A Distributional Perspective on Word Learning in Neural Language Models

Filippo Ficarra [1] **Ryan Cotterell** [1] **Alex Warstadt** [1]
[1]ETH Zürich
{fficarra, rcotterell, warstadt}@ethz.ch

Abstract

Language models (LMs) are increasingly studied as models of human language learners. Due to the nascency of the field, it is not well-established whether LMs exhibit similar learning dynamics to humans, and there are few direct comparisons between learning trajectories in humans and models. Word learning trajectories for children are relatively well-documented, and recent work has tried to extend these investigations to language models. However, there are no widely agreed-upon metrics for word learning in language models. We take a distributional approach to this problem, defining lexical knowledge in terms of properties of the learned distribution for a target word. We argue that distributional signatures studied in prior work fail to capture key distributional information. Thus, we propose an array of signatures that improve on earlier approaches by capturing knowledge of both where the target word can and cannot occur as well as gradient preferences about the word's appropriateness. We obtain learning trajectories for a selection of small language models we train from scratch, study the relationship between different distributional signatures, compare how well they align with human word learning trajectories and interpretable lexical features, and address basic methodological questions about estimating these distributional signatures. Our metrics largely capture complementary information, suggesting that it is important not to rely on a single metric. However, across all metrics language models' learning trajectories fail to correlate with children's.

FilippoFicarra/word_learning

1 Introduction

There is a long tradition of characterizing words in terms of their distributions (Wittgenstein, 1953). The **distributional hypothesis** (Harris, 1954; Lenci, 2008), which characterizes knowledge of a word in terms of "the company it keeps" (Firth, 1957), has proven surprisingly prescient. This is the idea behind static word representations (Deerwester et al., 1990; Landauer and Dumais, 1997; Hofmann, 1999; Mikolov et al., 2013; Pennington et al., 2014) estimated from data as well as modern (large) language models (OpenAI, 2022; Meta, 2024). While this distributional approach to training language models (LMs) is now well-established, only recently has distributional information been explored as a tool for *evaluating* lexical knowledge in LMs.

In the last few years, there has been growing interest in studying word learning in language models (Nikolaus and Fourtassi, 2021a; Chang and Bergen, 2022; Portelance et al., 2024, 2023; Vong et al., 2024; Zhuang et al., 2024b,a; Ma et al., 2024). Most of these studies are part of a larger research program to use LMs to inform the study of human language acquisition by serving as convenient, controllable, and effective models of human development (Dupoux, 2018; Linzen, 2019; Warstadt and Bowman, 2022). From this perspective, it is desirable to have LMs with human-like learning trajectories, as they can better serve as generalizable models of human learners. Word learning has a potentially important role in the success of this research program because it is one of the best proving grounds for comparing the learning trajectories of humans and LMs head to head. While some studies (e.g., Choshen et al., 2022) have tracked syntax learning in LMs using benchmarks like BLiMP (Warstadt et al., 2020), corresponding data for children is more limited in scope (Evanson et al., 2023). There is also child data on phonological learning (Lavechin et al., 2022) which can be explored further as audio-based LMs improve.

Fortuitously, word learning trajectories in text-based LMs can be easily compared against a wealth of child data in multiple languages thanks to the massive efforts of caregivers and scholars who report and curate child word learning data in the Wordbank database (Frank et al., 2017). Unfortunately, the caregiver reporting approach (Fenson et al., 2013) used in Wordbank is not immediately applicable to LMs, and there is no consensus on how to benchmark word learning in LMs. Zhuang et al. (2024b) explored word learning through different methods, including comparing LMs' word similarity scores to humans' (Finkelstein et al., 2001; Bruni et al., 2012; Hill et al., 2015; Gerz et al., 2016), classifying lexical entailment relations (Santus et al., 2016), predicting semantic features (Buchanan et al., 2019) and using minimal pairs to measure LM preferences for appropriate word usage (Marvin and Linzen, 2018). Other works rely on visual stimuli to ground evaluations for multimodal models (Nikolaus and Fourtassi, 2021a; Berger et al., 2022; Vong et al., 2024). Notably, Chang and Bergen (2022) and Portelance et al. (2023) take a distributional approach, characterizing lexical knowledge in terms of the LM's surprisal, an information-theoretic quantity which has been widely studied in psycholinguistics (Hale, 2001; Levy, 2008).

In this study, we take inspiration from Chang and Bergen's (2022) approach to tracking throughout training the **distributional signatures** of word learning, i.e., a metric characterizing for a single point in training the model's distributional knowledge about a particular word. We formalize their approach and build on it in several respects. Chang and Bergen consider only LMs' surprisal in a context where the target word is appropriate, and (implicitly) rely on a trivial approximation of the ground truth distribution in evaluating the quality of lexical knowledge. In contrast, we propose a family of distributional signatures allowing for the consideration of the LM's learned distribution in both appropriate and inappropriate contexts. We also introduce distributional signatures that are truly intrinsic to the model itself as well as strongly reference signatures that compare the learned distribution to a non-trivial ground truth, which we approximate using a large pretrained LM.

In our experiments, we train language models from scratch on three datasets resembling the input to children to varying degrees. We record the distributional signatures for a set of common words throughout training, and following Chang and Bergen we apply a threshold to the measured learning trajectories to obtain an age-of-acquisition (AoA) for each word. We then conduct analyses to answer the following questions:

1. Which methods allow us to reliably extract AoA scores?
2. How does the order of word acquisition in LMs compare to that of children?
3. What are the empirical properties of the learning trajectories for different distributional signatures?

We find that the learning trajectories for different distributional signatures are indeed different from each other, suggesting that earlier approaches failed to capture some aspects of word learning. While many signatures, like Chang and Bergen's (2022), give trajectories that are highly correlated with simple features like lexical frequency, other signatures are harder to predict and therefore may capture more nontrivial information. However, we find that learning trajectories for some distributional signatures fail to converge, making AoAs difficult to infer. Finally, no signature yields AoA scores that are strongly correlated with children's AoA, supporting the conclusion that with current methods, LMs' learning patterns are poorly aligned with humans', and underscoring a limitation of current LMs as models of human development. We therefore call for future work to evaluate and improve the human-likeness of LMs' learning trajectories using the distributional signatures we propose.

2 Preliminaries

Let Σ be an **alphabet**, a finite, non-empty set, of characters, e.g., Unicode symbols.[1] A **string** is a finite sequence of characters, drawn from an alphabet Σ. The set Σ^*, the Kleene closure of Σ, is the set of all strings with characters drawn from Σ including the empty string ε. We consider two distinguished types of strings. First, we define a **word**[2] as a character string $\boldsymbol{w} \in \Sigma^*$, which is believed to operate as a lexical item. Second, we call an arbitrary character string that precedes a word a **context**. We denote a context as $\boldsymbol{c} \in \Sigma^*$.

A **language model** p is a probability distribution over Σ^*. A language model's **prefix probability** is defined as the following sum

$$\overrightarrow{p}(\boldsymbol{y}) \stackrel{\text{def}}{=} \sum_{\boldsymbol{y}' \in \Sigma^*} p(\boldsymbol{y}\boldsymbol{y}'). \tag{1}$$

[1]Note that most modern language models operate over tokens, rather than over characters. Our presentation is in terms of characters for simplicity.

[2]Defining a word is a complex matter, and we concede to not having down it justice in this article. As a simple example, consider the English verb *to run*. If a child says, *I ran*, we probably think they used *to run* and this should, ideally, be taken into account in our framework. Yet, under our current set-up, we are not able to account for such inflection. Moreover, even beyond morphological inflection, it is hard to define *in se*; see Marantz (2001) for a longer discussion.

Throughout the paper, we will primarily be interested in a specific ratio of p's prefix probabilities, which we will use to define the probability of a word in a context:

$$\overrightarrow{p}(\boldsymbol{w} \mid \boldsymbol{c}) \stackrel{\text{def}}{=} \frac{\overrightarrow{p}(\boldsymbol{cw})}{\overrightarrow{p}(\boldsymbol{c})}, \tag{2}$$

i.e., the probability of a word given a context.[3] We call Eq. (2) a language model p's **word distribution**. We are also interested in the surprisal of a word in a context, denoted $-\log \overrightarrow{p}(\boldsymbol{w} \mid \boldsymbol{c})$.

Now we derive a language model p's **context distribution** by Bayes' rule as follows

$$\overrightarrow{p_\kappa}(\boldsymbol{c} \mid \boldsymbol{w}) = \frac{\overrightarrow{p}(\boldsymbol{w} \mid \boldsymbol{c})\overrightarrow{p}(\boldsymbol{c})}{\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p}(\boldsymbol{w} \mid \boldsymbol{c})\overrightarrow{p}(\boldsymbol{c})}. \tag{3}$$

Under the assumption that p is of finite expected length, then $\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p}(\boldsymbol{w} \mid \boldsymbol{c})\overrightarrow{p}(\boldsymbol{c})$ is always finite (Opedal et al., 2024, see Section 2.1).

Complementarily, we define a word $\boldsymbol{w}$'s **negative context distribution** as

$$\overrightarrow{p_\kappa}(\boldsymbol{c} \mid \neg\boldsymbol{w}) = \frac{(1-\overrightarrow{p}(\boldsymbol{w} \mid \boldsymbol{c}))\overrightarrow{p}(\boldsymbol{c})}{\sum_{\boldsymbol{c}\in\Sigma^*}(1-\overrightarrow{p}(\boldsymbol{w} \mid \boldsymbol{c}))\overrightarrow{p}(\boldsymbol{c})}. \tag{4}$$

In the remainder of this paper, we will distinguish three LMs: p, the underlying distribution that we take to have generated the observed strings; q, a parameterized model whose parameters we estimate; and r, a pre-trained reference LM, potentially larger and trained on more data. A standard method of constructing a language model p that approximates q is maximum-likelihood estimation. Suppose we observe a bag of N samples $\wr \boldsymbol{y}^{(n)} \wr_{n=1}^{N}$ where $\boldsymbol{y}^{(n)} \sim p$, then we choose a model q that minimizes the following cross-entropy $-\sum_{n=1}^{N} \log q(\boldsymbol{y}^{(n)})$.

3 Defining Lexical Knowledge

Our goal is to evaluate word learning in LMs by following the **trajectory** throughout training the LM of a distributional signature for each target word. However, both in terms of trajectory extraction and signature design, there are many design choices. In this section, we explore and discuss the implications of a range of choice points in defining the distributional signature that is tracked during training. In §4, we discuss how to extract a trajectory from a timestamped sequence of distributional signatures.

[3] In our formalism, $\overrightarrow{p}(\boldsymbol{w} \mid \boldsymbol{c})$ is not a probability distribution over words. Rather, $\overrightarrow{p}(\boldsymbol{w} \mid \boldsymbol{c})$ is simply the probability of the character string $\boldsymbol{w}$ following $\boldsymbol{c}$.

Chang and Bergen (2022). The most direct predecessor to this work, Chang and Bergen (2022), considered a single distributional signature: the surprisal under the LM of the target words in contexts where the word occurs in a test corpus. This is a natural quantity to track during training, as it is equivalent to the cross entropy loss per token restricted to only samples from a single class. In our notation, they consider

$$\widehat{\sigma}_+(\boldsymbol{w}) \stackrel{\text{def}}{=} -\frac{1}{M}\sum_{m=1}^{M} \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c}^{(m)}), \tag{5}$$

where $\overrightarrow{q}$ is the LM we are analyzing and the contexts $\boldsymbol{c}^{(m)}$ are contexts taken from a corpus that occur *before* the word $\boldsymbol{w}$, which we refer to as **positive contexts** for $\boldsymbol{w}$. We observe that—under the assumption that the positive contexts are sampled from the ground true context distribution, i.e., $\boldsymbol{c}^{(m)} \sim \overrightarrow{p_\kappa}(\cdot \mid \boldsymbol{w})$—Eq. (5) is a Monte Carlo estimator of the quantity

$$\sigma_+(\boldsymbol{w}) \stackrel{\text{def}}{=} -\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c} \mid \boldsymbol{w}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c}). \tag{6}$$

However, even in expectation, there is one salient manner in which Chang and Bergen's (2022) distributional signature misses potentially valuable distributional information about the target word: it fails to consider the LM's distributional knowledge about $\boldsymbol{w}$ in **negative contexts** where $\boldsymbol{w}$ is not found. Moreover, beyond this limitation, this distributional signature is the only element in a potentially very large design space; in the remainder of this section, we also explore additional distributional signatures.

Considering negative contexts. Knowing the distribution of $\boldsymbol{w}$ requires not just knowing when the word is appropriate in context, but also when it is *in*appropriate. Thus, we can instead study the LM's distribution in contexts sampled according to a word's negative context distribution $\overrightarrow{p_\kappa}(\cdot \mid \neg\boldsymbol{w})$, i.e., the context distribution over all those contexts that occur before a word that is *not* $\boldsymbol{w}$ and does not have $\boldsymbol{w}$ as a prefix; see §2. Thus, analogously to Eq. (6), we define the following distribtional signature:

$$\sigma_-(\boldsymbol{w}) \stackrel{\text{def}}{=} -\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c} \mid \neg\boldsymbol{w}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c}). \tag{7}$$

Again, under the assumption that negative contexts were sampled from p, i.e., $\boldsymbol{c}^{(m)} \sim \overrightarrow{p_\kappa}(\cdot \mid \neg\boldsymbol{w})$, we

	Positive	Negative	All
True	$-\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c} \mid \boldsymbol{w}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})$	$-\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c} \mid \neg\boldsymbol{w}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})$	$-\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})$
Intrinsic	$-\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{q_\kappa}(\boldsymbol{c} \mid \boldsymbol{w}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})$	$-\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{q_\kappa}(\boldsymbol{c} \mid \neg\boldsymbol{w}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})$	$-\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{q_\kappa}(\boldsymbol{c}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})$
Reference	$\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c} \mid \boldsymbol{w}) \left\lvert \log \frac{\overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})}{\overrightarrow{r}(\boldsymbol{w} \mid \boldsymbol{c})} \right\rvert$	$\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c} \mid \neg\boldsymbol{w}) \left\lvert \log \frac{\overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})}{\overrightarrow{r}(\boldsymbol{w} \mid \boldsymbol{c})} \right\rvert$	$\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c}) \left\lvert \log \frac{\overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})}{\overrightarrow{r}(\boldsymbol{w} \mid \boldsymbol{c})} \right\rvert$

Table 1: Overview of all distributional signatures measured in our experiments.

can derive a Monte Carlo estimator as follows

$$\widehat{\sigma}_-(\boldsymbol{w}) \stackrel{\text{def}}{=} -\frac{1}{M}\sum_{m=1}^{M} \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c}^{(m)}). \quad (8)$$

All contexts. Rounding out this series of related signatures, we design a distributional signature that considers the LM's predictions in all—both positive and negative—contexts

$$\sigma_\pm(\boldsymbol{w}) \stackrel{\text{def}}{=} -\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c}), \quad (9)$$

where $\overrightarrow{p_\kappa}(\cdot)$ is the unconditional distribution over contexts. Analogously, we arrive at the following Monte Carlo estimator where $\boldsymbol{c}^{(m)} \sim \overrightarrow{p_\kappa}(\cdot)$

$$\widehat{\sigma}_\pm(\boldsymbol{w}) \stackrel{\text{def}}{=} -\frac{1}{M}\sum_{m=1}^{M} \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c}^{(m)}). \quad (10)$$

Intrinsic Signatures. We now discuss a different class of distributional signatures. Rather than taking the expectation with respect to the true context distribution $\overrightarrow{p_\kappa}(\cdot \mid \boldsymbol{w})$, we consider an **intrinsic signature** where we take the expectation with respect to the model $\overrightarrow{q_\kappa}(\cdot \mid \boldsymbol{w})$. This yields three distributional signatures analogous to those above which are defined in the second row in Table 1, which we term $\sigma_{\mathrm{I}+}$, $\sigma_{\mathrm{I}-}$, and $\sigma_{\mathrm{I}\pm}$, respectively. We discuss estimating the intrinsic distribution signatures in App. A.

Comparing to a reference distribution. Additionally, the signatures σ_+, σ_-, and $\sigma_\pm$ estimate the relationship between the model and the underlying distribution p. However, the true language model p may not be achievable—both due to the finite training data or the model class itself. Thus, it also makes sense to compare q to **reference distribution** r, taken to be a larger LM trained on more data. Following this intuition, we define three **reference signatures**, listed in the third row of Table 1, and denote them as $\sigma_{\mathrm{R}+}$, $\sigma_{\mathrm{R}-}$ and $\sigma_{\mathrm{R}\pm}$, respectively. It is easily seen that the reference signatures are themselves distance metrics between the target model and the reference distribution.

4 Analyzing Trajectories

Given our goal of studying the word acquisition *process* of LMs, we would like to study the **trajectory** of a signature σ for various words throughout the training of the target LM. However, an entire trajectory may be too much information for some analyses. In this section, we consider a family of statistics that can be extracted from the trajectory and review the main choice points in doing so.

Determining AoA by Thresholding While many statistics are possible, we focus on **age of acquisition (AoA)**, which is a single number that should be interpreted as the point at which learning has advanced to a satisfactory degree. For human learners, Braginsky et al. (2016) take AoA as the age when 50% of children are such that their caregivers report them as understanding the word. Chang and Bergen (2022) apply this thresholding approach to LMs. Given a trajectory, they take the AoA to be the first time step at which the signature reaches a threshold defined as τ% of the way between some initial value representing the beginning of learning and some final value representing the endpoint of learning.[4] Unfortunately, thresholding in this way is only appropriate when $\widehat{\sigma}$ changes (roughly) monotonically across time steps. While this is true of some signatures we consider, we find empirically that $\widehat{\sigma}_+$, $\widehat{\sigma}_{\mathrm{I}\pm}$, $\widehat{\sigma}_{\mathrm{I}+}$ and $\widehat{\sigma}_{\mathrm{I}-}$ are exceptions. Thus, we adopt a different

[4]This comes with several choice points: Chang and Bergen explore a range of values for τ and report little change in qualitative results, while Ma et al. (2024) do observe important differences due to the choice of τ. In defining the initial and final values, a naive approach would be choosing the first and last values. Chang and Bergen (2022) select the initial value as the surprisal under a random chance baseline.

approach to extracting AoAs based on the notion of a **Cauchy sequence**. Intuitively, we say that the target word is learned at the point in the trajectory where the value of the signature becomes close to its neighboring points in the trajectory. Our approach is defined formally in App. D. For the sake of uniformity, we apply this approach to all signatures and leave an exploration of thresholding approaches for suitable signatures to future work.

Idealizing the Trajectory Empirical trajectories may be noisy due to estimation errors from a test corpus or local instabilities during training. We consider several techniques to idealize the trajectory and reduce noise before thresholding. One approach to idealize the trajectory and reduce noise before thresholding is parametric curve fitting. Braginsky et al. (2019) and Chang and Bergen (2022) assume trajectories form a sigmoid curve (for LMs, the x-axis should be the log of the number of steps). Outside of the context of word learning, Zhang et al. (2021) fit exponential learning curve inspired by psychometrics literature on learning to LM results (Heathcote et al., 2000; Leibowitz et al., 2010). However, this approach only applies if the typical shape of the trajectory is known. If this is not the case, one can instead smooth the curve with a filter such as a moving average or an overparameterized curve such as a generalized additive model (GAM) (Hastie and Tibshirani, 1986), as done by Chang et al. (2024). This approach has the benefits of preserving the shape of the curve with high fidelity and can be applied to curves of any shape. A disadvantage is that it does not guarantee that a threshold or convergence point is reached. As several of our signatures give trajectories that do not follow a consistent shape, we opt to apply a moving average to smooth the trajectories.

5 Methods

5.1 Language Models

We train several language models to analyze our proposed notions of word learning.

Training Data. We use three datasets previously released with train/test splits for training and evaluating our LMs: (i) **Unified**: This dataset was compiled by Constantinescu et al. (2024). It consists of approximately 600M words sampled from a combination of three corpora: Project Gutenberg,[5] Wikipedia, and OpenSubtitles (Lison and Tiedemann, 2016). Given that a typical 13-year-old person may be exposed to around 100M words (Gilkerson et al., 2017), this dataset is not as representative of the actual input to children, although it contains a large proportion of spoken language. (ii) **BabyLM**: This is the 100M text-only corpus from the second BabyLM Challenge (Choshen et al., 2024). The dataset is designed to be relatively developmentally plausible while also containing the amount of input that a typical adolescent is exposed to. It includes child-directed speech from CHILDES (MacWhinney, 2000) and children's stories from Project Gutenberg (Gerlach and Font-Clos, 2020), as well as dialogue such as BNC and the Switchboard Corpus (Stolcke et al., 2000), along with Simple English Wikipedia and Open Subtitles. (iii) **CHILDES**: This is the CHILDES subset taken from BabyLM consisting of 29M tokens of child-directed speech.

These datasets balance developmental plausibility against quantity. Our motivation for training on datasets like BabyLM and CHILDES is to observe whether more developmentally plausible training distributions result in more human-like word-learning trajectories.

To estimate the signatures for each word, we sample 100 positive and 100 negative contexts from the BabyLM test set. We use the same test contexts for all models regardless of training data in order to make cross-model comparisons more fair.

Models. We train GPT-2 from scratch following the training procedure given by Radford et al. (2019). To reduce variance in performance due to random seed, we train three variations of each model using different random seeds. To compute the reference signatures ($\widehat{\sigma}_{\mathrm{R}+}$, $\widehat{\sigma}_{\mathrm{R}-}$, $\widehat{\sigma}_{\mathrm{R}\pm}$) we use Llama-3.1-8B[6] as the reference distribution r.

Full details regarding the hyperparameters, training duration, and loss curves can be found in App. C. As we are interested in analyzing the learning trajectories for models, it is important that they are trained for a reasonable duration. For models trained on BabyLM and CHILDES we apply early stopping, i.e., we choose the best model on a held-out development set, as we found that models overfit eventually. For models trained on Unified we train for 30,000 steps, or 12 epochs, following (Constantinescu et al., 2024). We estimate that Chang and Bergen (2022) train models on about

[5] https://gutenberg.org

[6] https://ai.meta.com/blog/meta-llama-3-1/

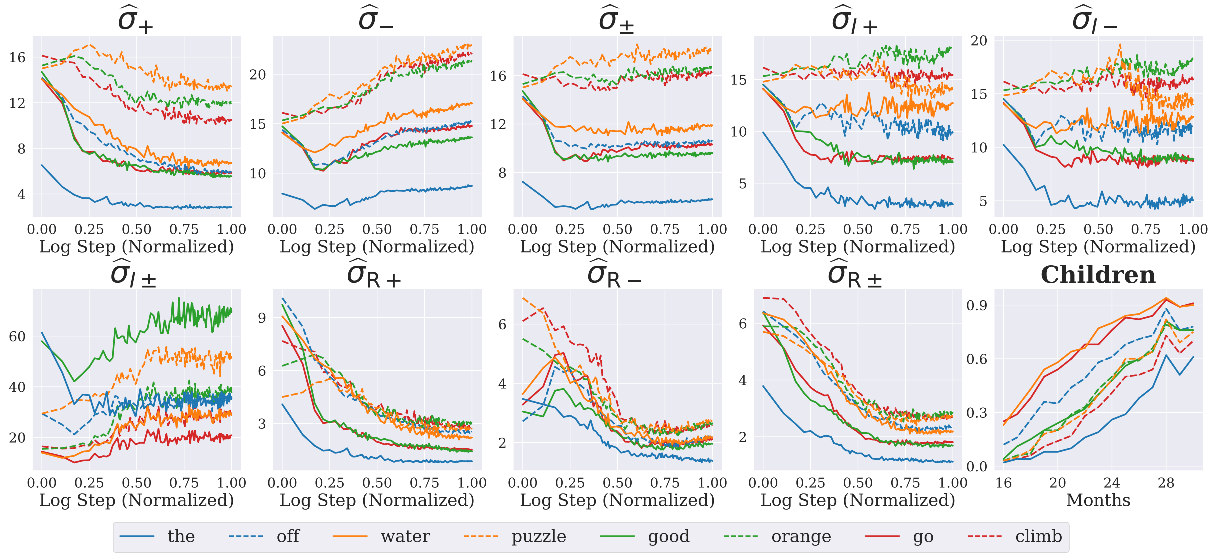


Figure 1: Trajectories for a sample of 8 words for LMs trained on the Unified dataset. We sample one high-frequency (solid line) and one low-frequency (dashed) word from each of four categories: function word, noun, adjective, verb. The y-axis represents the value of the estimator in all σ plots, while in the children plot, it represents the proportion of children who produced the word.

1.6×10^9 input tokens (counting repetitions).[7] By comparison, our models were trained for a duration of between 6.8×10^8 and 7.8×10^{10} tokens.

5.2 The Wordbank Corpus

Child *AoA* data comes from the North American English portion of the Wordbank database (Frank et al., 2017). For each word an month, Wordbank gives the proportion of children in the study that have produced the word by that point. The *AoA* is the first month by which at least 50% children have produced that word (Goodman et al., 2008; Braginsky et al., 2016). We remove words for which we were not able to sample 100 positive context types from the BabyLM dataset, leaving us with 305 words. The words in Wordbank were divided in 4 different **lexical categories**: NOUNS (101), PREDICATES (124), FUNCTION WORDS (45) and OTHER (49). Words in the category OTHER have inconsistent properties,[8] so we exclude them from some analyses, giving us 262 words.

6 Examining LM Learning Trajectories

Before quantitatively comparing LM and child word learning trajectories in §7, we perform several analyses focusing on LM trajectories alone.

[7]This estimate is based on the reported 100k steps with a batch size of 128 and a context window of 128.

[8]For example *babysitter, doctor, brother, grandma* were categorized as OTHER while they clearly belong to the category NOUNS.

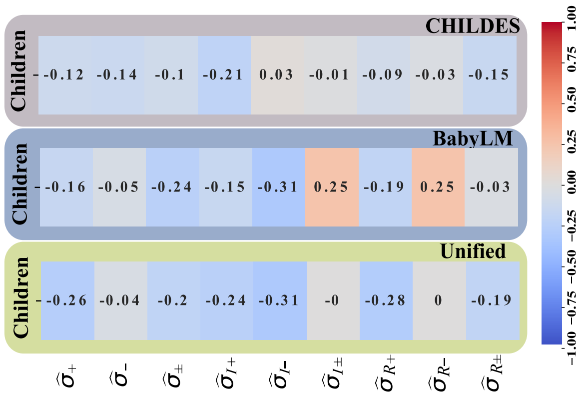


Figure 2: Correlation coefficients between the different signatures and and Children's *AoA* (C) across three datasets: Childes, BabyLM, and Unified.

6.1 Case studies

We perform several case studies by inspecting the learning trajectories and *AoA* scores for humans and each distributional signature from §3. We analyze the trajectories and *AoA* scores for LMs trained on the Unified dataset for a sample of 8 words: two FUNCTION WORDS, two NOUNS, two ADJECTIVES, and two VERBS. For each category, one word is chosen from the 10 most and 10 least frequent (with respect to the Unified dataset).

Fig. 1 shows the trajectories for these words, and Table A5 gives the *AoA* scores. For most signatures, we observe that the higher-frequency word from a category has an earlier *AoA* than the corresponding lower-frequency word. We also observe that most signatures yield a wide range of AoA scores, but others—particularly $\widehat{\sigma}_{-}$—show very similar (and late) AoAs for all words we inspect. Table A6 shows the first and last learned words for each signature. Generally, we find that high-frequency words and function words are learned first.

6.2 Convergence behavior

As we rely on convergence to extract AoA scores, we now examine how different signatures converge. Fig. 1 shows that the shape of the learning trajectories varies between signatures. Within a given signature, trajectory shapes are internally consistent to varying degrees. As expected, the reference signatures are mostly monotonic decreasing, indicating that the LMs' learned word distributions become closer to ground truth. Furthermore, for the corpus-based signatures, $\widehat{\sigma}_{+}$ trajectories are decreasing, and $\widehat{\sigma}_{-}$ are increasing.[9] On the other hand, the intrinsic signatures and $\widehat{\sigma}_{\pm}$ are not consistently increasing or decreasing.

Fig. A4 show how many words failed to converge under different methodologies. We find the vast majority of trajectories converge with $\epsilon = 0.15$. For lower values of ϵ, we see as many as half of all words' trajectories failing to converge for the CHILDES dataset. However, with the larger datasets BabyLM and Unified, we see high rates of convergence across the board. Finally, it is the intrinsic signatures and $\widehat{\sigma}_{\pm}$ that show the lowest rates of convergence. As discussed above, these are precisely the same signatures that do not have an internally consistent shape.

Finally, we consider whether the convergence and thresholding approaches to extracting AoAs give similar results. We compute the AoA scores for a given signature using a range of values of ϵ. Then, the figures in App. I show correlations for each pair of thresholds. With a few exceptions for extreme values, different thresholds still yield AoA scores that are highly correlated. Therefore, in all our results (including those discussed above) we apply an intermediate value of $\epsilon = 0.07$.

[9]One caveat: for the first few training steps, the trajectory sometimes goes in the other direction. Chang and Bergen (2022) observed this phenomenon, showing that the learned distribution approximates a uniform distribution initially followed by the unigram distribution. After this point, the trajectories are largely monotonic.

6.3 Comparing Signatures

Another important question is whether different signatures give similar AoA scores to each other. App. J shows the correlation matrix of AoA scores for each signature. First, the correlations are all notably higher for LMs trained on the Unified dataset. Together with the finding that convergence rates are higher for this dataset, this supports the conclusion that AoA scores become more consistent as training time increases. We find that most pairs of signatures are weakly or negatively correlated, with a few exceptions. In general, the various positive signatures ($\widehat{\sigma}_{+}$, $\widehat{\sigma}_{\mathrm{I}+}$, $\widehat{\sigma}_{\mathrm{R}+}$) have relatively strong correlations. Across all datasets, the most strongly correlated pair is $\widehat{\sigma}_{+}$ and $\widehat{\sigma}_{\mathrm{R}+}$. The negative signatures, but $\widehat{\sigma}_{\mathrm{I}-}$, have weak correlations with other signatures, except for the pair $\widehat{\sigma}_{-}$, $\widehat{\sigma}_{\pm}$ which have very similar estimators.

7 Human vs. LM Learning Trajectories

We now examine the similarities and differences between word learning in LMs and humans.

7.1 Comparing Human and LM $AoAs$

We begin simply by measuring the Pearson correlation between human AoAs and the LM AoAs from each signature. These values are plotted in Fig. 2. Overall, we see very weak or negative correlations. We find that the signature that correlates most changes depending on the datasets, but no correlation exceeds 0.31 (both positive and negative). The strongest positive correlations are from BabyLM for $\widehat{\sigma}_{\mathrm{I}\pm}$ and $\widehat{\sigma}_{\mathrm{R}-}$, while the strongest negative correlation are from Unified and BabyLM for the $\widehat{\sigma}_{\mathrm{I}-}$ signature.

7.2 Predicting $AoAs$ from Features

We now examine which factors predict human and LM AoAs, and compare whether these factors have similar effects. Braginsky et al. (2016) identify several interpretable features that predict human AoAs. Chang and Bergen (2022) previously fit linear models to predict LM AoAs using these features. We extend this analysis to our set of signatures.

Specifically, we take each the AoA scores from children and from each signature as our dependent variable, and additionally consider the following predictors[10] that Braginsky et al. (2016)

[10]The outcomes of regressions using single predictors can be misleading due to correlations among the predictors. Therefore, regressions with multiple predictors have been conducted as shown in the last two columns of Table 2.

Metadata		Single-Predictor					Multi-Predictor	
***AoA* type**	**#words**	**Log freq.**	**Conc.**	**#chars**	**MLU**	**Lex. cat.**	**Full**	**Full $\setminus$ Log freq.**
Children	*262*	*0.004*	*0.26*	*-0.003*	*0.032*	*0.20*	***0.417***	*0.363*
$\widehat{\sigma}_{+}$	262	**0.614**	**0.298**	0.135	0.072	**0.304**	**0.616**	**0.392**
$\widehat{\sigma}_{-}$	251	0.063	0.012	0.023	0.0	0.047	0.083	0.065
$\widehat{\sigma}_{\pm}$	245	0.542	0.265	**0.142**	0.04	0.294	0.546	0.379
$\widehat{\sigma}_{\mathrm{I}+}$	215	0.4	0.274	0.107	**0.162**	0.201	0.463	0.382
$\widehat{\sigma}_{\mathrm{I}-}$	197	0.234	0.168	0.028	0.035	0.126	0.256	0.179
$\widehat{\sigma}_{\mathrm{I}\pm}$	201	0.052	0.012	0.005	0.006	0.05	0.063	0.05
$\widehat{\sigma}_{\mathrm{R}+}$	262	0.572	0.295	0.118	0.088	0.296	0.582	0.377
$\widehat{\sigma}_{\mathrm{R}-}$	262	0.013	0.0	0.001	0.007	0.003	0.033	0.013
$\widehat{\sigma}_{\mathrm{R}\pm}$	256	0.292	0.159	0.083	0.04	0.122	0.3	0.2

Table 2: Summary of model results for Child AoA and LMs trained on the Unified dataset. Note: *lexical category* does not contain the category OTHER which includes words that could be assigned to NOUNS, PREDICATES or FUNCTION WORDS.

studied: (i) **log frequency** with respect to each LM's training dataset for LMs and with respect to CHILDES for children, (ii) **number of characters** (iii) **concreteness** judgments, collected from human subjects by Brysbaert et al. (2014), that indicate the extent to which a word is concrete, measured on a scale from 1 (very abstract) to 5 (very concrete), (iv) **mean length of utterances (MLU)** in CHILDES that contain the word for children and with respect to each LM's training dataset for LMs, and (v) **lexical category** NOUN, PREDICATE, FUNCTION WORD, and OTHER, annotated by Frank et al. (2017, 2021).

Do similar factors influence LMs and Children $\boldsymbol{AoA}$s? Regressions for children and for LMs trained on Unified are given in Table 2. For children, the adjusted R^2 with all features reaches 0.417. The strongest single predictors of children's AoA are *concreteness* and *lexical category*. *Log frequency* is a notably weak predictor on its own, though it does add meaningful predictive power when added to a model including all other features. These results largely reproduce those of (Braginsky et al., 2016) and Chang and Bergen (2022), the latter of whom reports an adjusted R^2 of 0.43 for predicting child AoA from all features using a larger vocabulary of 571 words.

When it comes to predicting LMs' AoA, we identify two main patterns: First, the signatures $\widehat{\sigma}_{\mathrm{I}\pm}$ and $\widehat{\sigma}_{-}$ exhibit negligible relationships with any of the predictors. Second, among the other signatures, *log frequency* is consistently the most predictive factor, similar to the findings of Chang and Bergen (2022). Predictors *concreteness* and *lexical category* are the next most predictive factors. The figures in App. G show scatterplots of different AoAs versus each of the predictors. For brevity, we discuss results on the Unified dataset: For all the signatures and datasets, language models (LMs) demonstrate opposite behavior regarding the effects of *log frequency* and *concreteness* on children, with more frequent words having a lower AoA. While children tend to acquire concrete words earlier, language models seem to struggle more with processing concrete words and perform better with abstract ones. Moreover, while children do not display a significant correlation with *number of characters*, most signatures reveal positive correlations with it. The exceptions are $\widehat{\sigma}_{\mathrm{R}-}$ and $\widehat{\sigma}_{\mathrm{I}\pm}$ (BabyLM and CHILDES), which show slightly negative correlations. Lastly, *MLU* exhibits a similar pattern in both children and language models.

Does more developmentally plausible training data result in more human-like learning patterns? From Fig. 2, we can see that the models trained on the BabyLM tend to have the most human-like learning trajectories, according to some signatures (though, as stated above, LM trajectories are far from human-like across the board). This is surprising given that CHILDES, which comes from discourses between caregivers and young children, is the data most closely resembling the input to the young children studied in Wordbank. However, since all three datasets differ greatly in size, we cannot determine whether this result is due to data domain or dataset size. By analyzing Table A8, we find that there is a positive effect of the training set size on the predictability of AoA. We speculate that this may explain why Chang and Bergen (2022), who trained models on much larger datasets than ours, reported higher predictability for model AoA scores (for $\widehat{\sigma}_+$). We also note that in the CHILDES dataset, *log frequency* is not significantly more predictive than other factors, in contrast with other datasets. Overall, the results do not exhibit any notable human-like patterns.

8 Discussion and Conclusion

Our main objective was to explore more fully the space of distributional tests of word learning. We showed that the distributional test adopted by Chang and Bergen (2022) and Portelance et al. (2023) can be viewed as an estimator of a more general distributional signature. This insight also enabled us to define a larger family of signatures that follow a clear typology. However, the question remains, which of these evaluations should be the focus of researchers interested in studying word learning in LMs? One of our main findings in §6.3 is that many of these signatures are complementary. This is true with respect to children's AoAs as well as in comparison to each other. Arguably, considering both positive and negative contexts gives a more complete picture of the LM's distributional knowledge, and comparing the LM's distribution against an LLM allows the signature to better reflect the gradience of the ground truth distribution, which is not observable. Nonetheless, each signature we propose has a clear interpretation and may be useful for specific applications, though usable AoA scores cannot always be extracted.

In §7 we found that we could not predict children's AoAs well from any metric of model AoA. This result might be somewhat surprising in light of Portelance et al.'s (2023) finding that LM surprisal improves predictions of children's AoAs. However, we note that that work uses $\widehat{\sigma}_+$ at the end of training as a predictor AoA, rather than the AoA of the model under that signature. Our results do further corroborate Chang and Bergen's (2022) conclusions on this question, and significantly expand them to a wider variety of signatures. They also add to a growing body of work finding specific differences in the language learning patterns of humans and LMs in other domains (e.g., Evanson et al., 2023; Constantinescu et al., 2024). On the other hand, Zhuang et al. (2024b,a) show that multimodal LMs can show more human-like learning trajectories and also introduce a novel training objective that further improves human-likeness.

Future work should apply our distributional tests to these and other potentially more human-like training procedures. Besides learning in a world grounded in sensory experience, children also learn through interaction both with the physical world and with other agents (Clark, 2018; Nikolaus and Fourtassi, 2023). Moreover, children, unlike LMs, have constraints on production, going through one-word and two-word utterance phases (Bloom, 1970). These factors no doubt influences the kinds of words children use early in development and may account for the precedence of concrete words. There are only a few examples of training regimes for LMs inspired by interaction (Lazaridou et al., 2020; Nikolaus and Fourtassi, 2021b; Ma et al., 2024). Furthermore, the reliance on stochastic gradient descent and cross-entropy loss likely skew learning trajectories in LMs in ways that are not entirely human-like. There are many opportunities for exploring more human-like LM training, and we expect word learning will be an important evaluation of human-likeness as these are explored.

Having better mapped out the space of evaluations for lexical knowledge, our work paves the way for comparing learning trajectories of language models and humans. Our findings provide strong empirical support that there are large differences between how these language learners develop throughout learning and draw attention to the fact that there is significant work to be done to explore pretraining methods and datasets that result in more developmentally plausible language models.

9 Limitations

Our study has several limitations. First, while we are interested in the possibility that LMs can be used as cognitive models and we attempt to use developmentally plausible data, our LMs are not trained in a way that is maximally similar to how humans learn. They lack exposure to speech, grounding, and interaction with other agents, all of which may have a large influence on word learning. Second, while our proposed true and reference signatures are weighted by a distribution $\overrightarrow{p_\kappa}$, we only estimate this distribution using Monte Carlo estimation. Future work should explore whether better estimators, for example based on LLMs, yield qualitatively different results. Third, the specifics of our findings could be sensitive to our training setup. Future work test whether different pretraining pipelines give qualitatively different results. Finally, our study focuses on extracting AoAs from learning trajectories, but AoA is just one statistic that can be extracted from the learning trajectory.

References

Uri Berger, Gabriel Stanovsky, Omri Abend, and Lea Frermann. 2022. A computational acquisition model for multimodal word categorization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3819–3835.

Lois Bloom. 1970. *Language Development: Form and Function in Emerging Grammars*. MIT Press.

Mika Braginsky, Daniel Yurovsky, Virginia A. Marchman, and Michael C. Frank. 2019. Consistency and variability in children's word learning across languages. *Open Mind: Discoveries in Cognitive Science*, 3:52–67.

Mika Braginsky, Daniel Yurovsky, Virginia A. Marchman, and Mike C. Frank. 2016. From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 38:1691–1696.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145, Jeju Island, Korea. Association for Computational Linguistics.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Erin M. Buchanan, K. D. Valentine, and Nicholas P. Maxwell. 2019. English semantic feature production norms: An extended database of 4436 concepts. *Behavior Research Methods*, 51(4):1849–1863.

Tyler A. Chang and Benjamin K. Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2024. Characterizing learning curves during language model pre-training: Learning, forgetting, and stability. *Transactions of the Association for Computational Linguistics*, 12:1346–1362.

Leshem Choshen, Ryan Cotterell, Michael Y. Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. [Call for papers] the 2nd babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *Preprint*, arXiv:2404.06214.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 8281–8297.

Eve V. Clark. 2018. Conversation and Language Acquisition: A Pragmatic Approach. *Language Learning and Development*, 14(3):170–185.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. 2024. Investigating critical period effects in language acquisition through neural language models. *Transactions of the Association for Computational Linguistics*.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

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

Linnea Evanson, Yair Lakretz, and Jean Rémi King. 2023. Language acquisition: do children and language models follow similar learning stages? In *Findings of the Association for Computational Linguistics*, pages 12205–12218.

Larry Fenson, Virginia Marchman, Donna Thal, Phillip Dale, Steven Reznick, and Elizabeth Bates. 2013. *MacArthur-Bates Communicative Development Inventories*. Springer.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing Search in Context: The Concept Revisited. In *Proceedings of the 10th international conference on World Wide Web*, pages 406–414.

J. R. Firth. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*.

Michael C. Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A. Marchman. 2021. *Variability and consistency in early language learning: The Wordbank project.* Variability and consistency in early language learning: The Wordbank project. The MIT Press.

Micheal C. Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A. Marchman. 2017. Wordbank: an open repository for developmental vocabulary data. *Journal of Child Language*, 44(3):677–694.

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1).

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182, Austin, Texas. Association for Computational Linguistics.

Jill Gilkerson, Jeffrey A. Richards, Steven F. Warren, Judith K. Montgomery, Charles R. Greenwood, D. Kimbrough Oller, John H. L. Hansen, and Terrance D. Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, 26(2):248–265.

Judith C. Goodman, Philip S. Dale, and Ping Li. 2008. Does frequency count? parental input and the acquisition of vocabulary. *Journal of Child Language*, 35(3):515–531.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Zellig Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Trevor Hastie and Robert Tibshirani. 1986. Generalized Additive Models. *Statistical Science*, 1(3):297 – 310.

Andrew Heathcote, Scott Brown, and Douglas J. K. Mewhort. 2000. The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, 7(2):185–207.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *UAI*, volume 99, pages 289–296.

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

Marvin Lavechin, Maureen De Seyssel, Lucas Gautheron, Emmanuel Dupoux, and Alejandrina Cristia. 2022. Reverse engineering language acquisition with child-centered long-form recordings. *Annual Review of Linguistics*, 8(1):389–407.

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. 2020. Multi-agent Communication meets Natural Language: Synergies between Functional and Structural Language Learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7663–7674, Online. Association for Computational Linguistics.

Nathaniel Leibowitz, Barak Baum, Giora Enden, and Amir Karniel. 2010. The exponential learning equation as a function of successful trials results in sigmoid performance. *Journal of Mathematical Psychology*, 54(3):338–340.

Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1):1–31.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Tal Linzen. 2019. What can linguistics and deep learning contribute to each other? *Language*, 95(1):e99–e108.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 923–929.

Ziqiao Ma, Zekun Wang, and Joyce Chai. 2024. Babysit a language model from scratch: Interactive language learning by trials and demonstrations. *Preprint*, arXiv:2405.13828.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk: The database, Vol. 2, 3rd ed.* Lawrence Erlbaum Associates Publishers.

Alec Marantz. 2001. Words. *WCCFL XX Handout, USC.*

Rebecca Marvin and Tal Linzen. 2018. Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

AI @ Meta. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Mitja Nikolaus and Abdellah Fourtassi. 2021a. Evaluating the acquisition of semantic knowledge from cross-situational learning in artificial neural networks. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 200–210. Association for Computational Linguistics.

Mitja Nikolaus and Abdellah Fourtassi. 2021b. Modeling the interaction between perception-based and production-based learning in children's early acquisition of semantic knowledge. In *Proceedings of the 25th conference on computational natural language learning*, pages 391–407, Online. Association for Computational Linguistics.

Mitja Nikolaus and Abdellah Fourtassi. 2023. Communicative Feedback in language acquisition. *New Ideas in Psychology*, 68:100985.

Andreas Opedal, Eleanor Chodroff, Ryan Cotterell, and Ethan Wilcox. 2024. On the role of context in reading time prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3042–3058.

OpenAI. 2022. ChatGPT: Optimizing language models for dialogue.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Eva Portelance, Yuguang Duan, Michael C. Frank, and Gary Lupyan. 2023. Predicting age of acquisition for children's Early vocabulary in five languages using language model surprisal. *Cognitive Science*, 47(9).

Eva Portelance, Michael C. Frank, and Dan Jurafsky. 2024. Learning the meanings of function words from grounded language using a visual question answering model. *Cognitive Science*, 48(5).

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.

Enrico Santus, Anna Gladkova, Stefan Evert, and Alessandro Lenci. 2016. The CogALex-V shared task on the corpus-based identification of semantic relations. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 69–79. The COLING 2016 Organizing Committee.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Wai Keen Vong, Wentao Wang, A. Emin Orhan, and Brenden M. Lake. 2024. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511.

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. Publisher: CRC Press.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ludwig Wittgenstein. 1953. *Philosophical investigations*. Wiley-Blackwell.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1112–1125.

Chengxu Zhuang, Evelina Fedorenko, and Jacob Andreas. 2024a. Lexicon-level contrastive visual-grounding improves language modeling. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 231–247.

Chengxu Zhuang, Evelina Fedorenko, and Jacob Andreas. 2024b. Visual grounding helps learn word meanings in low-data regimes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1311–1329.

A Estimation

We now discuss how to estimate the various distributional signatures we introduced in §3.

A.1 An Intrinsic Metric

We develop an intrinsic metric, i.e., a metric that does not relay, in expectation, on the true language model p. Thus, we consider the following information-theoretic quantity that resembles Eq. (6), but where the expectation is taken with respect to the model itself:

$$\sigma_{\mathrm{I}+} \stackrel{\text{def}}{=} -\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{q_\kappa}(\boldsymbol{c} \mid \boldsymbol{w}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c}). \tag{1}$$

In contrast to Chang and Bergen's (2022) distribution signature, Eq. (1) is not grounded in an external language model. Thus, it measures a notion of knowledge internal to the language model itself. We can also, by analogy to Eq. (9), define an intrinsic metric that considers just negative contexts

$$\sigma_{\mathrm{I}-} \stackrel{\text{def}}{=} -\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{q_\kappa}(\boldsymbol{c} \mid \neg\boldsymbol{w}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c}) \tag{2}$$

and one that considers all contexts

$$\sigma_{\mathrm{I}\pm} \stackrel{\text{def}}{=} -\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{q_\kappa}(\boldsymbol{c}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c}). \tag{3}$$

A.2 A Practical Estimator

We now discuss a scheme to estimate Eq. (1). First, we note that, by Bayes' rule, we have

$$\overrightarrow{q_\kappa}(\boldsymbol{c} \mid \boldsymbol{w}) = \frac{\overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})\overrightarrow{q}(\boldsymbol{c})}{\sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})\overrightarrow{q}(\boldsymbol{c})}. \tag{4}$$

Instead, we consider the following approximation. Given a bag of contexts $\mathcal{C} = \wr \boldsymbol{c}^{(m)} \wr_{m=1}^{M}$ that proceed a word $\boldsymbol{w}$, we construct the following empirical approximation

$$\widetilde{q}_\kappa(\boldsymbol{c} \mid \boldsymbol{w}) = \frac{\mathbb{1}\{\boldsymbol{c} \in \mathcal{C}\}\,\overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c})\,\overrightarrow{q_\kappa}(\boldsymbol{c})}{\sum_{m=1}^{M} \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c}^{(m)})\,\overrightarrow{q_\kappa}(\boldsymbol{c}^{(m)})}. \tag{5}$$

Plugging Eq. (5) into Eq. (1), we arrive at

$$\widehat{\sigma}_{\mathrm{I}+} \stackrel{\text{def}}{=} -\sum_{m=1}^{M} \widetilde{q}_\kappa(\boldsymbol{c}^{(m)} \mid \boldsymbol{w}) \log \overrightarrow{q}(\boldsymbol{w} \mid \boldsymbol{c}^{(m)}),$$

In the limiting case, i.e., when $\mathcal{C}$ includes all of Σ^*, we have $\widehat{\sigma}_{\mathrm{I}+} \to \sigma_{\mathrm{I}+}$. Note that Eq. (6) is not a standard Monte Carlo estimator as the contexts $\boldsymbol{c}^{(m)}$ may not have been drawn from $\overrightarrow{q_\kappa}(\cdot \mid \boldsymbol{w})$, but it is still consistent. An analogous estimator can be derived for Eq. (2) and Eq. (3).

B $\boldsymbol{\sigma}_{\mathrm{R}+}, \boldsymbol{\sigma}_{\mathrm{R}-}$ and $\boldsymbol{\sigma}_{\mathrm{R}\pm}$ are distance metrics

The **reference** signatures as introduced in §3 can be easily shown to be distance metrics.
Let $x_c \stackrel{\text{def}}{=} \log \overrightarrow{p_\kappa}(\boldsymbol{c} \mid \boldsymbol{w})$ and $y_c \stackrel{\text{def}}{=} \log \overrightarrow{r}(\boldsymbol{w} \mid \boldsymbol{c})$, we can rewrite the signatures as follows:

$$\sigma_{\mathrm{R}+} = \sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c} \mid \boldsymbol{w})\,|x_c - y_c|$$

$$\sigma_{\mathrm{R}-} = \sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c} \mid \neg\boldsymbol{w})\,|x_c - y_c|$$

$$\sigma_{\mathrm{R}\pm} = \sum_{\boldsymbol{c}\in\Sigma^*} \overrightarrow{p_\kappa}(\boldsymbol{c})\,|x_c - y_c|$$

Since $\overrightarrow{p_\kappa}(\boldsymbol{c} \mid \boldsymbol{w})$, $\overrightarrow{p_\kappa}(\boldsymbol{c} \mid \neg\boldsymbol{w})$ and $\overrightarrow{p_\kappa}(\boldsymbol{c})$ are all greater than zero, the expressions above represent weighted Manhattan distances, dmonstrating that they are indeed distance metrics.

C Training Details

Training was conducted in parallel across *8 GPUs*, with *gradient accumulation steps* set to 16 and a *batch size per device* of 4. As a result, our model was trained with an *effective batch size* of 512.

Hyperparameter	Value
# of heads	12
# of layers	12
learning rate	7e-4
learning rate scheduler	linear
precision	fp16

Table A3: Training Hyperparameters for GPT-2

Dataset	42	123	28053
Childes	2,800	2,800	2,600
BabyLM	4800	7200	6000
Childes	30,000	30,000	30,000

Table A4: Final steps for the model trained with seeds 42, 123 and 28053.

We saved the checkpoints used for our analysis at increasing intervals throughout the training:

- Every 50 steps for steps $\in (0, 1, 000]$
- Every 200 steps for steps $\in (1, 000, 10, 000]$
- Every 500 steps for steps $\in (10, 000, 30, 000]$

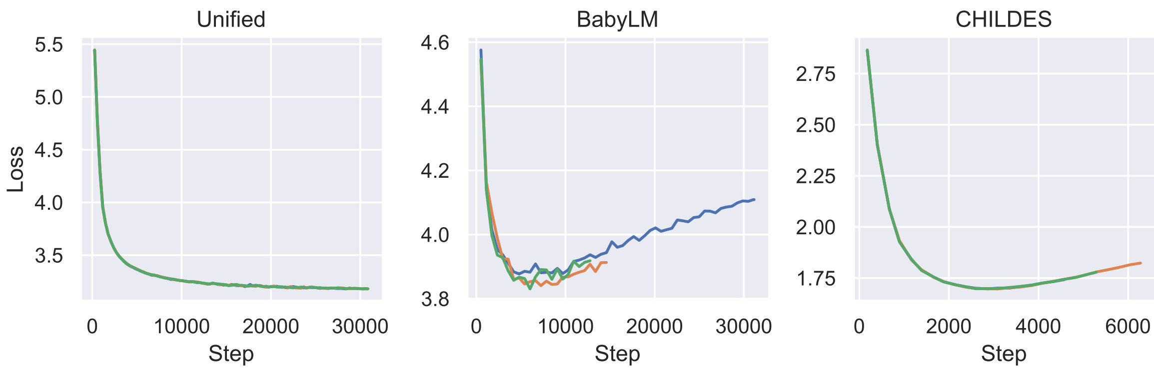


Figure A3: Validation losses for models trained on Unified, BabyLM, and CHILDES. The curves show the necessity for an earlier stopping step for seed 42 (blue), 123 (orange), and 28053 (green).

D Extract through convergence

When it comes to complex signatures that are nonmonotonic or do not have consistently shaped trajectories, one way to extract AoA is to find the convergence point. Let $\sigma(\boldsymbol{w}, t)$ the value that the signature of the word $\boldsymbol{w}$ assumes at time-step t.

Fix a tolerance parameter $\epsilon > 0$. Then, the age of acquisition AoA is

$$AoA = \alpha\left(\sigma, \boldsymbol{w}\right) = \operatorname*{argmin}_{t \in \{1,\ldots,T\}} \left(\max_{s,s' \in \{t,\ldots,T\}} \left|\sigma(\boldsymbol{w}, s) - \sigma(\boldsymbol{w}, s')\right| < \epsilon \right) \tag{6}$$

This definition mirrors the definition of the convergence of a Cauchy sequence. However, because T is finite, for small enough ϵ, we do not, in general, observe true convergence in the analytic sense. Thus, the tolerance parameter ϵ is best viewed as a hyperparamter, and our findings are dependent on the choice of ϵ. However, given that nearly all learning algorithms are analyzed by letting $T \to \infty$, there is a sense in which our definition of AoA is well-founded. Specifically, if we assume the convergence of the learning algorithm as $T \to \infty$ implies the convergence of σ, then, for every $\epsilon > 0$, there exists a number of epochs such that we will achieve $\alpha\left(\sigma, \boldsymbol{w}\right)$.

To address the shape instability, we decided to smooth the curves with a moving average. Once the denoised signal is obtained, we then apply the previously mentioned convergence method to extract AoA from our signatures.

Fig. A4 shows the percentage of words that did not reach convergence for various ϵ values. A word is marked as non-converged if σ failed to converge for even a single seed.

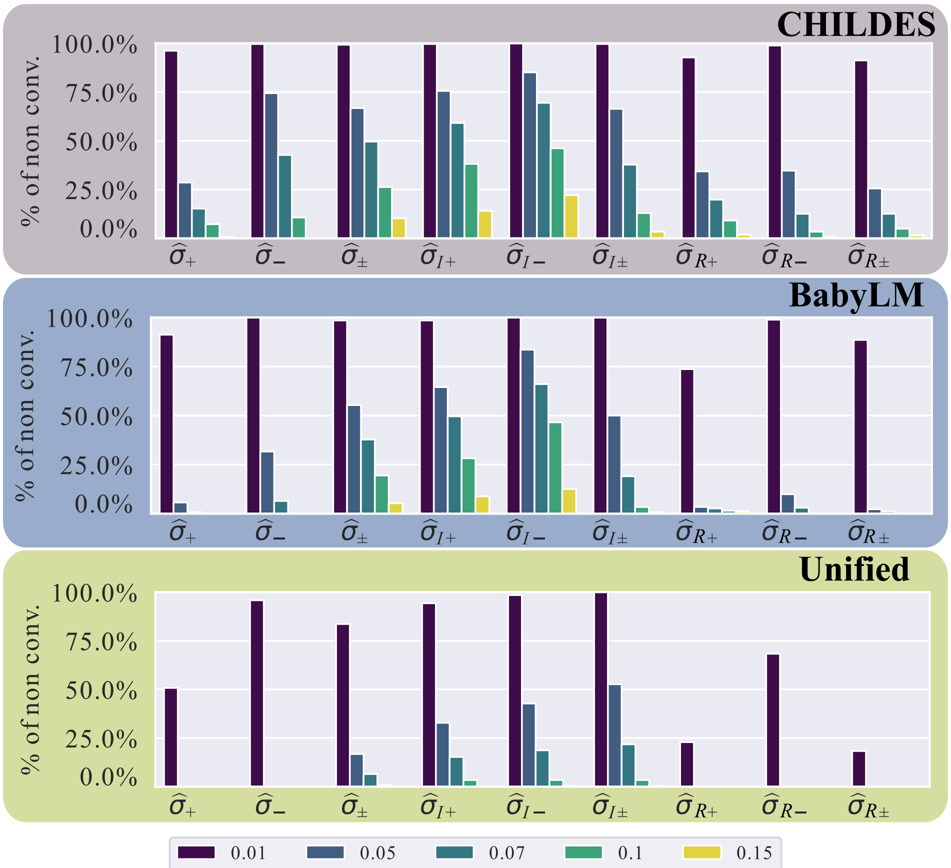


Figure A4: Percentage of non-converged words across various ϵ.

E Case studies' $AoAs$

Table A5 presents the AoA for each word illustrated in Fig. 1. The AoA values for the signatures were obtained from the model trained on the Unified dataset, using seed 42 and extracted using $\epsilon = 0.07$.

Word	Children	$\widehat{\sigma}_+$	$\widehat{\sigma}_-$	$\widehat{\sigma}_\pm$	$\widehat{\sigma}_{\mathrm{I}+}$	$\widehat{\sigma}_{\mathrm{I}-}$	$\widehat{\sigma}_{\mathrm{I}\pm}$	$\widehat{\sigma}_{\mathrm{R}+}$	$\widehat{\sigma}_{\mathrm{R}-}$	$\widehat{\sigma}_{\mathrm{R}\pm}$
the	27.79	0.45	0.94	0.88	0.49	0.89	0.89	0.47	0.80	0.57
off	22.77	0.68	0.91	0.82	0.93	0.95	0.96	0.67	0.89	0.67
water	20.00	0.69	0.91	0.89	not conv.	not conv.	0.92	0.68	0.66	0.67
puzzle	24.79	0.87	0.91	0.92	0.95	0.91	0.95	0.76	0.82	0.64
good	24.54	0.57	0.88	0.45	0.43	0.67	0.94	0.57	0.83	0.59
orange	23.26	0.84	0.92	0.96	not conv.	not conv.	0.92	0.80	0.85	0.64
go	23.33	0.47	0.89	0.59	0.42	0.77	0.92	0.52	0.79	0.54
climb	26.04	0.73	0.92	0.93	0.96	not conv.	0.93	0.75	0.59	0.65

Table A5: AoA for words in Fig. 1.

F First and Last Acquired Words

Table A6 reports the first 10 and the last 10 words that were acquired according to each signature. The words refers to the model trained on the Unified dataset, using seed 42 and extracted using $\epsilon = 0.07$.

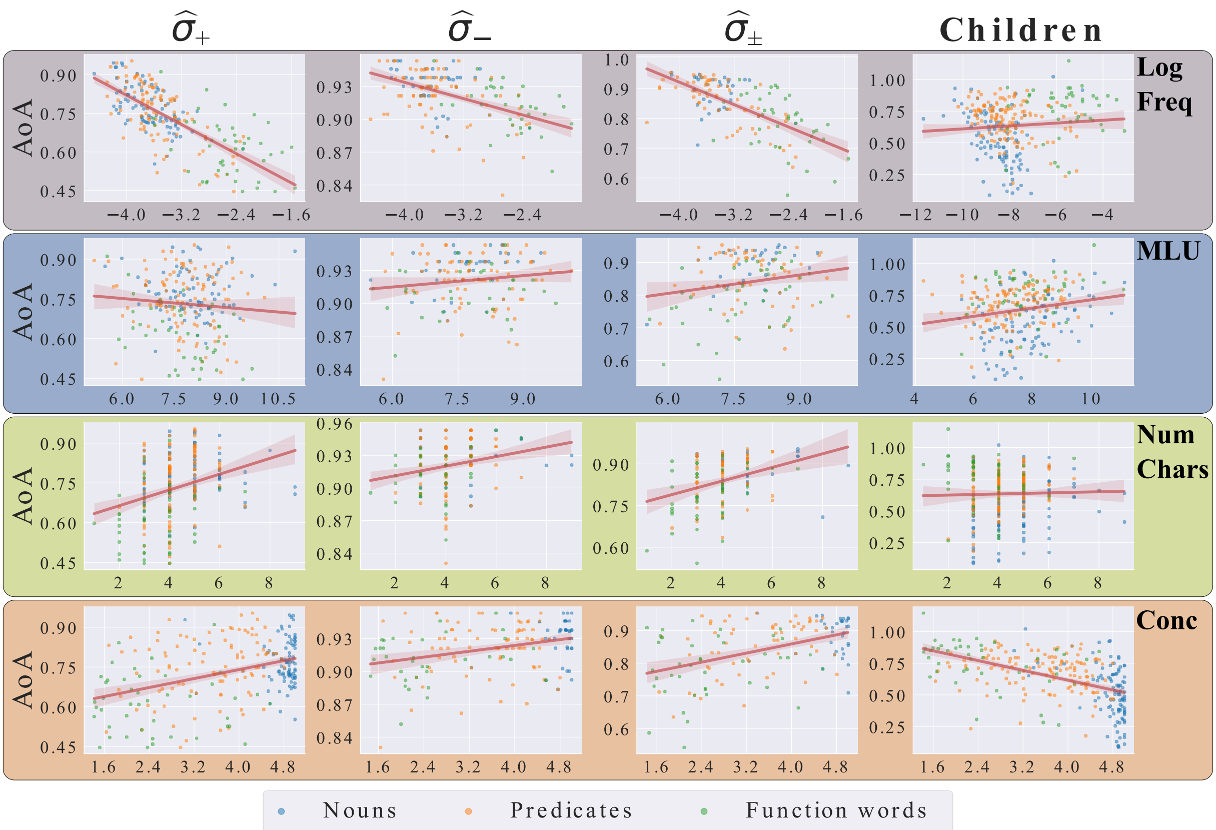
σ	First acquired words	Last acquired words
$\widehat{\sigma}_{+}$	not, do, you, have, there, can, this, to, that, am	yes, washing, brush, toy, cow, clock, wash, puzzle, flower, egg
$\widehat{\sigma}_{-}$	all, so, a, he, can, this, there, on, out, for	red, paint, the, dinner, dry, milk, pretty, feed, cup, blue
$\widehat{\sigma}_{\pm}$	under, so, like, for, all, a, at, on, out, here	chocolate, elephant, doll, teacher, truck, gas, washing, kitchen, lips, basket
$\widehat{\sigma}_{\mathrm{I}+}$	need, can, have, gonna, this, is, you, what, not, wanna	green, swim, touch, sleep, broken, dirty, present, park, ear, frog
$\widehat{\sigma}_{\mathrm{I}-}$	have, a, take, that, so, can, here, all, need, now	rock, cake, money, bread, wash, cup, knife, build, teacher, sky
$\widehat{\sigma}_{\mathrm{I}\pm}$	can, am, need, this, a, that, there, stairs, look, first	clock, bedroom, coat, park, your, thank, sheep, away, walk, rain
$\widehat{\sigma}_{\mathrm{R}+}$	make, man, not, have, this, here, little, do, to, put	toy, yes, brush, flower, egg, plate, camera, star, block, washing
$\widehat{\sigma}_{\mathrm{R}-}$	truck, write, say, block, rock, hard, big, a, friend, knife	arm, apple, read, present, star, buy, snow, gas, brush, slow
$\widehat{\sigma}_{\mathrm{R}\pm}$	a, not, am, man, have, on, can, do, make, so	yes, apple, train, empty, frog, basket, toy, brush, draw, gonna

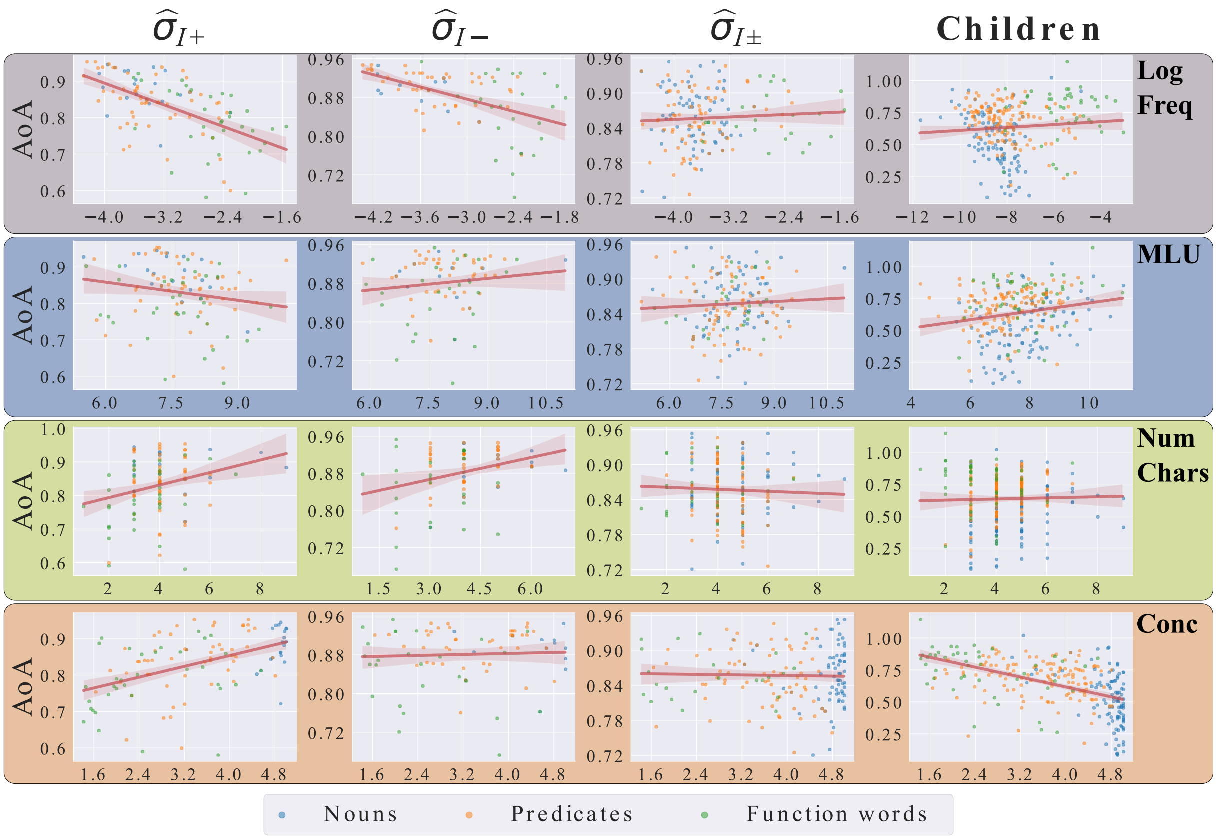
Table A6: First and last acquired words for each signature.

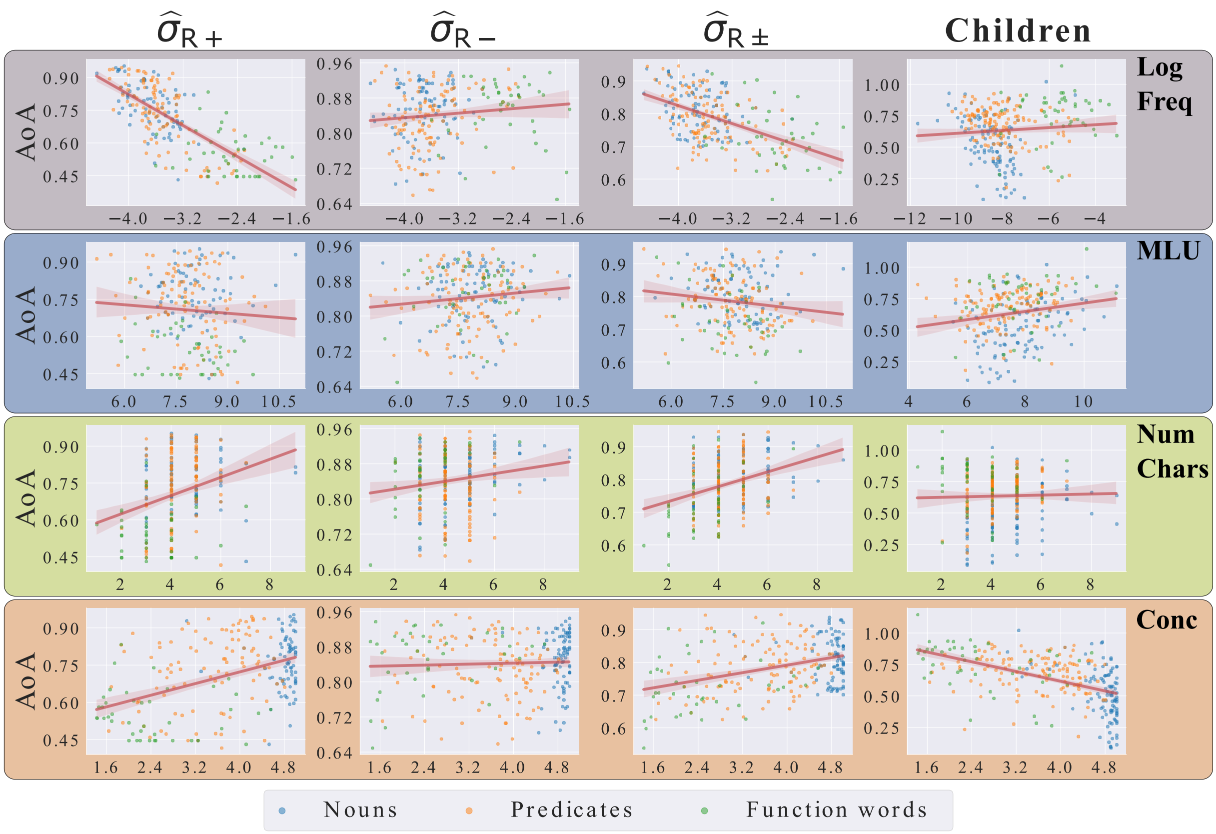
G AoA vs Predictors

In the following subsections, we show how Log Frequency, MLU, Number of Characters, and Concreteness each influence σ's AoA across different datasets. Each AoA value presented in the plots represents the average AoA across multiple seeds, extracted with $\epsilon = 0.07$. Only words that achieved convergence across all seeds were included in the analysis.

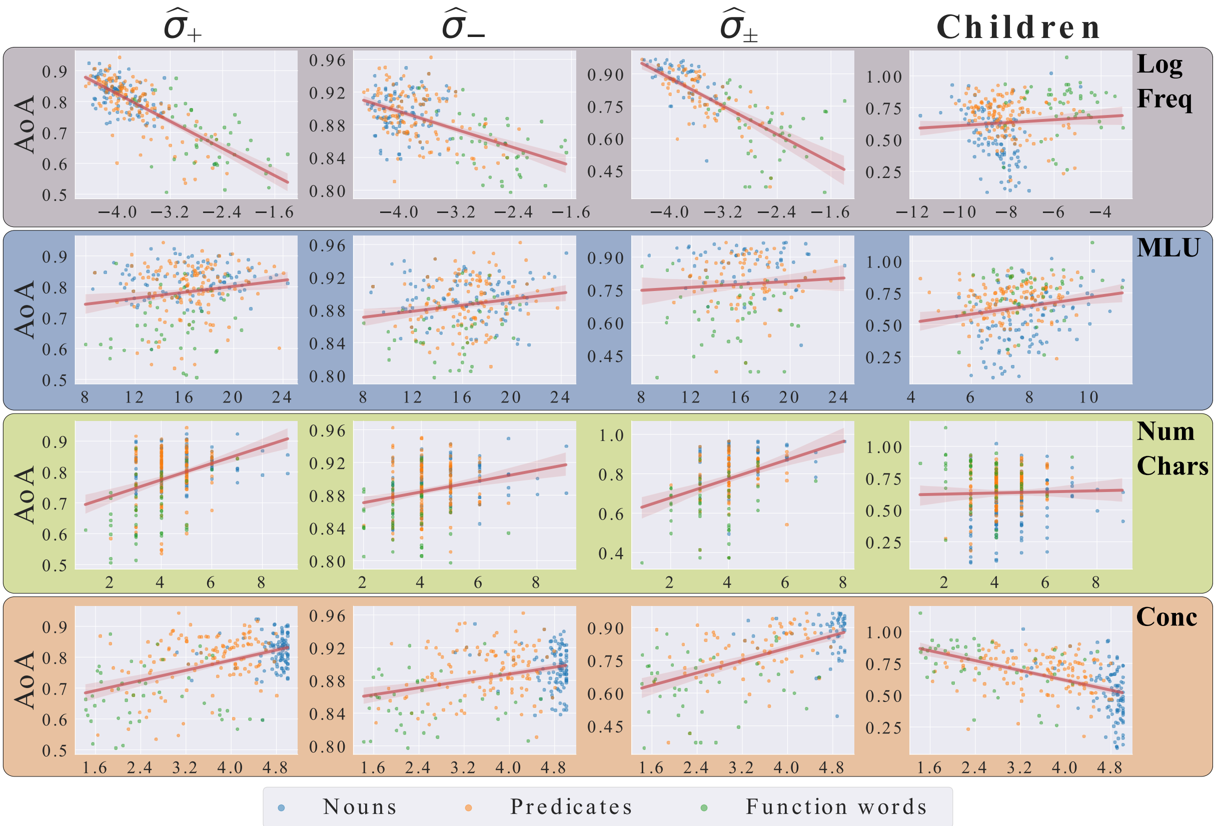
G.1 CHILDES



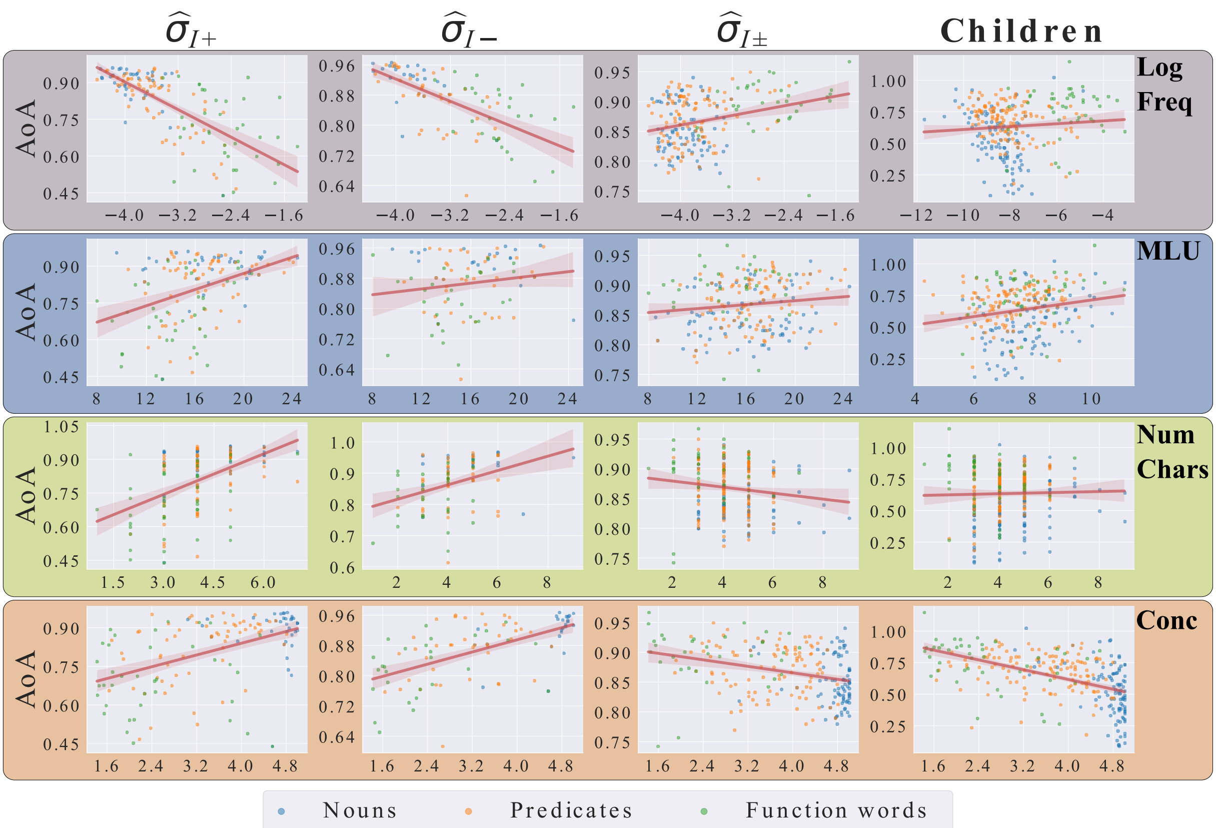

$\widehat{\sigma}_{I+}$
$\widehat{\sigma}_{I-}$
$\widehat{\sigma}_{I\pm}$
Children
Log Freq
MLU
Num Chars
Conc
AoA
Nouns
Predicates
Function words


$\widehat{\sigma}_{\mathrm{R}+}$
$\widehat{\sigma}_{\mathrm{R}-}$
$\widehat{\sigma}_{\mathrm{R}\pm}$
Children
Log Freq
MLU
Num Chars
Conc
AoA
Nouns
Predicates
Function words

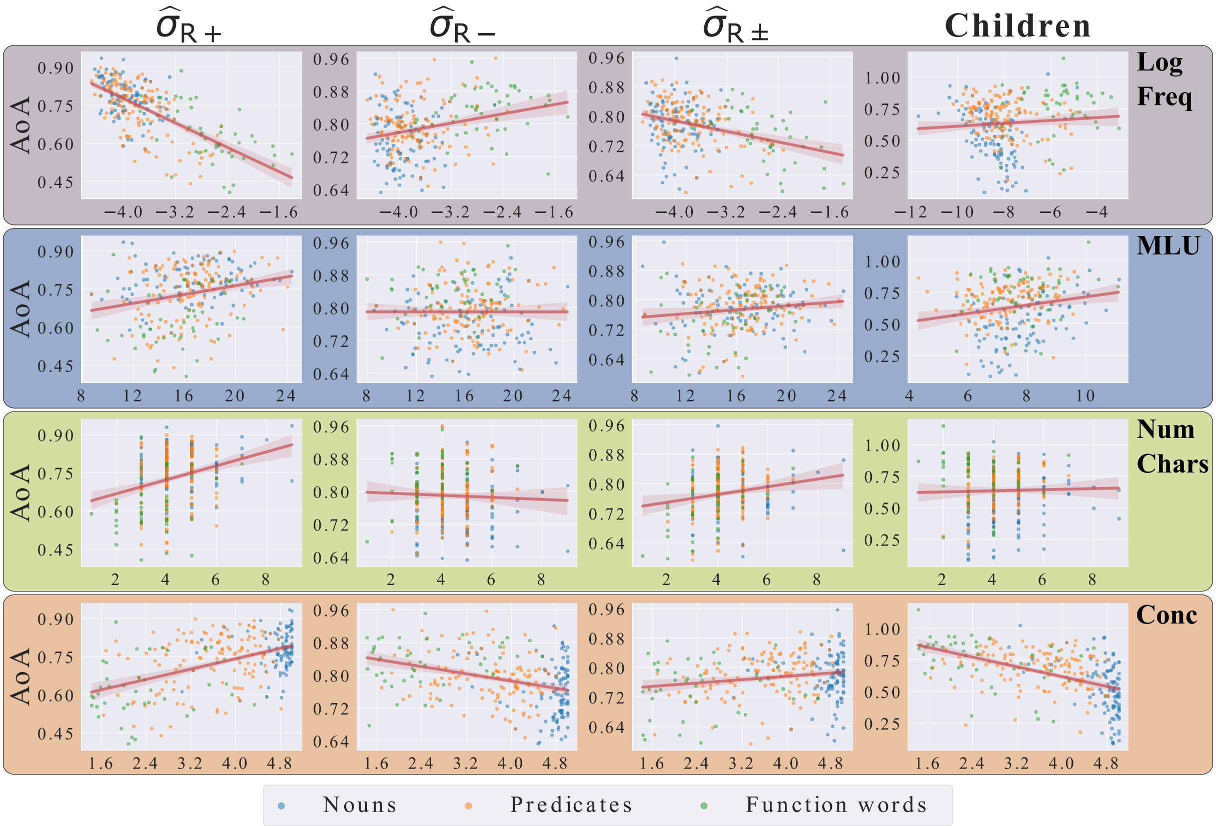
G.2 BabyLM



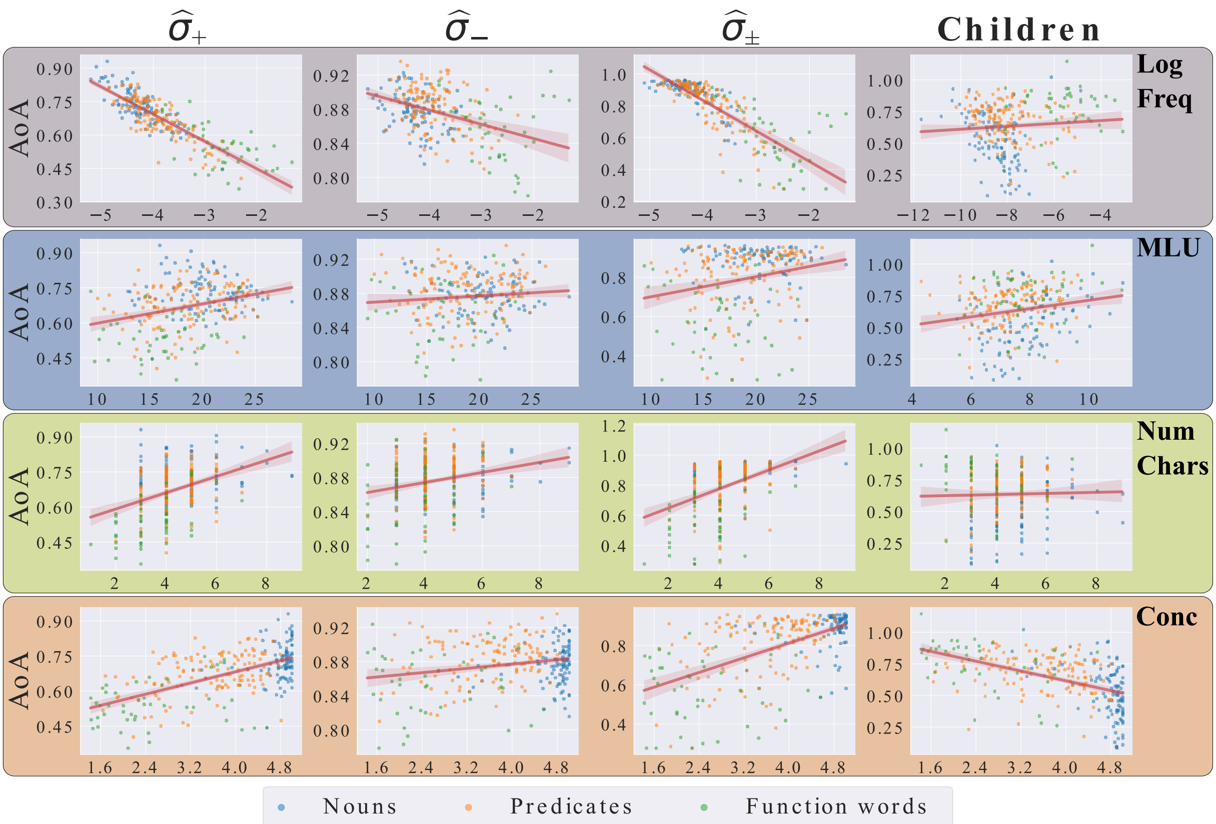
$\widehat{\sigma}_{+}$
$\widehat{\sigma}_{-}$
$\widehat{\sigma}_{\pm}$
Children
Log Freq
MLU
Num Chars
Conc
AoA
Nouns
Predicates
Function words

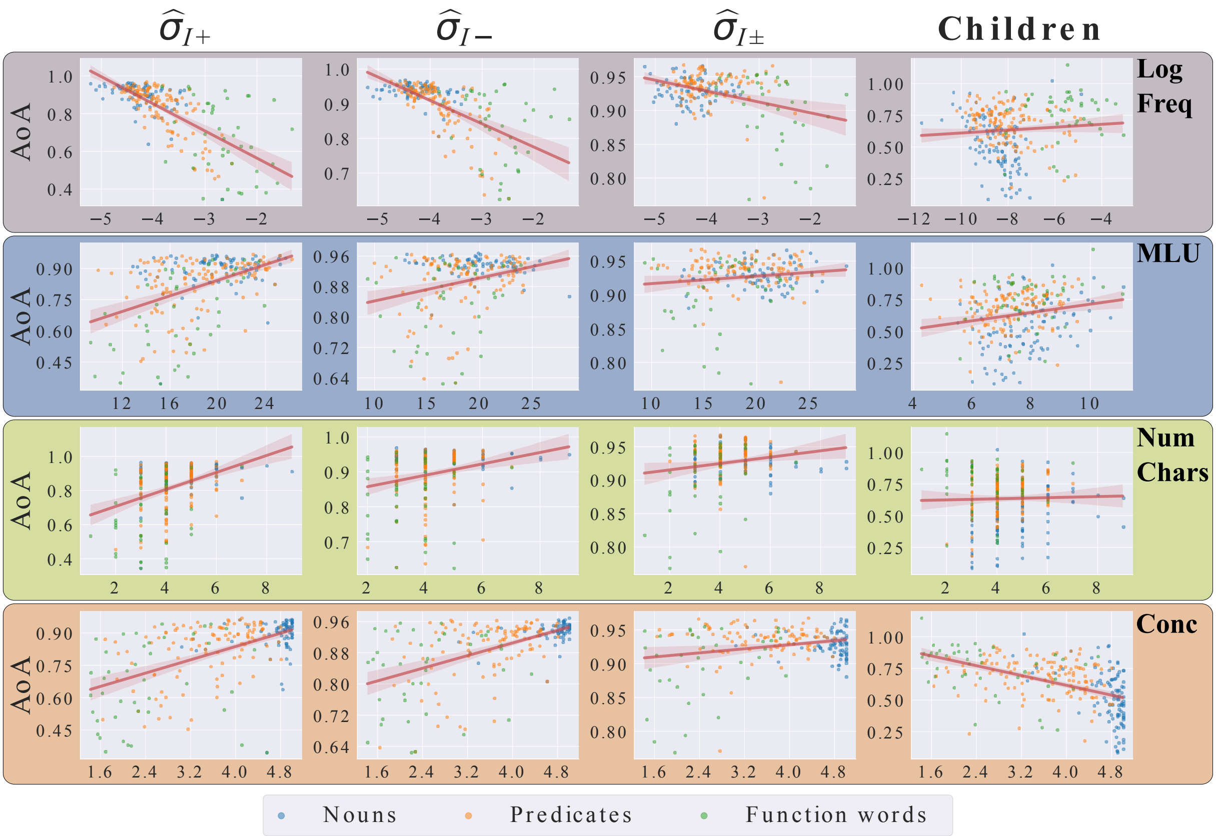


$\widehat{\sigma}_{I+}$
$\widehat{\sigma}_{I-}$
$\widehat{\sigma}_{I\pm}$
Children
Log Freq
MLU
Num Chars
Conc
AoA
Nouns
Predicates
Function words

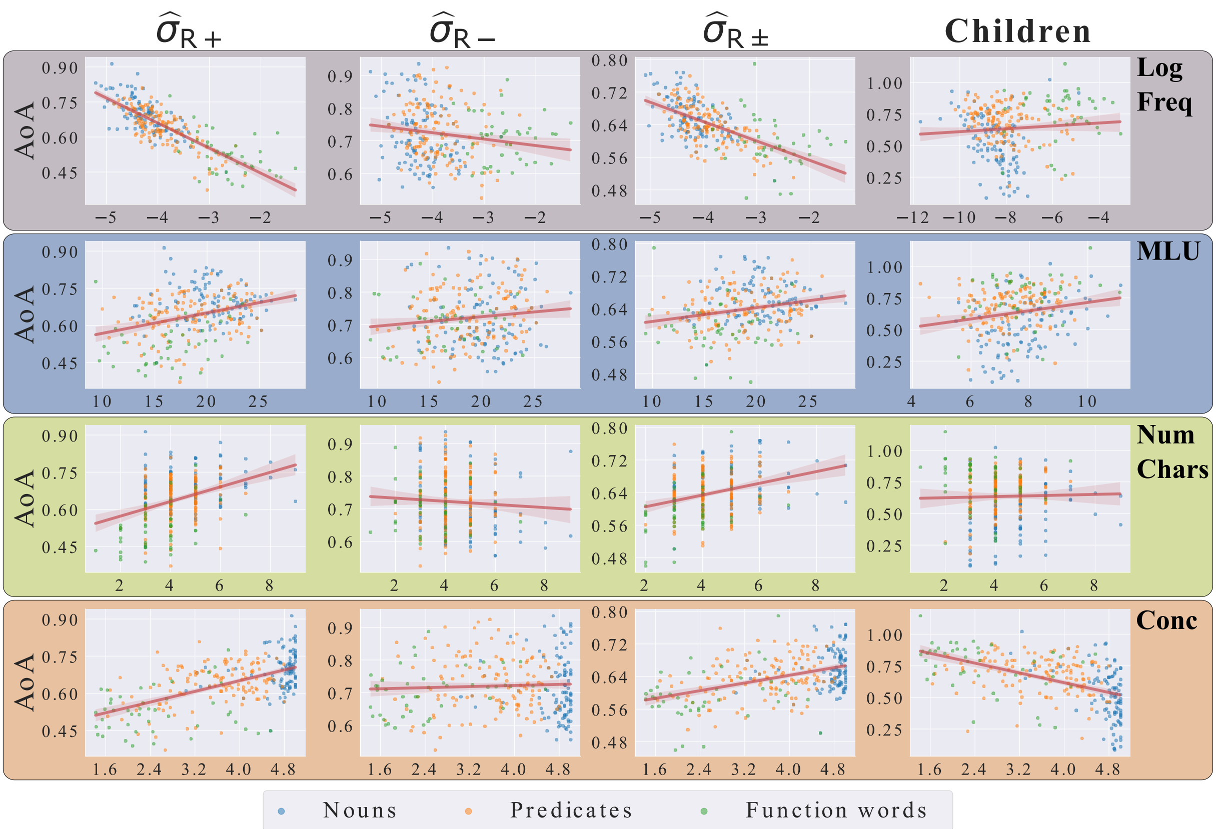

$\widehat{\sigma}_{R+}$
$\widehat{\sigma}_{R-}$
$\widehat{\sigma}_{R\pm}$
Children
Log Freq
MLU
Num Chars
Conc
AoA
Nouns
Predicates
Function words

G.3 Unified


$\widehat{\sigma}_{+}$
$\widehat{\sigma}_{-}$
$\widehat{\sigma}_{\pm}$
Children
Log Freq
MLU
Num Chars
Conc
AoA
Nouns
Predicates
Function words



$\widehat{\sigma}_{I+}$
$\widehat{\sigma}_{I-}$
$\widehat{\sigma}_{I\pm}$
Children
Log Freq
MLU
Num Chars
Conc
AoA
Nouns
Predicates
Function words



$\widehat{\sigma}_{R+}$
$\widehat{\sigma}_{R-}$
$\widehat{\sigma}_{R\pm}$
Children
Log Freq
MLU
Num Chars
Conc
AoA
Nouns
Predicates
Function words

H Regression analysis

```r
full <- paste(predictors, collapse = "+")
reduced <- paste(original_predictors[-1], collapse = "+")
vif_values <- vif(lm(AoA ~ full, data = data))
if (max(vif_values) > 5){
    print("Multicollinearity detected\n")
}
predictors <- c("log_frequency", "concreteness", "n_chars", "mlu", "lexical_class")
for (i in 1:length(predictors)) {
    formula <- paste("AoA ~", predictors[[i]])
    model <- lm(formula, data = data)
    cat(paste(predictors[[i]], "Adjust Rsquared:", summary(model)$adj.r.squared))
}
m_full <- lm(AoA ~ full, data = data)
m_reduced <- lm(AoA ~ reduced, data = data)
cat("Full model Adjust Rsquared:", summary(m_full)$adj.r.squared
cat("Reduced model Adjust Rsquared:", summary(m_reduced)$adj.r.squared
```

Listing 1: simplified R code for the regression analysis.

In the following regressions we denote Log Frequency as **LF**, Concreteness as **Co**, Number of Characters as **NC**, Mean Length of Utterances as **MLU**, and Lexical Category as **LC**. No VIF value has exceeded 5, indicating no multicollinearity among the predictors.

H.1 Children regressions

Model	Predictor	Estimate	p-value	Adj. R^2
LF	Intercept	0.723676	< 2e-16 ***	0.004
	Log Frequency	0.011405	0.162	
C	Intercept	1.003250	< 2e-16 ***	0.2575
	Concreteness	-0.096730	< 2e-16 ***	
NC	Intercept	0.615790	< 2e-16 ***	-0.003
	Number of Characters	0.004372	0.661	
MLU	Intercept	0.38715	1.76e-6 ***	0.03233
	MLU	0.03268	1.7e-3 **	
LC	Function Words	0.75029	< 2e-16 ***	0.2012
	Nouns	-0.2284	6.67e-12 ***	
	Predicates	-0.06586	0.0337 *	
Full	Log Frequency	-0.047086	6.37e-07 ***	**0.4177**
	Number of Characters	0.025407	0.00342 **	
	Concreteness	-0.091746	3.51e-10 ***	
	MLU	0.046211	6.70e-08 ***	
	Function Words	0.26867	0.00454 **	
	Nouns	-0.193857	1.72e-05 ***	
	Predicates	-0.084683	0.01456 *	
Full \ LF	Number of Characters	0.038773	5.50e-07 ***	0.3625
	Concreteness	-0.073379	4.91e-07 ***	
	MLU	0.043550	9.72e-07 ***	
	Function Words	-0.042864	5.50e-07 ***	
	Nouns	-0.115122	0.0085 **	
	Predicates	-0.001887	0.9527	

Table A7: Children regressions. For each model the Adjusted R^2, the estimate for the predictors and the p-value for predictor significance.

H.2 Signatures regressions

The table below reports the Adjusted R^2 values for each predictor across all datasets and signatures, as introduced in App. G. Consistent with previous analyses, the AoA values were computed using a ϵ = 0.07. The table shows the number of words included in the regression analysis, counting only those that remained after outlier removal and successfully achieved convergence across all three seeds. Among all the signatures, $\widehat{\sigma}_+$ is the one that demonstrates the strongest predictive power across nearly all predictors.

Dataset	σ	#words	LF	Co	NC	MLU	LC	Full	Full \ LF
CHILDES	$\widehat{\sigma}_+$	222	0.365	0.106	0.065	0.004	0.162	0.376	0.192
	$\widehat{\sigma}_-$	141	0.108	0.032	0.025	-0.001	0.037	0.105	0.046
	$\widehat{\sigma}_\pm$	124	0.236	0.074	0.06	0.012	0.081	0.252	0.129
	$\widehat{\sigma}_{\mathrm{I}+}$	103	0.173	0.110	0.035	0.014	0.068	0.193	0.125
	$\widehat{\sigma}_{\mathrm{I}-}$	77	0.115	-0.003	0.039	0.003	0.073	0.132	0.088
	$\widehat{\sigma}_{\mathrm{I}\pm}$	163	-0.001	-0.002	-0.001	-0.0	-0.002	-0.006	-0.005
	$\widehat{\sigma}_{\mathrm{R}+}$	210	**0.412**	**0.146**	0.069	0.002	**0.199**	**0.422**	**0.232**
	$\widehat{\sigma}_{\mathrm{R}-}$	229	0.006	-0.0	0.013	0.006	0.011	0.048	0.028
	$\widehat{\sigma}_{\mathrm{R}\pm}$	228	0.182	0.103	**0.078**	**0.016**	0.093	0.216	0.162
BabyLM	$\widehat{\sigma}_+$	257	**0.48**	0.168	0.092	0.026	0.217	**0.483**	0.278
	$\widehat{\sigma}_-$	243	0.124	0.052	0.029	0.014	0.083	0.125	0.1
	$\widehat{\sigma}_\pm$	162	0.413	0.222	0.091	-0.001	0.29	0.433	0.323
	$\widehat{\sigma}_{\mathrm{I}+}$	130	0.383	**0.224**	**0.179**	**0.107**	**0.221**	0.421	**0.363**
	$\widehat{\sigma}_{\mathrm{I}-}$	80	0.288	0.2	0.103	0.027	0.135	0.297	0.229
	$\widehat{\sigma}_{\mathrm{I}\pm}$	209	0.056	0.066	0.016	0.004	0.063	0.085	0.08
	$\widehat{\sigma}_{\mathrm{R}+}$	253	0.407	0.204	0.082	0.057	0.188	0.425	0.291
	$\widehat{\sigma}_{\mathrm{R}-}$	254	0.036	0.061	-0.0	-0.001	0.072	0.079	0.079
	$\widehat{\sigma}_{\mathrm{R}\pm}$	257	0.087	0.023	0.025	0.011	0.023	0.097	0.06
Unified	$\widehat{\sigma}_+$	262	**0.614**	**0.298**	0.135	0.072	**0.304**	**0.616**	**0.392**
	$\widehat{\sigma}_-$	251	0.063	0.012	0.023	0.0	0.047	0.083	0.065
	$\widehat{\sigma}_\pm$	245	0.542	0.265	**0.142**	0.04	0.294	0.546	0.379
	$\widehat{\sigma}_{\mathrm{I}+}$	215	0.4	0.274	0.107	**0.162**	0.201	0.463	0.382
	$\widehat{\sigma}_{\mathrm{I}-}$	197	0.234	0.168	0.028	0.035	0.126	0.256	0.179
	$\widehat{\sigma}_{\mathrm{I}\pm}$	201	0.052	0.012	0.005	0.006	0.05	0.063	0.05
	$\widehat{\sigma}_{\mathrm{R}+}$	262	0.572	0.295	0.118	0.088	0.296	0.582	0.377
	$\widehat{\sigma}_{\mathrm{R}-}$	262	0.013	0.0	0.001	0.007	0.003	0.033	0.013
	$\widehat{\sigma}_{\mathrm{R}\pm}$	256	0.292	0.159	0.083	0.04	0.122	0.3	0.2

Table A8: Table reporting the Adj. R^2 for the linear models predicting LM's AoA.

I Thresholds' Correlation

The Pearson correlations reported in Fig. A5 illustrates how the AoA values extracted using varying ϵ correlate. This analysis aims to determine whether the choice of ϵ significantly impacts the results. As discussed in §6.2, with few exceptions, the results across different ϵ values show high correlation. Therefore, our analysis will remain consistent regardless of the choice of ϵ.

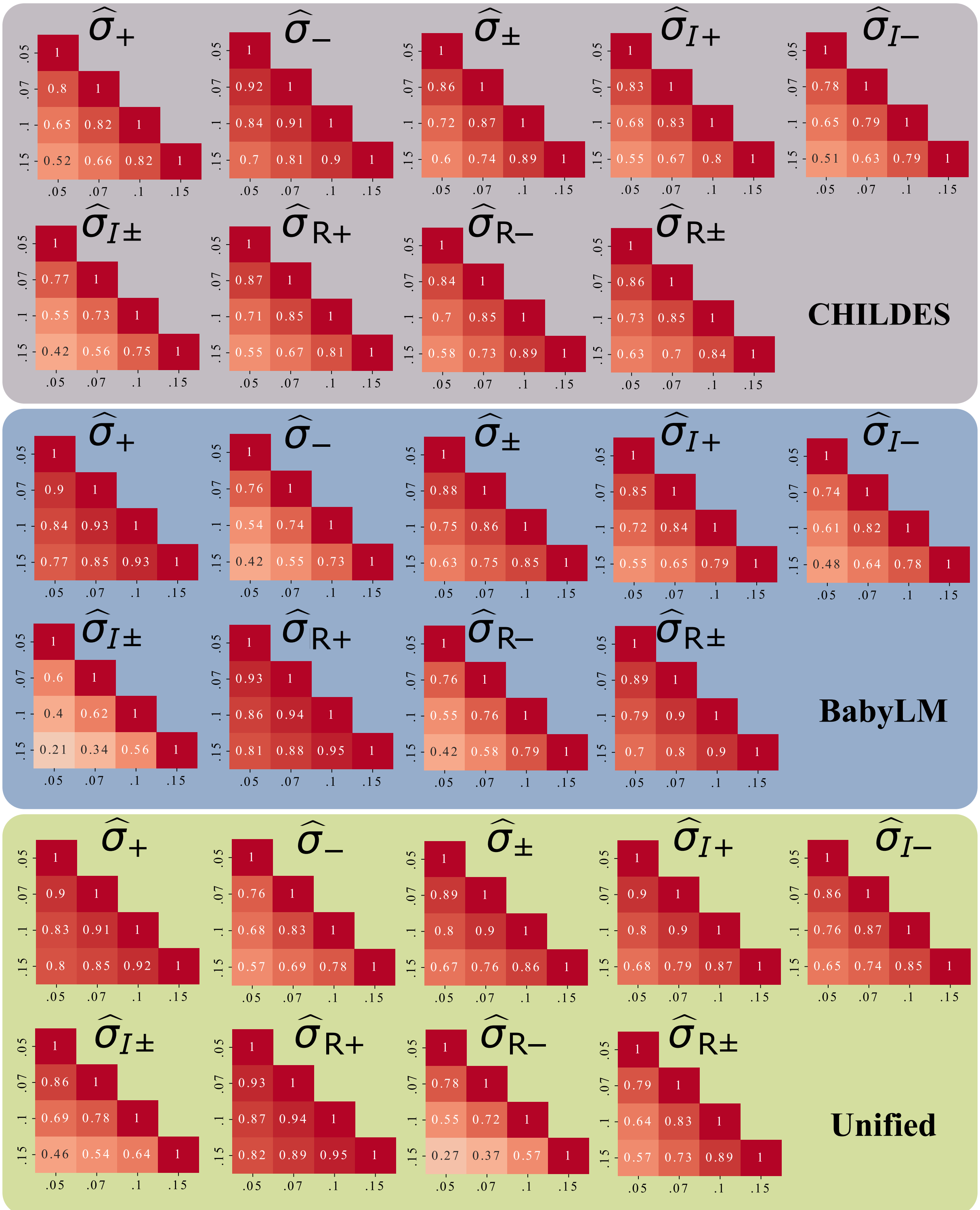


Figure A5: For each σ, we present a Pearson correlation coefficient matrix comparing different ϵ. Warmer colors indicate stronger positive correlations, cooler color indicate stronger negative correlations.

J Signatures' AoA Correlation

The figures in this section show how each signature's AoA values correlate with one another across each model trained with different datasets. As a result, the dataset itself influences the correlation patterns among the different signatures. For example, the Unified dataset displays only positive correlations, whereas Childes and BabyLM datasets exhibit negative correlations for $\widehat{\sigma}_{\mathrm{I}\pm}$ and $\widehat{\sigma}_{\mathrm{R}-}$. The AoA values were extracted using $\epsilon = 0.07$.