question subject	Laura ended who the finger can make boats.	Laura ended the finger that can make boats.
Irregular verb	Will grown quickly at home.	Will grew quickly at home.
Island effects: adjunct	Who should William have the baby without watching?	Who should William have without watching the baby?
Island effects: corrdinate structure constraint	Who must Philip turn and the dinosaur.	Who must Philip and the dinosaur turn?
Local attractor in question with auxilliary	Is the whale gets the person?	Is the whale getting the person?
NPI licensing: matrix question	Her boat does ever play with the growth?	Does her boat ever play with the growth?
NPI licensing: only licensor	Even Mark ever finds some suit.	Only Mark ever finds some suit.
Quantifiers: existential there	There are most books about soft birds.	There are many books about soft birds.
Quantifiers: superlative	No pig could stand on top of at least six days.	No pig could stand on top of more than six days.