Curriculum Development Using Age of Acquisition and Surprisal

Siddharth Sriraman (ssriraman8@gatech.edu)
Georgia Institute of Technology,
Atlanta, GA, USA

Manognya Bhattaram (mbhattaram3@gatech.edu)
Georgia Institute of Technology,
Atlanta, GA, USA

Vishnu Krishnan (vishnuk@gatech.edu)
Georgia Institute of Technology,
Atlanta, GA, USA

Shanmukh Karra (shanmukh.karra@gatech.edu)
Georgia Institute of Technology,
Atlanta, GA, USA

Abstract

The Age of Acquisition (AoA) of a word refers to an estimate of the age at which children produce the word with a probability greater than some threshold, i.e., the age at which children learn a particular word. Existing databases of children's vocabulary development, such as Stanford's Wordbank repository (Frank et al., 2016), only contain words with AoAs ranging from 16 to 30 months. By training language models on child-directed speech, researchers have shown that a model's average surprisal for a given word aids in predicting AoA of that word beyond traditional predictors (Portelance et al., 2023). Our project trains language models (LMs) on AO-CHILDES (Huebner & Willits, 2021), a child-directed speech dataset, to model any word's AoA using its average surprisal, extending beyond the words present in Wordbank. We then build a curriculum (Bengio et al., 2009) based on increasing AoA to train LMs such as BabyBERTa (Huebner et al., 2021) with the ordering of words mirroring a child learning English. The AO-CHILDES dataset also provides a ground truth curriculum to use.

Keywords: curriculum learning; large language models; age of acquisition

Introduction

Over the last few years, language models (LMs) have become increasingly complex and powerful, and have demonstrated state-of-the-art performance in many natural language processing (NLP) tasks. With this increasing complexity, training LMs to achieve optimal performance has become time-consuming and expensive. There has been extensive research in designing more efficient hardware and smarter algorithms to alleviate this. Our project takes a different view: using a cognitive science-backed approach to improve training performance. An area that has not been as widely studied is how samples are fed into models while training. Our project relies on combining two key concepts: age of acquisition and curriculum learning, to design a training strategy that mimics child-like learning.

Across multiple languages, it has been shown that children learn language in a consistent manner, specifically in the ordering of the words they learn (Łuniewska et al., 2015). This phenomenon is quantified through a word's Age of Acquisition (AoA), the estimated age at which a child is said to have learned a word. It can be more concretely defined as the age at which the probability of a child uttering that word exceeds some threshold. Recent research (Portelance et al., 2023) has shown that models such as Long Short-Term Memory (LSTMs) can characterize AoA through its surprisal values, and can aid in AoA prediction beyond traditionally studied predictors. Our first goal is to build a reasonable predictor of AoA through language model surprisal values.

NLP models are traditionally trained by feeding samples of tokenized text from a corpus in a random order. Constructing an ordering of samples where easier-to-learn samples are learned first before more difficult ones has been shown to improve model performance (Xu et al., 2021). This training strategy is referred to as curriculum learning. While most curriculum learning methods rely on heuristics and model-based metrics like error to define sample complexity, our project uses a psycholinguistic factor, i.e AoA, to build a curriculum. The second goal of the project is to build an AoA-ordered curriculum of documents from any corpus that mimics the order in which a child would learn a language. We study the extent to which this ordering can be achieved purely from surprisal values, without the need for any ground truth AoA values for words.

Related Work

AoA Prediction

There has been extensive research in cognitive science on both obtaining ground truth AoA for words and on the factors which shape a word's AoA.

Wordbank, a repository of child vocabulary growth, provides the trajectory of learning of words by children across 38 languages. For each word, it contains data on the proportion of children at each age between 16 to 30 months who produce that word. A simple measure of a word's AoA would be the youngest age at which over 50% of children produce that word. But many words do not reach this 50% mark in the given age range. So, they instead model an estimated AoA for each word using a Bayesian Generalised Linear Model. We use these as ground truth AoAs for building our generalized AoA predictor in this project, which extends beyond the words present in Wordbank.

For AoA prediction, traditional methods model AoA as a function of lexical properties of a word, such as the mean length of the utterances (MLU) it occurs in, concreteness, part of speech, frequency of occurrence (Braginsky et al., 2019). Among these, frequency is a strong predictor of AoA, words that are more frequent are learnt before those that are less frequent. Words with longer MLUs also tend to have larger AoAs (Goodman et al., 2008). Both these observations, however, depended on the part of speech of the word.

Recent works have focused on using surprisal values from language models as a proxy for AoA (Portelance et al., 2023). They used Long Short-Term Memory (LSTM) and n-gram models to obtain word surprisals. While surprisals were not a strong predictor of AoA on its own, it improved prediction beyond the previously known AoA predictors described above, but only by a small degree. But when controlling for part of speech, surprisals positively correlated with function words and predicates in many languages.

There has also been research in defining AoA based on the development of surprisal values when training language models (Chang & Bergen, 2021). As models get trained, the surprisal values for a word reduces till it reaches its optimal minimum value. They defined AoA for a word as the training step count at which a word reaches a surprisal halfway between its minimum and maximum value. They showed that this AoA correlates well with traditional predictors like log-frequency, MLU and concreteness, but they do not directly compare it against the ground truth AoA of these words. Our project extends this to build an AoA predictor based on training step count.

Curriculum Learning

The original work introducing curriculum learning defined an ordering purely using a rule-based difficulty criterion (Bengio et al., 2009). Other methods include manual labeling of samples (Pentina et al., 2015), which is time-consuming.

Another way of forming a curriculum involves defining difficulty based on the errors by other models trained on that sample (Xu et al., 2021). This cross-review method is a generalized way to build a curriculum where a dataset is split into N subsets, each of which is a stage of a curriculum. A separate model is trained on each subset. The difficulty of each example is evaluated through the sum of F1 scores or errors of each of the other *N-1* models not trained on this example.

Another aspect of curriculum learning involves how examples are sampled as training progresses. One way to do this, for example, is to only sample examples from up to the given stage when processing a particular stage of a curriculum. This ensures that in the initial training steps, only easier examples are processed, and progressively more difficult examples are sampled as training continues.

In the context of language learning, there has also been research done in building child-like curricula. AO-CHILDES is a corpus consisting of 3,304 child-directed American-English speech transcripts, consisting of nearly 5 million words. AO-CHILDES already provides a ground truth age-ordering of documents, which we use to evaluate our curriculum ordering against. The BabyBERTa LM (Huebner et al., 2021) tested curriculum learning using the AO-CHILDES dataset, and found that a random sampling baseline, which is not curriculum ordered, still outperforms curriculum learning. This is due to the lack of diversity of vocabulary when following the curriculum. However, they noted that an age-ordered curriculum performs significantly better than a reverse age-ordered one.

However, recently, the BabyBERTa model trained with an age-ordered curriculum combined with a medium- and high-complexity corpus has been shown to perform consistently better on BLiMP tasks than a random sampling baseline (Opper et al., 2023).

Method

AoA Prediction

To predict AoA as a function of model surprisal, we first train a language model on AO-CHILDES to obtain the average surprisal of a word over all its occurrences in the corpus. Let C be the set of documents in the corpus that a token v occurs in. For a language model characterized by the probability distribution P over tokens, the surprisal of v is given by:

$$surp(v) = \frac{1}{|C|} \sum_{w \in C;\, w_i = v} -\log P(w_i \mid w_{i-1}, w_{i-2} \ldots w_1)$$

where w is a document (sequence of tokens) in C.

We used two methods to obtain the AoA value based on surprisal:

- **Final AoA**: Extract the average surprisal values over the corpus for each token on the final trained model.
- **Midpoint AoA**: The method used by Chang et al. (2022), where we track the development of average surprisal for each token as the model gets trained. We then find the training step count at which the surprisal reaches halfway between a random

sampling baseline (no training) and its minimum value once trained.

The final AoA method uses the surprisal value as the AoA scale, while the midpoint AoA uses step count as the scale. Similar to Chang et al. (2022), we trained a bidirectional LSTM model from scratch to obtain surprisals. We also tested with Pythia (Biderman et al., 2023), a pre-trained large language model (LLM) to get surprisal values. We used the 14M-parameter, 6-layer configuration of the model.

Curriculum Building

After predicting the AoA values for each word and storing them, the next step was to build a curriculum by ordering the sentences of AO-CHILDES. The challenge was to order the sentences ensuring that the sentences containing words with lower AoA values would appear at the start of the ordering, and sentences containing words with higher AoA values would appear at the end of the ordering. To achieve this, we implemented three strategies:

- Sentence AoA based on token with maximum AoA.
- Sentence AoA based on average AoA of tokens in the sentence (without removing any stopwords).
- Sentence AoA based on average AoA of tokens in the sentence (removing stopwords with NLTK).

The first strategy was quite straightforward and prioritizes the most complex word in the sentence. The sentence AoA is simply the AoA of the word with the highest AoA value within the sentence. This strategy assumes that mastering the most challenging word will facilitate understanding the entire sentence, but ignoring the AoA of the other words in the sentence may lead to an incorrect estimate of the overall difficulty.

In the second strategy, the AoA of a sentence is calculated as the average AoA of all its words. This method takes into account the contribution of all the words in the sentence to the sentence's overall complexity. However, this strategy may result in an undesired reduction of sentence AoA as stopwords ("the", "is", "and", etc.) would likely have lower AoA value than regular words.

The third strategy is similar to the second strategy, but it excludes stopwords. Stopwords are words that are filtered out before processing natural language data, and NLTK provides a list of common English stopwords. This strategy thus prioritizes words in the sentence that carry more meaning.

Results

AoA Prediction

We trained a bidirectional LSTM (BiLSTM) with a single hidden layer and an embedding size of 256 for 10,000 steps. An example of the midpoint AoA method for the word "cat" is shown in Figure 1. Chang et al. (2022) also provide the midpoint and final AoAs for BiLSTMs, GPT2 and BERT models. The only difference is that those AoAs were computed based on surprisals over the WikiText-3 (Merity et al., 2017) and BookCorpus (Zhu et al., 2015) datasets. Since Pythia is pre-trained on the Pile dataset (Gao et al., 2020), we only obtain the final AoA values.

Once we compute the surprisal-based AoAs, we compute their Pearson correlation with the ground truth AoAs we obtained from Wordbank. We tabulated the results of different models in Table 1.

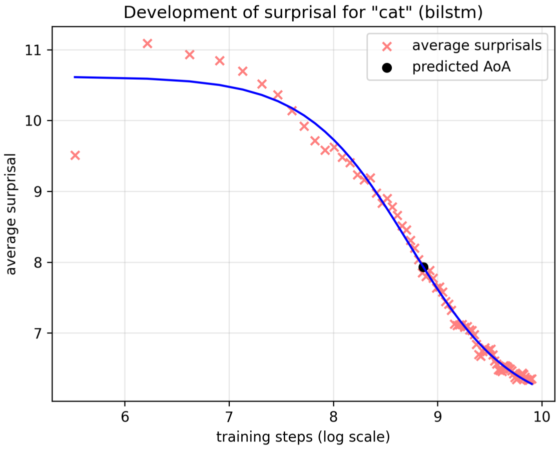

Figure 1: Development of surprisal values during training (a sigmoid curve is fit to find the step at which surprisal is halfway to its minimum from the starting maximum value).

Table 1: Pearson correlation of different models and AoA methods with Wordbank ground truth AoA (all words).

Model	**Final AoA Method**	**Midpoint AoA Method**
BiLSTM (AO-CHILDES)	0.181 (0.01)	0.084 (0.27)
BiLSTM*	**0.275 (0.0)**	**0.273 (0.0)**
GPT-2*	0.246 (0.0)	0.266 (0.0)
BERT*	**0.272 (0.0)**	**0.286 (0.0)**
Pythia 14M (Pile)	0.085 (0.12)	-

* trained on WikiText-3 and BookCorpus

Lower AoAs usually refer to easier words to learn that should have lower average surprisals. The values we originally obtained were showing reasonable but negative correlation. We suspect this could be due to an implementation detail where the values might have been negated at some point. To account for this, we negated them again in Table 1.

We used Pearson correlation since our goal is not to build an accurate predictor of AoA, rather just a variable that correlates well with AoA to allow us to build a curriculum ordering.

Clearly, the BiLSTM and BERT models generated AoAs that correlated most strongly with the ground truth AoAs. As expected, the BiLSTM trained on just AO-CHILDES did not perform as well as the one trained on a much larger corpus. We also do not see a very large variance in correlation when using the midpoint AoA method versus the final AoA method. The Pythia model surprises did not correlate as well with the ground truth AoA, this might be due to the nature of the dataset it was trained on.

Table 2: Pearson correlation of different models and AoA methods with Wordbank ground truth AoA (nouns and function words).

Model	Final AoA Method	Midpoint AoA Method
BiLSTM (AO-CHILDES)	-0.294 (0.0008)	-0.085 (0.38)
BiLSTM*	**-0.322 (0.0)**	**-0.328 (0.0)**
GPT-2*	-0.290 (0.0)	**-0.329 (0.0)**
BERT*	**-0.318 (0.0)**	**-0.341 (0.0)**
Pythia 14M (Pile)	0.086 (0.21)	-

* trained on WikiText-3 and BookCorpus

Looking deeper into the BiLSTM model, we also drew some interesting insights when analyzing correlation by the lexical class of the word. When controlling for nouns and function words alone, we observed stronger negative correlation of the surprisals of these models with the ground truth AoA. This was seen across both the AoA methods, as shown in Table 2 (correlations lower than -0.30 are highlighted). The BERT model showed the highest negative correlation of -0.34. When controlling for just predicate words, we saw correlation near zero with the ground truth AoA. Hence, the learning of function words and nouns are predictable by AoA, but not predicate words. Overall, combining all lexical classes, AoAs based on surprisals are reasonably well-correlated with ground truth AoAs to facilitate building age-ordered curricula.

Curriculum Building

The AO-CHILDES dataset provides built-in ground truth AoA values for each sentence, which serve as a ground truth curriculum. To investigate the effectiveness of the three curriculum building (sentence-ordering) strategies, we measured the similarity in ranking between each of the three curriculums and the ground truth curriculum by using Spearman's correlation coefficient as the comparison metric. To get the word AoAs, we used the BiLSTM (trained on WikiText-3 and BookCorpus) with the Final AoA method shown above. The results we obtained are shown in Table 3.

In all three cases the p-values were very close to zero, so the correlations were statistically significant. The strategy of calculating sentence AoA based on the token with the maximum AoA resulted in a Spearman correlation of 0.15. This suggests a very weak positive correlation, indicating that this strategy is the least effective among the three in aligning with the ground truth curriculum. This is expected, as it does not take into account the distribution of AoAs in the sentence, i.e, a single token with a larger AoA in a long sentence can skew the entire sentence AoA to be large.

Strategies two and three (average AoA of all words) outperformed strategy one, emphasizing the importance of AoA analysis beyond just the most complex word. The strategy of calculating sentence AoA based on the average AoA of all the words in the sentence resulted in a Spearman correlation of 0.54, which suggested a moderately positive correlation, indicating that this strategy is the most effective among the three in aligning with the ground truth curriculum.

Surprisingly, the strategy of calculating sentence AoA based on the average AoA of the words in the sentence but with stopwords removed using NLTK only resulted in a Spearman correlation of 0.36. We expected the removal of the stopwords to provide the highest correlation, as we thought the stopwords (which have lower AoA values) would lower the sentence AoA when computing the average. However, this was not the case, which suggests that the stopwords are important to include in the average AoA calculation.

Table 3: Comparison of curricula.

Curriculum type	Spearman's correlation with AO-CHILDES
Sentence AoA based on word with maximum AoA	0.15
Sentence AoA based on average AoA of all words in sentence	**0.54**
Sentence AoA based on average AoA of words in sentence (stopwords removed)	0.36

Discussion

Overall, using psycholinguistic parameters like AoA in language model development revealed fascinating insights. Firstly, we were able to show that language model surprisals (obtained as either step count of final value) reasonably correlate with ground truth AoAs of words. Since this metric is model-based, it allows us to order any arbitrary set of words in the vocabulary in increasing AoA, exceeding the set of words in the range of 16 to 30 months provided by Wordbank.

Secondly, this ordering allows us to build an age-ordered curriculum. The curriculum generated using weighted sentence AoA (including the stopwords) correlates very well in ranking when compared to a ground truth age-ordered curriculum like AO-CHILDES. This shows we can build a well-structured age-ordered curriculum from model surprisals, which are obtained as purely distributional statistics over data, without any external sources of age-ordering information.

Furthermore, we are currently evaluating the different curricula built by pre-training language models such as BabyBERTa (Huebner et al., 2021) and MiniLM (Wang et al., 2020) on them and testing their respective performances. The curriculum pre-trained models are evaluated against the non-curriculum pre-trained models on benchmark tasks such as BLiMP (Warstadt et al., 2020), to analyze if LMs benefit from child-like learning. We are also analyzing how learning develops by training models exposed to different stages of the curriculum (mapping to ages in children) and analyzing their performance on these tasks. With this, we aim to study how research areas in cognitive science such as language acquisition can be ported over to improve the efficiency of training language models.

Conclusion

Our research is a study in developing more effective ways to train language models. We use the Age of Acquisition metric as a criterion to build a curriculum for training language models, shedding light on the cognitive science aspects of language acquisition. This approach not only showcases the potential for incorporating psycholinguistic factors in language model development, but also contributes to the ongoing exploration of efficient training methodologies. We showed that AoA can be represented through model surprisal, and using that in turn to build a curriculum strongly matches a real age-ordered curriculum, with no a priori information used for age ordering. Our findings highlight the importance of comprehensive AoA analysis of sentences for effective curriculum design, offering insights into the learning progression of children acquiring language. This work opens avenues for further research in refining curriculum learning strategies, understanding their effectiveness in training large language models, and extending the applicability of our approach to diverse linguistic contexts.

References

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. *Proceedings of the 26th Annual International Conference on Machine Learning*, 41--48. https://doi.org/10.1145/1553374.1553380

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., & van der Wal, O. (2023). Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. *Proceedings of the 40th International Conference on Machine Learning, 202*. https://arxiv.org/pdf/2304.01373.pdf

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and Variability in Children's Word Learning Across Languages. *Open Mind: Discoveries in Cognitive Science, 3*, 52–67. https://doi.org/10.1162/opmi_a_00026

Chang, T. A., & Bergen, B. K. (2021). Word Acquisition in Neural Language Models. *Transactions of the Association for Computational Linguistics, 10*, 1--16. https://doi.org/10.1162/tacl_a_00444

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2016). Wordbank: an open repository for developmental vocabulary data. *Journal of Child Language, 44*(3), 677–694. https://doi.org/10.1017/s0305000916000209

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of child language, 35*(3), 515–531. https://doi.org/10.1017/S0305000907008641

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. *ACLWeb*, 624–646. https://doi.org/10.18653/v1/2021.conll-1.49

Huebner, P. A., & Willits, J. A. (2021). Using lexical context to discover the noun category: Younger children have it easier. *Psychology of Learning and Motivation, 75*, 279–331. https://www.sciencedirect.com/science/article/pii/S0079742121000256

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, & Connor Leahy. (2020). The Pile: An 800GB Dataset of Diverse Text for Language Modeling. https://arxiv.org/abs/2101.00027

Łuniewska, M., Haman, E., Armon-Lotem, S., Etenkowski, B., Southwood, F., Anđelković, D., Blom, E., Boerma, T., Chiat, S., de Abreu, P. E., Gagarina, N., Gavarró, A., Håkansson, G., Hickey, T., de López, K. J., Marinis, T., Popović, M., Thordardottir, E., Blažienė, A., & Sánchez, M. C. (2015). Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order of words? *Behavior Research Methods, 48*(3), 1154–1177. https://doi.org/10.3758/s13428-015-0636-6

Opper, M., Morrison, J., & Siddharth, N. (2023). On the effect of curriculum learning with developmental data for grammar acquisition. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning* (pp. 318–327). Association for Computational Linguistics. https://aclanthology.org/2023.conll-babylm.31/

Pentina, A., Sharmanska, V., & Lampert, C. H. (2015). Curriculum Learning of Multiple Tasks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 5492–5500. https://doi.org/10.1109/CVPR.2015.7299188

Portelance, E., Duan, Y., Frank, M. C., & Lupyan, G. (2023). Predicting Age of Acquisition for Children's Early Vocabulary in Five Languages Using Language Model Surprisal. *Cognitive Science, 47*(9). https://doi.org/10.1111/cogs.13334

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *Advances in Neural Information Processing Systems, 33*, 5776–5788. https://proceedings.neurips.cc/paper_files/paper/2020/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8, 377–392. https://doi.org/10.1162/tacl_a_00321

Xu, B., Zhang, L., Mao, Z., Wang, Q., & Zhang, Y. (2021). Curriculum Learning for Natural Language Understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 3307–3320. https://doi.org/10.1109/TASLP.2021.3121986

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, pages 19–27*. https://doi.org/10.1109/ICCV.2015.11

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *Proceedings of the Fifth International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1609.07843