Revisiting Age of Acquisition in Curriculum Learning: Disentangling Lexical Features and Semantic Structure

Ian Gifford[*1,3], **Aaron Shah**[*1], **Catherine Chen**[*1,3], **Taimaa Kassab Bachi**[*1], **Eva Portelance**[2,3]
[1]McGill University, Montreal, Canada
[2]HEC Montréal, Montreal, Canada
[3]Mila - Quebec AI Institute, Montreal, Canada
{ian.gifford, aaron.shah, catherine.chen4, taimaa.bachi}@mail.mcgill.ca,
eva.portelance@hec.ca

Abstract

Previous work has found that ordering training data by children's Age of Acquisition (AoA) for words increases the stability of distributional word embeddings, suggesting that early-learned words play a privileged role in shaping semantic structure. In this study, we determine whether AoA itself drives these effects, or whether they emerge from correlated lexical factors such as frequency, concreteness, and phonological complexity. Using incremental Word2Vec training, we construct curricula ordered by AoA and by individual lexical features, while systematically controlling for vocabulary growth and deterministic ordering effects. We show that AoA-ordered curricula produce greater early-phase stability than shuffled baselines, even under controlled exposure conditions. We find that the advantage observed with AoA can be largely explained by correlated factors like overall word frequency. Despite limited gains on general similarity benchmarks, AoA-ordered embeddings outperform shuffled embeddings on a proxy domain-specific task: predicting human AoA norms. This advantage persists after debiasing timestamp effects, implying that AoA curricula induce developmentally meaningful semantic structure.

1 Introduction

Children's language acquisition progresses through systematic vocabulary growth that follows a structured trajectory. Large-scale cross-linguistic analyses of over 30,000 children show that the ordering of early vocabulary is consistent, and that properties of the linguistic environment systematically predict acquisition timing (Braginsky et al., 2019). In other words, lexical exposure is structured, cumulative, and developmentally ordered.

Age of Acquisition (AoA) is defined as the earliest age at which a word is understood (Kuperman et al., 2012). Numerous studies report that early-learned words are recognized, named, or read aloud faster and more accurately than later-learned words (Juhasz, 2005; Brysbaert et al., 2000).

Connectionist accounts of language development propose that early-learned words disproportionately shape the structure of semantic space because they are learned during periods of increased network plasticity (Ellis and Ralph, 2000; Steyvers and Tenenbaum, 2005). Consistent with these theories, DeChant and Bauer (2021) demonstrated that ordering training data by AoA yields greater representational stability in incremental word embeddings. That is, words maintained more consistent semantic neighborhoods across independent training runs. The justification of AoA effects, however, remains contested. The cumulative frequency hypothesis proposes that words learned early are processed more quickly simply because they have been encountered more often over a lifetime, meaning that apparent AoA effects may reflect differences in total exposure rather than learning order itself (Ghyselinck et al., 2004). AoA is also highly entangled with other lexical features. Early words tend to be more frequent, more concrete (Brysbaert et al., 2014), and phonologically simpler. Disentangling AoA from frequency, concreteness, and phonological complexity is therefore essential.

Thus, in this paper, we address the following two research questions: (1) Do AoA-based learning curricula for word embeddings still present an advantage over random baselines when we control for frequency of exposure? and (2) If AoA curricula truly represent an advantage for semantic space acquisition, what external factors may explain these gains?

We show that AoA ordering does in fact yield early stability advantages during training. These effects persist after controlling for curriculum artifacts such as frequency of exposure and sentence aggregation strategies. Although much of

[*]Equal contribution.

the AoA effect is explained by correlated lexical features, a residual AoA-specific signal remains. Finally, AoA-based curricula improve performance on child-language-related evaluation tasks, suggesting that developmentally structured input shapes semantic representation in meaningful ways.

2 Related work

Recent works examined whether neural language models exhibit developmental patterns observed in human language acquisition. Hosseini et al. (2024) showed that transformer language models trained on developmentally realistic amounts of input can predict human neural responses to language. Similarly, Evanson et al. (2023) investigated whether children and neural language models follow comparable stages of language acquisition, finding that some learning stages are shared. Chang and Bergen (2022) analyzed when words became learnable in neural language models and how timing differs in humans.

A related line of works links lexical acquisition to long-term semantic stability. Cassani et al. (2021) reported that words learned earlier tend to exhibit less semantic change historically. Baumann and Hartmann (2026) built on these results by decomposing the effect into specific aspects of semantic change.

Embedding stability has previously been studied as a property of distributional semantic models. Wendlandt et al. (2018) define stability in terms of nearest-neighbor consistency across training runs. Antoniak and Mimno (2018) and Hellrich and Hahn (2016) demonstrate that embedding neighborhoods can vary substantially under small corpus perturbations and retraining, emphasizing that semantic conclusions drawn from unstable embedding neighborhoods may lack reproducibility. Further work has shown that embedding instability can propagate into downstream task performance (Leszczynski et al., 2020). In our developmental setting, however, stability also takes on a theoretical interpretation: early stability may reflect the rapid formation of a reliable semantic structure in which early-acquired words function as semantic anchors for later vocabulary growth (Steyvers and Tenenbaum, 2005).

3 Lexical features

In DeChant and Bauer (2021)'s original approach, training curricula were constructed using AoA norms as a cognitively grounded proxy for the structure of early vocabulary growth, allowing them to test whether words learned earlier exhibit more stable representations.

Here, we extend this approach by incorporating additional lexical features that are linked to AoA. We drew on findings from Braginsky et al. (2016, 2019), who showed that a range of lexical predictors, most notably input frequency, but also semantic factors such as concreteness, can be used to predict when words are typically learned across languages.

Motivated by this work, we constructed lookup tables for four classes of lexical features: age of acquisition, concreteness, frequency, and phonological complexity.

3.1 Age of Acquisition

AoA values were obtained primarily from the AoA norms of Kuperman et al. (2012). To increase coverage, we supplemented this dataset with the AI-generated AoA estimates of Green et al. (2025), which were fine-tuned to align with Kuperman et al. estimates. When a word appeared in both sources, the human-rated Kuperman et al. value was retained and the AI-generated estimate was discarded.

In addition, we generated morphologically inflected variants using the `lemminflect` library when such forms were not already present in the dataset. These new entries were assigned AoA values slightly higher than those of their base forms by adding a small constant (1×10^{-5}) to the base form's AoA value. This ensured that each added inflected form was ranked directly after its base form (e.g., *children* immediately follows *child*) while preserving the overall rank ordering. This procedure is consistent with DeChant and Bauer (2021).

3.2 Concreteness

Concreteness ratings were taken from Brysbaert et al. (2014), who collected crowd-sourced concreteness judgments on a 5-point scale. We excluded bigrams and retained only items known by at least 85 percent of raters. As with the AoA data, we augmented the concreteness lookup table with inflectional variants using `lemminflect`. These inflected forms were assigned slightly lower concreteness values than their base forms by subtracting a small constant (1×10^{-5}) from the base form's rating (e.g., *happiness* immediately follows *happy*)

to reproduce the treatment of inflected forms used for AoA.

3.3 Word Frequency

Lexical frequency was drawn from the `wordfreq` Python package, which compiles approximately 400,000 English words from many sources, including Google Books Ngrams 2012, Reddit, NewsCrawl 2014, and Wikipedia (Speer, 2022). We used language-wide frequency norms rather than corpus-specific frequencies to obtain a standardized estimate of word frequency independent of our training corpus.

For each word, we extracted its Zipf-scaled frequency value, which corresponds to the base-10 logarithm of occurrences per billion words.

3.4 Phonological Complexity

Phonological complexity was estimated using the `pronouncing` library, a simple interface for the CMU Pronouncing Dictionary. Each word was converted to its ARPAbet phoneme sequence. If more than one pronunciation is available, then the first one was retained.

From the phoneme sequence, we extracted the number of phonemes, the number of syllables, and the length of the largest consonant cluster. Phonological complexity was computed as a weighted sum of these components (see Formula 1 in Appendix C).

4 Data and curricula creation

The training corpus was constructed by combining all 10 million sentences from the BabyLM 2025 dataset with a random 10 percent subsample of the Refined BookCorpus, resulting in approximately 17 million sentences total (Charpentier et al., 2025; Singh, 2024).

The CHILDES (MacWhinney, 2000) portion of the BabyLM dataset was normalized: speaker tags and bracketed annotations were removed, and non-lexical placeholders for unintelligible speech were discarded. All text was lowercased and segmented into sentences using `spaCy` prior to merging.

To limit the influence of highly repeated material, sentence frequency was capped at five occurrences. This procedure follows DeChant and Bauer (2021), who similarly stripped the corpus of duplications.

4.1 Curricula creation

Curricula were built by sorting all sentences in the preprocessed corpus according to sentence-level scores, which were computed from a chosen lexical feature and aggregation method. We build on the curriculum construction methodology of DeChant and Bauer (2021), extending it by introducing alternative construction strategies and curricula based on additional lexical features beyond AoA.

Sentence-level lexical scores were computed for age of acquisition, lexical frequency, concreteness, and phonological complexity. Each token was assigned a lexical score using the corresponding lookup table, with missing entries excluded from scoring.[1] Stopwords were also excluded. Per-token scores were aggregated into a single sentence score using one of three methods: mean, minimum, or maximum. If a sentence contained no eligible scoring units, either because all tokens were stopwords or absent from the lookup tables, the sentence was not included in the resulting curriculum.

The ordered corpus is then divided into contiguous tranches, which determine when sentences are introduced during training and allowed us to capture snapshots of the embeddings at different stages. Tranches also define boundaries for local shuffling between training rounds. Originally, Dechant and Bauer capped tranches after 500 new unique words were introduced, however this meant that tranches could vary in size as a function of each word's occurrence frequency. To isolate the effects of the curricula from this tranching strategy with an enforced bias towards frequency of exposure, we introduce several other methods of tranche construction: Sentence-count tranches contain a fixed number of consecutive sentences, with total word counts varying by sentence length; Word-count tranches accumulate sentences until a target number of word tokens is reached; Finally, we construct matching tranches by aligning tranche sizes with those of a reference curriculum. In this last case, each tranche contains the same number of words as the corresponding tranche in a reference AoA curriculum. This procedure allows for direct comparisons against shuffled curricula while preserving the variable tranche sizes of the reference curriculum, rather than enforcing a fixed number of words or sentences across all tranches.

All curricula are built from the same preprocessed pool of sentences and differ only in their

[1] Excluding missing entries may bias sentence-level scores toward words with available lexical scores. However, this effect is limited in practice because later training ignores words occurring fewer than 20 times (`min_count = 20`). We also found minimal overlap between words absent from the lookup tables and words retained in the final training vocabulary.

ordering or tranche construction strategy.

5 Methodology

5.1 Training

We trained distributional word representations using the Word2Vec implementation in the `Gensim` library. Models were initialized with a fixed learning rate ($\alpha = 0.025$) and a vector dimensionality of 50. Training used the skip-gram architecture with negative sampling (20 negative samples per positive instance), a context window of 5, and a minimum count of 20 words. Prior to training, an initial vocabulary was constructed from the training data in order to initialize the model parameters and fix the set of trainable word types. These training settings match those used in the prior work of DeChant and Bauer (2021). For each curriculum, we performed five independent training runs, using different random seeds, resulting in different model initializations and sentence ordering within each tranche.

5.2 Stability metric

To evaluate the stability of word embeddings under different curriculum types, we measured k-nearest neighbor (kNN) overlap across multiple independent training runs. Wendlandt et al. (2018) define stability as the proportion of shared nearest neighbors for the same word across runs, providing a quantitative measure of embedding consistency. We used $k = 30$ to balance robustness and computational efficiency.[2] We computed the average stability over all words to get a single stability measure for the model.

We report this measure for embeddings after each tranche to show the evolution of the vector space during training. For all stability plots, we report the mean across five independent runs. Although 95 percent confidence intervals were computed, they were omitted from figures because they were not visually distinguishable (95th percentile width < 0.005 across all plots). When graphing, we plot a rolling average for easier analysis (see corresponding Appendix sections B and C for raw stability measures).

Additionally, we analyzed stability as a function of cumulative unique words introduced. By mapping tranche index to unique word counts, we compare embedding stability specifically as vocabulary grows.

[2]Wendlandt et al. found that using $k = 10$ performed similarly to substantially larger values of k. While DeChant and Bauer (2021) used $k = 20$ for kNN metrics, our larger corpus size warranted a larger overlap range.

6 Results

6.1 AoA curriculum and controlling for confounds

We first replicated the approach of DeChant and Bauer (2021) with our augmented data, where each tranche introduces 500 new words at a time, for both the AoA and shuffled curricula. While we observed improvements in stability, these effects were more limited than reported in the original study. In Figure 1, we observe that the AoA curriculum performs better only initially before performing in line with the shuffled curriculum around tranche 450. This is a novel observation likely due to the much larger size of our corpus. Another difference is the strong initial spike at the beginning for the shuffled curriculum. We suspect that this pattern arises from early-training artifacts that are amplified by the highly uneven distributions of sentences per tranche produced by this tranching method in the AoA and shuffled curricula (see Figures A.1 and A.2). See Figure B.1 in the appendix for further explanation.

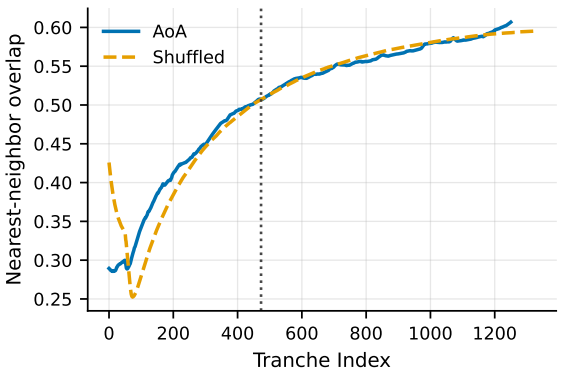

Figure 1: Proportion of nearest neighbor overlap (stability) of AoA and shuffled curricula across tranches introducing 500 new unique words. (Stability curves shown as rolling means; see raw: Fig B.1).

To isolate the effects of the AoA curriculum, we controlled for potential confounding factors. First, we noticed that AoA tranches are initially larger than those in the shuffled curriculum (see Appendix A). Because model stability is evaluated after each tranche, the model exposed to more data early may appear more stable. To address this, we created a matched shuffled curriculum in which tranche

sizes corresponded to the AoA curriculum. In this setup, the AoA curriculum continues to introduce 500 new words per tranche, while the shuffled curriculum introduces the same total number of words per tranche (see Appendix Figures B.2 and B.3 for results). We noticed that early improvements persisted. Additionally, both curves exhibit spikes at similar locations in Figure B.3, suggesting these fluctuations may be driven by the varying tranche size rather than by differences in lexical ordering.

Based on the previous observation, we controlled tranche sizes based on word counts. Each tranche introduced a fixed 40,000 word tokens. This approach produced smoother plots that precisely show where the stability curves intersect (see Appendix Figures B.4 and B.5). The crossover is earlier than in the previous graphs, so we zoom in on the first 20 percent of training and see that the early stability effect still persists. We also controlled tranche sizes based on sentence counts and observed similar results.

With the standardized 40,000-word tranches, we further plotted stability as a function of the cumulative number of unique words seen, rather than total words processed, shown in Figure 2, to evaluate model stability relative to vocabulary growth while still maintaining uniform training increments. With this method, the early stability improvements are more pronounced. Even when both curricula expose the model to the same number of unique word types, the AoA curriculum still shows higher early stability. This pattern indicates that the effect cannot be fully explained by differences in the rate of vocabulary growth alone, but may instead depend on which words are introduced earlier in training. Introducing earlier-acquired words first may help structure the embedding space in a way that supports more consistent representations as additional vocabulary is incorporated. Because this visualization makes the AoA effect most clearly observable, all subsequent decompositions are plotted using cumulative unique words on the x-axis for consistency and comparability.

Finally, we examined whether AoA itself contributes to model stability beyond general curriculum effects. With a descending AoA curriculum, in which later-acquired words were introduced first, a reverse effect was observed in Figure 3, confirming the importance of early-acquired words. To further isolate curriculum-building effects, we introduced a permuted-AoA curriculum by shuffling the scores in the AoA lookup table. The resulting stability patterns across multiple permutations were consistent and notably weaker than those observed in the original AoA curriculum, shown in Figure 4. Thus, in subsequent analyses, we will use the permuted-AoA curriculum as a model of general curriculum effects distinguishable from those attributable to early-acquired words.

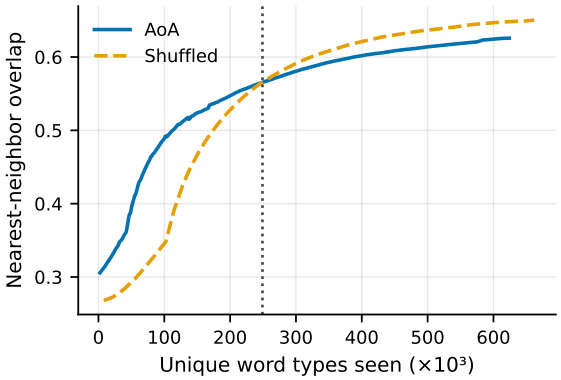

Figure 2: Stability of AoA and shuffled curricula with fixed 40k-word tranches plotted against unique words seen. (See raw: Fig B.6).

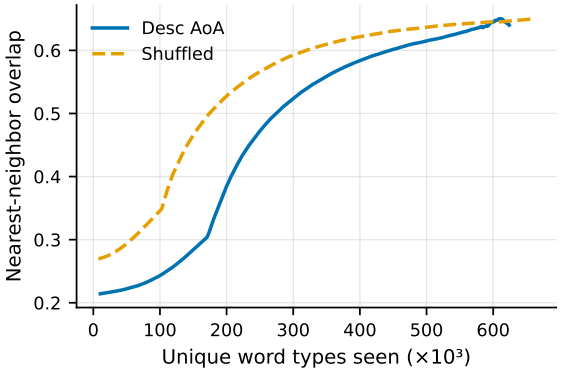

Figure 3: Stability of descending AoA and shuffled curricula with fixed 40k-word tranches plotted against unique words seen. (See raw: Fig B.7).

Overall, these results indicate that the AoA curriculum has an impact on embedding stability, even after controlling for tranche size, tranche construction method, and other curriculum-related effects.

6.2 Decomposition of AoA effects

Across developmental studies, AoA has been linked to lexical features such as frequency, concreteness, and phonological complexity (Braginsky et al., 2019; Verhagen et al., 2022). To understand the extent to which the observed AoA stability im-

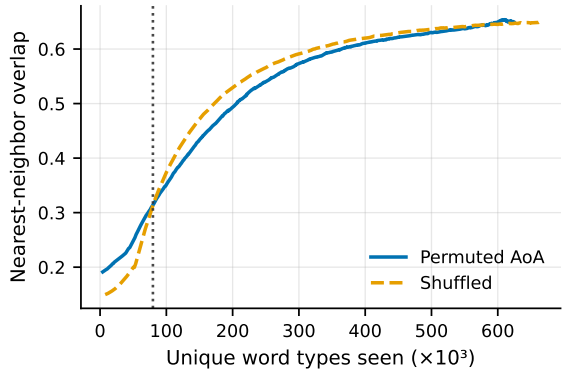

Figure 4: Stability of permuted-AoA and shuffled curricula with fixed 40k-word tranches plotted against unique words seen. (See raw: Fig B.8).

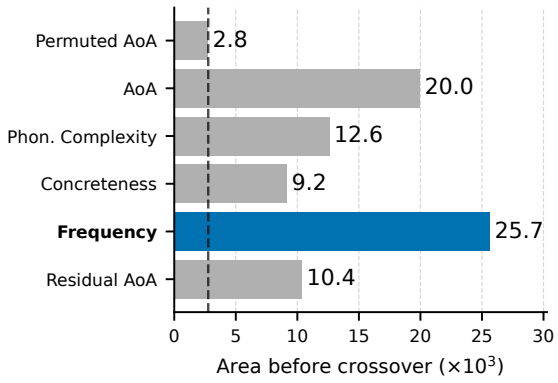

Figure 5: Pre-crossover area between curriculum and shuffled stability curves. Area values reflect the integrated stability advantage over cumulative unique words prior to crossover. (See plots in Appendix C).

provements are driven by these correlated variables, we constructed alternative curricula based on each lexical feature independently. Each curriculum was built analogously to the AoA condition and evaluated under identical settings as the AoA experiment that produced the most significant results, that is: mean aggregation, 40 thousand words per tranche, and plotted against unique words seen.

To quantify the persistence and magnitude of curriculum stability effects, we measured (1) the crossover point between the curriculum and shuffled stability curves and (2) the area between curves prior to crossover, calculated on the raw stability values. Statistics are summarized in Table 1 and Figure 5, respectively.[3]

Curriculum	Crossover Point (Unique Words)
Permuted AoA (baseline)	80k
AoA	249k
Phonological Complexity	138k
Concreteness	233k
Frequency	**312k**
Residual AoA	123k

Table 1: Crossover points between curriculum and shuffled stability curves, measured in cumulative unique words seen during training.

6.2.1 Phonological Complexity

We first constructed a curriculum ordered by increasing phonological complexity, with results shown in Appendix Figures C.1 and C.2. This manipulation did not reproduce the characteristic AoA pattern. The early-phase stability increase was substantially weaker than under the AoA condition, supported by the earlier crossover point and smaller pre-crossover area. This suggests that phonological complexity explains little of the AoA stability advantage.

6.2.2 Concreteness

Next, we constructed a curriculum ordered by decreasing concreteness, with results shown in Appendix Figures C.3 and C.4. Here, early acceleration in stability remained weaker than in the AoA curriculum, supported by the relatively small pre-crossover area. The crossover point, however, is closer to that observed under AoA ordering (16k word difference). This pattern indicates that concreteness may contribute more meaningfully to the AoA effect, although it does not fully account for it.

6.2.3 Frequency

We then constructed a curriculum ordered by decreasing frequency, with results shown in Figure 6. This condition most closely resembled the AoA results. Early improvement in stability was stronger than both the phonological complexity and concreteness conditions. The effect persisted the longest and with the greatest magnitude, slightly surpassing AoA.

A distinctive feature of the frequency curriculum was the presence of quasi-regular spikes, most pronounced during early training (see Figure C.5).

[3]While formal uncertainty estimates were not computed for the derived crossover and area statistics themselves, the narrow confidence intervals indicate that these values are stable estimates derived from the mean curves.

These spikes were not observed in other curricula. They arise naturally from introducing high frequency words at the very beginning of training. Because such words appear across many sentences, large portions of the embedding space are updated simultaneously.

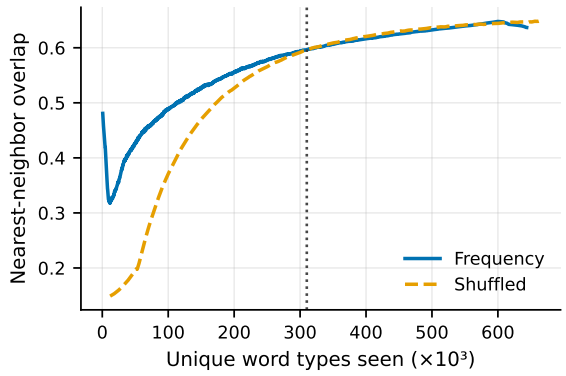

Figure 6: Stability of frequency-ordered and shuffled curricula with fixed 40k-word tranches plotted against unique words seen. (See raw: Fig C.5).

6.2.4 Residual AoA effects

To further isolate the contribution of these correlated features, we constructed a residual curriculum designed to remove their shared variance with AoA. Specifically, we trained a regression model to predict AoA values from frequency, phonological complexity, and concreteness. Linear, polynomial, and generalized additive models (GAMs) were considered, with GAMs best representing the non-linear data ($\rho = 0.62 \pm 0.01$, $R^2 = 0.41 \pm 0.01$; mean ± SD across 10-fold cross-validation). The predicted AoA values were subtracted from the original Kuperman norms, yielding residuals that reflect the portion of AoA not explained by these three features. These residual values were then used as the lookup table for curriculum construction.

The resulting residual curriculum (results in Figure 7) performed similarly to the permuted AoA baseline (see Figure 4). Although the early phase acceleration differed slightly, the crossover point remained relatively low with a difference of only 43k words, indicating that most of the AoA stability effect can be accounted for by frequency, concreteness, and phonological complexity. The remaining unexplained variance likely reflects additional lexical or cognitive variables not included in our model, as well as limitations of the available datasets.

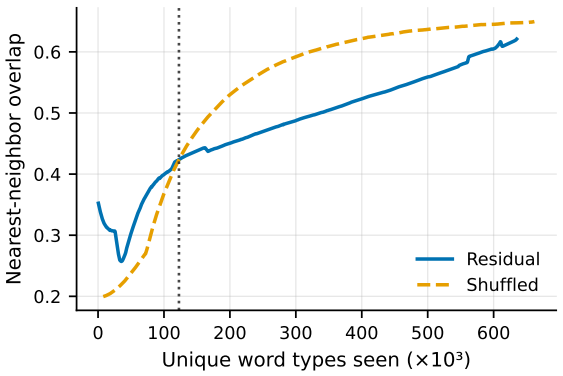

Figure 7: Stability of residual and shuffled curricula with fixed 40k-word tranches plotted against unique words seen. (See raw: Fig C.6).

7 Further analysis

7.1 Insights for language acquisition

AoA effects are a well-established phenomenon in the word recognition literature: words learned earlier in life are typically recognized and produced more quickly than those learned later (Izura and Ellis, 2002). Despite the consistency of this pattern, modern theories disagree on whether the AoA effect is due to AoA as an independent variable or as a byproduct of frequency, also called the cumulative-frequency hypothesis (Ghyselinck et al., 2004).

Ellis and Ralph (2000) supports a connectionist account, proposing that AoA effects emerge naturally in incremental learning systems because items encountered early are encoded under higher plasticity. Our results are consistent with this connectionist perspective. The permuted AoA baseline demonstrates that not all deterministic curricula produce equivalent benefits. Rather, which items are introduced early matters. At the same time, the frequency-based curriculum most closely approximates the AoA curriculum in its effects, suggesting that cumulative frequency plays a substantial role.

One mechanism through which AoA ordering exerts its influence appears to be the stabilization of local semantic neighborhoods. Our stability measurements directly support this by implying that earlier words function as consistent reference points for subsequently learned items. This is broadly consistent with growth-based accounts of semantic structure, such as the model proposed by Steyvers and Tenenbaum (2005), in which early nodes acquired central positions as the network expands.

With this interpretation, early acquired words act as semantic anchors that scaffold later representational organization.

To quantify this effect, we examined how often each word appears in other words' top-20 nearest neighbor lists. This measure captures a form of distributional centrality. The top 1 percent of words under the AoA curriculum appear on average 220.76 times in other words' nearest neighbor lists, compared to 196.01 under the shuffled curriculum (calculated with 3 sets of final embeddings from different runs). This measurable increase in centrality supports the semantic anchor hypothesis: early ordered training produces a set of more central lexical items.

Importantly, this notion of semantic centrality differs from semantic neighborhood density. Prior work has shown that the embeddings of early-acquired words tend to occupy relatively sparse semantic neighborhoods, potentially reducing lexical competition during acquisition (Alhama et al., 2020). In contrast, our centrality metric captures how frequently a word functions as a reference point within the larger semantic space. Early-acquired words may therefore be globally central while remaining locally sparse. This interpretation aligns closely with growth-based semantic network theories in which early nodes acquire increasingly central organizational roles as semantic systems expand (Steyvers and Tenenbaum, 2005; Hills et al., 2010). The fact that embeddings trained with an AoA curriculum exhibit greater semantic centrality suggests that they may be organizing a semantic space in a more human-like manner.

Our findings also align with stability-plasticity tradeoffs in incremental learning systems (Parisi et al., 2019). We notice a crossover point in Figure 2 that corresponds to about 250,000 unique words, 56,560,000 total tokens, or a mean sentence AoA score of 5.52 years of age. Before this point in training, AoA ordering produces greater representational stability during early vocabulary formation, suggesting rapid establishment of a coherent semantic space. However, as vocabulary size increases, shuffled curricula eventually match or exceed the stability achieved by AoA ordering. This reflects a tradeoff: early structured input promotes rapid and stable semantic organization with limited resources, but later in learning, a broader distribution of input may support flexibility.

Importantly, these parallels with developmental research are theoretical rather than biological claims. The observed dynamics reflect properties of incremental distributional learning systems, not direct evidence about neural mechanisms. Nevertheless, the convergence between embedding-based analyses and semantic network theories of lexical development suggests that the AoA effect may emerge naturally from incremental distributional learning systems.

7.2 Performance on domain-specific tasks

DeChant and Bauer (2021) reported that AoA-ordered curricula did not yield improvements on the intrinsic benchmarks of Simlex999 and wordsim-353 despite influencing training dynamics. We further investigate this by evaluating performance on a domain-specific task related to child language acquisition by assessing the ability of pretrained embeddings to predict AoA norms from Kuperman et al. (2012). We compare separate regression models trained with the embeddings generated from AoA and shuffled curricula while evaluating performance with leave-one-out cross validation, reporting both mean absolute deviation (MAD) in months of predicted AoA to actual AoA estimate and Spearman correlation. To better reflect real-world applications, we used 300-dimensional embeddings. Both linear regression and GAMs were considered, although GAMs better captured the non-linear relationship in the data. Under the AoA curriculum, the model achieved a Spearman correlation of 0.926 and MAD of 10.86 months, whereas the shuffled curriculum produced a Spearman of 0.555 and MAD of 22.61 months, corresponding to a reduction in prediction error of approximately 52 percent. These results are shown in Figure 8.

To verify that the successful predictions of the AoA-based embeddings were due to the semantic features and not artifacts of the curriculum building, we controlled for potential timestamp effects encoded in the embeddings. Following the debiasing procedure of Ravfogel et al. (2020), we trained a linear ridge regression model to predict the final embeddings from the tranche index where a word was introduced. Subtracting this predicted component from the AoA embeddings removes timestamp effects while preserving semantic information. After debiasing, performance remained strong (see Appendix D), with MAD increasing slightly to 11.02 months. This still represents a 51.3 percent relative improvement over the shuffled curriculum. The persistence of the effect indicates that the advantage is not merely due to residual traces of

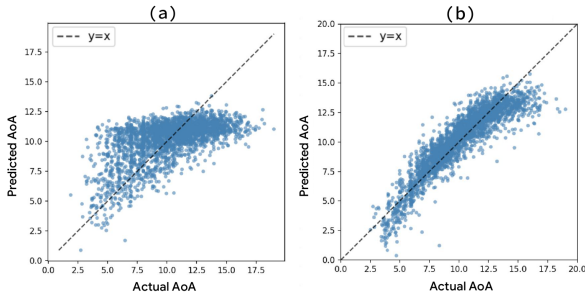

Figure 8: Predicting human AoA norms from word embeddings under different training curricula (predicted vs. actual AoA). (a) Embeddings trained with the **shuffled curriculum** yield a Spearman correlation of $\rho = 0.548$ and a mean absolute deviation (MAD) of 22.61 months. (b) Embeddings trained with the **AoA-ordered curriculum** yield a substantially stronger correspondence with human norms, with $\rho = 0.926$ and MAD of 10.86 months.

the Kuperman dataset during training, but rather reflects structural properties induced by AoA-ordered curriculum learning itself. While prior work has predicted AoA through exposure trajectories over increasingly difficult corpora (Botarleanu et al., 2022), our results show that the final embeddings themselves retain information useful for predicting developmental ordering. Thus, for domain-specific tasks related to child language acquisition, AoA-ordered curricula produce embeddings that capture relevant semantic structures more effectively than shuffled curricula.

Because AoA-based embeddings appear to capture developmentally meaningful semantic structure, they may also prove useful in cognitively motivated evaluation frameworks. For example, CogniVal (Hollenstein et al., 2019) provides a benchmark suite for assessing whether embeddings reflect human semantic representations by using them to predict brain activity data. In addition, these embeddings may be useful in cognitive metrics that rely on vector similarity, such as the divergent association task by Olson et al. (2021) which uses semantic distance between embeddings to measure creativity in people.

8 Conclusion

This study investigated AoA as a curriculum learning signal for distributional word embeddings. After replicating the stability advantages of AoA-ordered curricula reported by DeChant and Bauer (2021), we extended their findings controlling for underlying curriculum learning effects and by decomposing AoA into its underlying lexical components. We showed that AoA ordering consistently produces early-phase stability advantages over shuffled baselines. Moreover, reverse and permuted AoA curricula confirmed that this effect depends specifically on which words are introduced early, rather than on deterministic ordering alone. Our decomposition analysis revealed that much of the AoA advantage can be explained by correlated lexical features, particularly frequency and, to a lesser extent, concreteness. When shared variance was removed, residual AoA effects were substantially reduced, suggesting that AoA operates largely through these associated lexical properties. Additionally, we demonstrate that, although AoA curriculum embeddings have not previously led to consistent gains on general embedding benchmarks, they yield substantial gains on a domain-specific task: AoA norm prediction, where they lead to 52 percent reduction in prediction error over comparable regular embeddings.

These findings provide evidence that distributional learning mechanisms mirror core theories of human language acquisition. Early-learned, high-frequency, and concrete words appear to function as semantic anchors that stabilize local neighborhoods during the rapid early phase of vocabulary growth. As training progresses and vocabulary size increases, this early advantage diminishes, reflecting a stability–plasticity tradeoff characteristic of incremental learning systems. Exceptionally, frequency is the only curriculum learning method, among those tested, that increases early stability without sacrificing late-stage stability. This suggests that frequency-based curriculum construction may be beneficial for training on general text corpora.

More broadly, our results suggest that curriculum structure influences the trajectory through which representational spaces emerge. Understanding how input ordering shapes learning dynamics may therefore offer a principled bridge between computational models and theories of human vocabulary development.

We show that curriculum learning can provide a principled framework for modeling developmental constraints in artificial systems, even when it does not yield improvements on standard intrinsic benchmarks.

Limitations

Several limitations qualify our conclusions. First, our analysis is restricted to static Word2Vec embeddings; results may differ for contextualized models or alternative architectures. Second, the lexical features used to decompose AoA effects are themselves imperfect proxies. Frequency estimates are corpus-aggregated rather than developmentally grounded, phonological complexity relies on dictionary pronunciations, and AI-augmented AoA norms may introduce systematic bias. Finally, our findings concern distributional learning dynamics and should not be interpreted as direct claims about neurocognitive mechanisms of language acquisition.

Future work should evaluate AoA curricula in neural language models with contextualized representations, incorporate child-directed frequency measures, and explore multilingual settings where acquisition trajectories differ.

Acknowledgments

We acknowledge the support of IVADO, whose funding made this work possible. EP is an IVADO Professor. IG, AS, CC and TKB were supported by an IVADO R3AI Regroupment 3 grant. We thank our CoNLL reviewers for their careful reading and constructive feedback. Finally, we are grateful to the McGill AI Lab for fostering the collaborative environment that brought the four authors together.

References

Raquel G. Alhama, Caroline Rowland, and Evan Kidd. 2020. Evaluating word embeddings for language acquisition. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 38–42, Online. Association for Computational Linguistics.

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Andreas Baumann and Stefan Hartmann. 2026. The chicken and the egg: unraveling aspects of semantic change and how they relate to lexical acquisition. *Cognition*, 266:106301.

Robert-Mihai Botarleanu, Mihai Dascalu, Micah Watanabe, Scott Andrew Crossley, and Danielle S. McNamara. 2022. Age of exposure 2.0: Estimating word complexity using iterative models of word embeddings. *Behavior Research Methods*, 54(6):3015–3042.

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

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.

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.

Marc Brysbaert, Ilse Van Wijnendaele, and Simon De Deyne. 2000. Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, 104(2):215–226.

Giovanni Cassani, Federico Bianchi, and Marco Marelli. 2021. Words with consistent diachronic usage patterns are learned earlier: A computational analysis using temporally aligned word embeddings. *Cognitive Science*, 45(4):e12963.

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

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams. 2025. Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data. In *Proceedings of the First BabyLM Workshop*, pages 399–420, Suzhou, China. Association for Computational Linguistics.

Chad DeChant and Daniel Bauer. 2021. Learning word representations in a developmentally realistic order. Technical report, Columbia University.

Andrew Ellis and Matthew Ralph. 2000. Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26:1103–1123.

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: ACL 2023*, pages 12205–12218, Toronto, Canada. Association for Computational Linguistics.

Mandy Ghyselinck, Michael B Lewis, and Marc Brysbaert. 2004. Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica*, 115(1):43–67.

Clarence Green, Anthony Pak-Hin Kong, Marc Brysbaert, and Kathleen Keogh. 2025. Crowdsourced and ai-generated age-of-acquisition (aoa) norms for vocabulary in print: Extending the kuperman et al. (2012) norms. *Behavior Research Methods*, 57(11):304.

Johannes Hellrich and Udo Hahn. 2016. Bad Company—Neighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2785–2796, Osaka, Japan. The COLING 2016 Organizing Committee.

Thomas T. Hills, Josita Maouene, Brian Riordan, and Linda B. Smith. 2010. The associative structure of language: Contextual diversity in early word learning. *Journal of Memory and Language*, 63(3):259–273.

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019. CogniVal: A framework for cognitive word embedding evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 538–549, Hong Kong, China. Association for Computational Linguistics.

Eghbal A. Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. 2024. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, 5(1):43–63.

Cristina Izura and Andrew Ellis. 2002. Age of acquisition effects in word recognition and production in first and second language. *Psicológica*, 23.

Barbara J. Juhasz. 2005. Age-of-acquisition effects in word and picture identification. *Psychological Bulletin*, 131(5):684–712.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4):978–990.

Megan Leszczynski, Avner May, Jian Zhang, Sen Wu, Christopher R. Aberger, and Christopher Ré. 2020. Understanding the downstream instability of word embeddings. *CoRR*, abs/2003.04983.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Jay A. Olson, Johnny Nahas, Denis Chmoulevitch, Simon J. Cropper, and Margaret E. Webb. 2021. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25):e2022340118.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Nishant Singh. 2024. Refined bookcorpus dataset.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Mark Steyvers and Joshua B. Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78.

Josje Verhagen, Mees Van Stiphout, and Elma Blom. 2022. Determinants of early lexical acquisition: Effects of word- and child-level factors on dutch children's acquisition of words. *Journal of Child Language*, 49(6):1193–1213.

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102, New Orleans, Louisiana. Association for Computational Linguistics.

A Tranche Sizes

Here we report tranche sizes, measured in number of word tokens, for the curricula used in DeChant and Bauer (2021). As shown in Figure A.1, the AoA-ordered curriculum exhibits relatively larger tranche sizes at the beginning of training. In contrast, the shuffled curriculum displays a more gradual and approximately linear increase in tranche size across training, as shown in Figure A.2. This difference reflects the distributional properties of the underlying curricula when each tranche introduces 500 new unique word types.

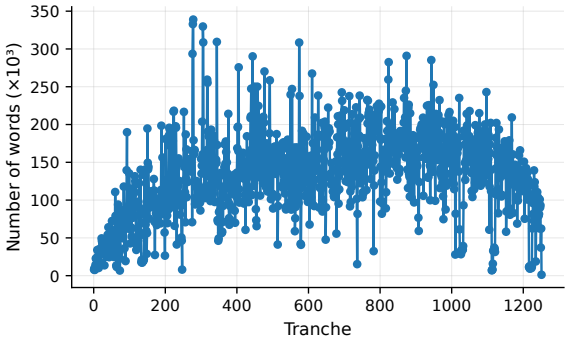

Figure A.1: Total number of words per tranche in the AoA-ordered curriculum, with 500 new unique words introduced per tranche.

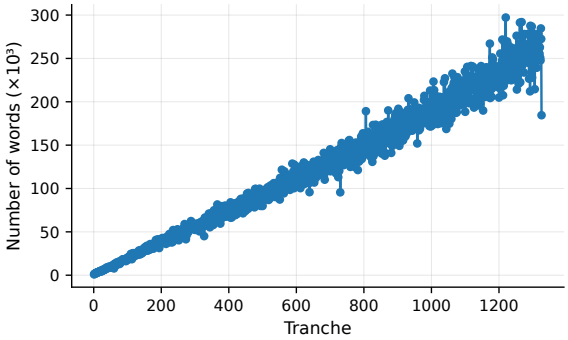

Figure A.2: Total number of words per tranche in the shuffled curriculum, with 500 new unique words introduced per tranche.

B Controlling for Confounds

This section presents the stability plots referenced in Section 6.1. The primary analyses in the main text use rolling-average smoothing to improve visual interpretability and to facilitate estimation of crossover points between curricula. This appendix section includes the corresponding raw stability curves for all plots discussed in the main text, as well as smoothed and raw versions of all other referenced stability plots.

We include the unsmoothed stability curves in Appendix Sections B and C for transparency and to highlight certain properties of the graphs. In particular, the replica AoA condition of DeChant and Bauer (2021) shown in Figure B.1 exhibits more local spikes throughout relative to the AoA results controlled by tranche-size shown in Figure B.4. Additionally, the frequency-ordered curriculum shows more uniform spikes across training shown in Figure C.5, which we attribute to the fact that highly frequent words occur in diverse contexts, resulting in regular embedding space reorganizations.

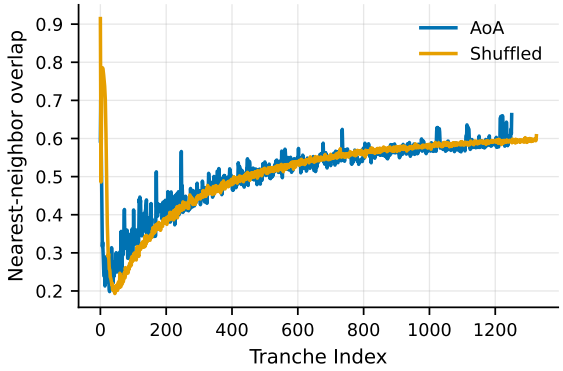

Figure B.1: Raw stability of AoA and shuffled curricula across tranches introducing 500 new unique words (see smoothed in Figure 1). We disregard the initial spike when analyzing curves, as it is an artifact of early optimization rather than a curriculum-specific effect. During early training, embedding updates are dominated by low-level distributional signals such as word frequency and global co-occurrence statistics, which remain relatively consistent across independent runs before richer semantic structure emerges. This effect is likely amplified by the very small tranche sizes used in the initial experiments (see Figure A.2).

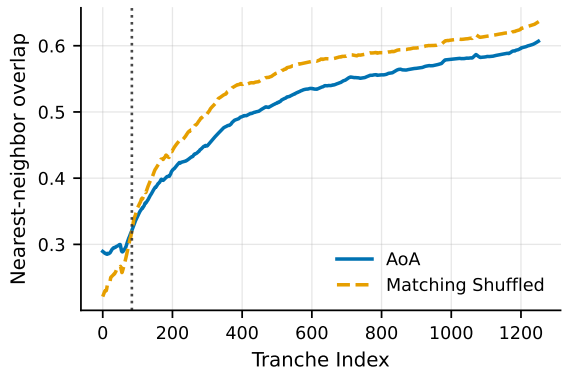

Figure B.2: Stability across AoA tranches introducing 500 new unique words. Each shuffled tranche has the same total word tokens as each corresponding AoA tranche. (See raw: Fig B.3).

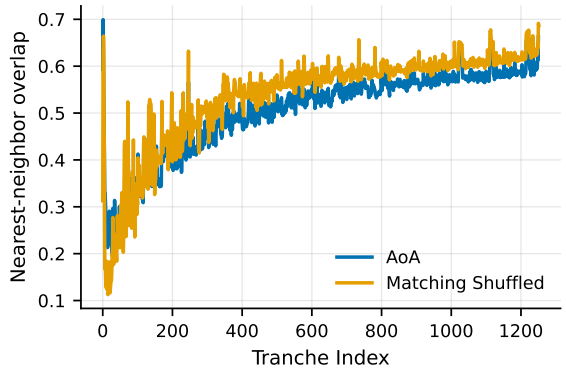

Figure B.3: Raw stability across AoA tranches introducing 500 new unique words. (See smoothed: Fig B.2).

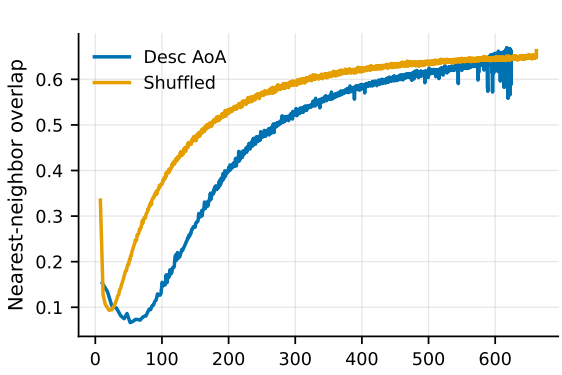

Figure B.6: Raw stability of AoA and shuffled curricula with fixed 40k-word tranches plotted against unique words seen. (See smoothed: Fig 2).

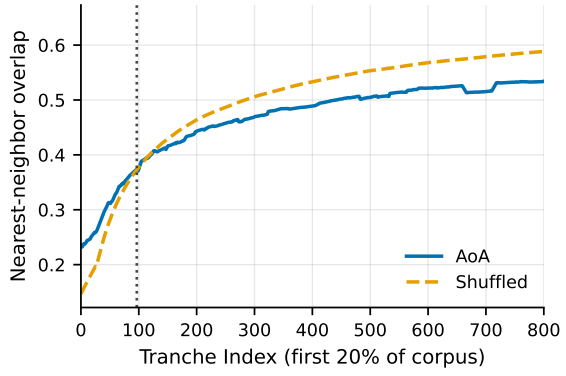

Figure B.4: Stability of AoA and shuffled curricula across fixed 40k-word tranches. To highlight early-stage dynamics, the x-axis is restricted to the first 20 percent of the curriculum. (See raw: Fig B.5).

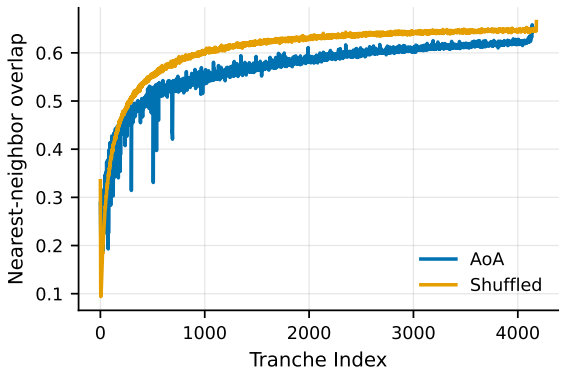

Figure B.5: Raw stability of AoA and shuffled curricula with fixed 40k-word tranches. (See smoothed: Fig B.4).

Figure B.7: Raw stability of descending AoA and shuffled curricula with fixed 40k-word tranches plotted against unique words seen. (See smoothed: Fig 3).

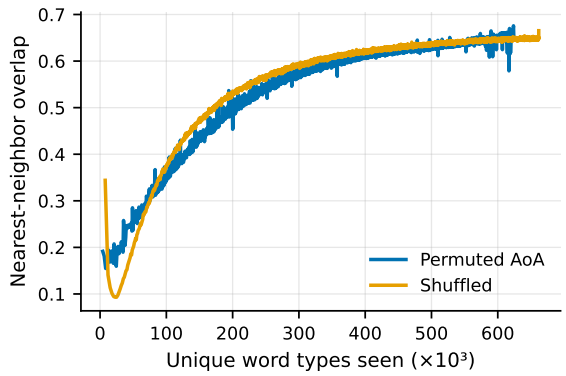

Figure B.8: Raw stability of permuted-AoA and shuffled curricula with fixed 40k-word tranches plotted against unique words seen. (See smoothed: Fig 4).

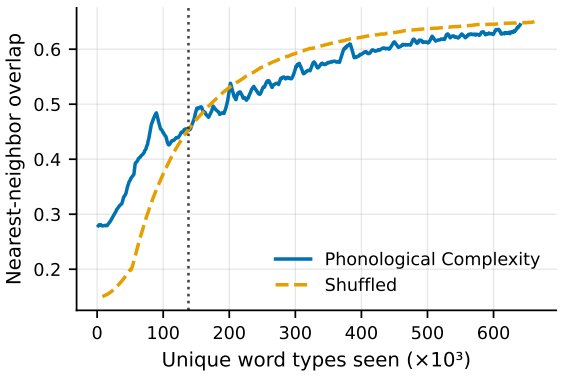

Figure C.1: Stability of phonological-complexity-ordered and shuffled curricula with fixed 40k-word tranches plotted against unique words seen. (See raw: Fig C.2).

C Decomposing AoA Effects

Here we report results for curricula ordered by phonological complexity, concreteness, and frequency, as well as the residual curriculum.

We compute phonological complexity for individual words using the formula:

$$\mathrm{PC} = N_{\text{phonemes}} + 0.5\, N_{\text{syllables}} + 1.5\, C_{\max} \quad (1)$$

where N_{phonemes} is the number of phonemes, $N_{\text{syllables}}$ is the number of syllables, and $C_{\max}$ is the maximum consonant cluster length.

We selected these weights heuristically to reflect the intuition that consonant cluster complexity contributes more strongly to articulatory difficulty than syllable or phoneme count alone, while phoneme count serves as the primary measure of phonological length. We did not adjust these weights to maximize stability or correlation with AoA, and therefore the resulting metric should be interpreted as an approximate proxy rather than a linguistically validated measure.

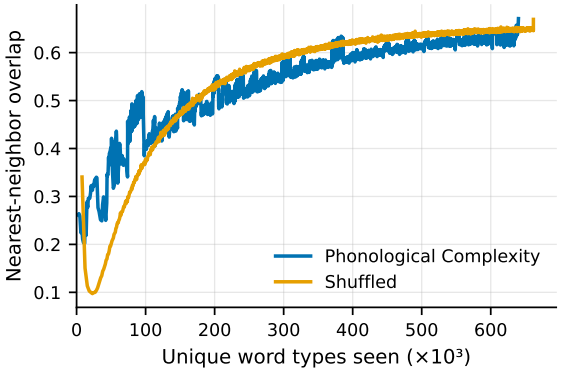

Figure C.2: Raw stability of phonological-complexity-ordered and shuffled curricula with fixed 40k-word tranches plotted against unique words seen. (See smoothed: Fig C.1).

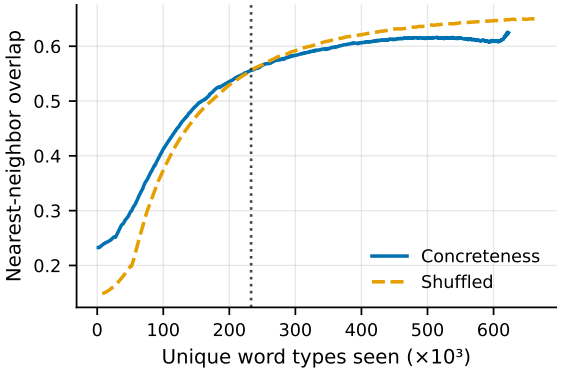

Figure C.3: Stability of concreteness-ordered and shuffled curricula with fixed 40k-word tranches plotted against unique words seen. (See raw: Fig C.4).

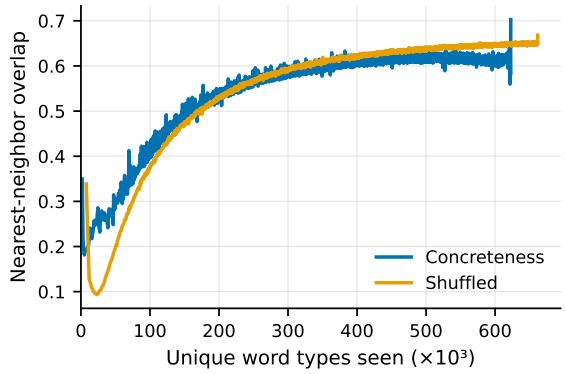

Figure C.4: Raw stability of concreteness-ordered and shuffled curricula with fixed 40k-word tranches plotted against unique words seen. (See smoothed: C.3).

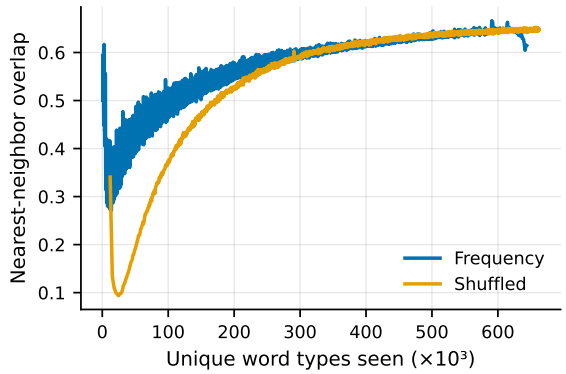

Figure C.5: Raw stability of frequency-ordered and shuffled curricula with fixed 40k-word tranches plotted against unique words seen. (See smoothed: Fig 6).

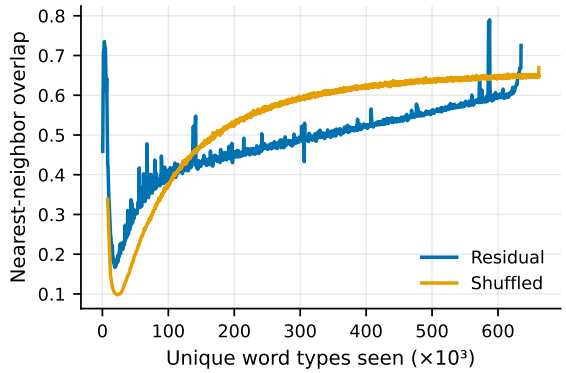

Figure C.6: Raw stability of residual and shuffled curricula with fixed 40k-word tranches plotted against unique words seen. (See smoothed: Fig 7).

D Domain-Specific Tasks

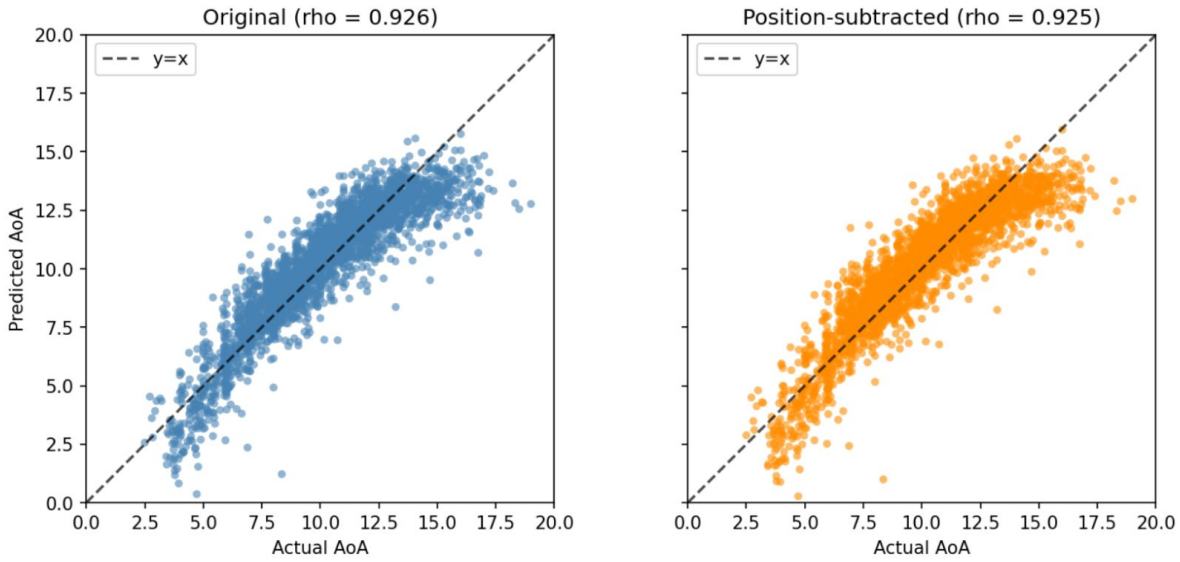

Figure D.1: Predicting human AoA norms from timestamp-debiased AoA-ordered embeddings (predicted vs. actual AoA). The left panel shows $\rho = 0.926$ and the right panel shows $\rho = 0.925$. The nearly identical predictive performance after removing the tranche-index (timestamp) component demonstrates that the embeddings' strong alignment with human AoA norms is not driven by residual curriculum artifacts, but instead reflects semantic structure induced by AoA-ordered curriculum learning itself.