Word Acquisition in Neural Language Models

Tyler A. Chang[1,2],   **Benjamin K. Bergen**[1]

[1]Department of Cognitive Science
[2]Halıcıoğlu Data Science Institute
University of California, San Diego
{tachang, bkbergen}@ucsd.edu

Abstract

We investigate how neural language models acquire individual words during training, extracting learning curves and ages of acquisition for over 600 words on the MacArthur-Bates Communicative Development Inventory (Fenson et al., 2007). Drawing on studies of word acquisition in children, we evaluate multiple predictors for words' ages of acquisition in LSTMs, BERT, and GPT-2. We find that the effects of concreteness, word length, and lexical class are pointedly different in children and language models, reinforcing the importance of interaction and sensorimotor experience in child language acquisition. Language models rely far more on word frequency than children, but like children, they exhibit slower learning of words in longer utterances. Interestingly, models follow consistent patterns during training for both unidirectional and bidirectional models, and for both LSTM and Transformer architectures. Models predict based on unigram token frequencies early in training, before transitioning loosely to bigram probabilities, eventually converging on more nuanced predictions. These results shed light on the role of distributional learning mechanisms in children, while also providing insights for more human-like language acquisition in language models.

1 Introduction

Language modeling, predicting words from context, has grown increasingly popular as a pretraining task in NLP in recent years; neural language models such as BERT (Devlin et al., 2019), ELMo (Peters et al., 2018), and GPT (Brown et al., 2020) have produced state-of-the-art performance on a wide range of NLP tasks. There is now a substantial amount of work assessing the linguistic information encoded by language models (Rogers
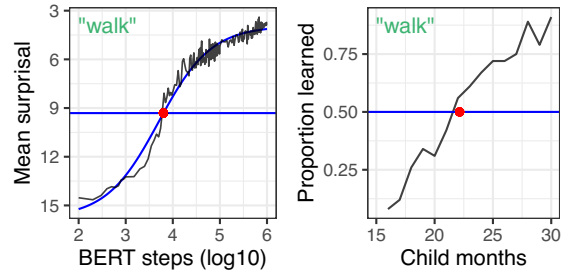

Figure 1: Learning curves for the word "walk" in a BERT language model and human children. Blue horizontal lines indicate age of acquisition cutoffs. The blue curve represents the fitted sigmoid function based on the language model surprisals during training (black). Child data obtained from Frank et al. (2017).

et al., 2020); in particular, behavioral approaches from psycholinguistics and cognitive science have been applied to study language model predictions (Futrell et al., 2019; Ettinger, 2020). From a cognitive perspective, language models are of theoretical interest as distributional models of language, agents that learn exclusively from statistics over language (Boleda, 2020; Lenci, 2018).

However, previous psycholinguistic studies of language models have nearly always focused on fully-trained models, precluding comparisons to the wealth of literature on human language acquisition. There are limited exceptions. Rumelhart and McClelland (1986) famously studied past tense verb form learning in phoneme-level neural networks during training, a study which was replicated in more modern character-level recurrent neural networks (Kirov and Cotterell, 2018). However, these studies focused only on sub-word features. There remains a lack of research on language acquisition in contemporary language models, which encode higher level features such as syntax and semantics.

As an initial step towards bridging the gap between language acquisition and language model-

ing, we present an empirical study of word acquisition during training in contemporary language models, including LSTMs, GPT-2, and BERT. We consider how variables such as word frequency, concreteness, and lexical class contribute to words' ages of acquisition in language models. Each of our selected variables has effects on words' ages of acquisition in children; our language model results allow us to identify the extent to which each effect in children can or cannot be attributed in principle to distributional learning mechanisms.

Finally, to better understand how computational models acquire language, we identify consistent patterns in language model training across architectures. Our results suggest that language models may acquire traditional distributional statistics such as unigram and bigram probabilities in a systematic way. Understanding how language models acquire language can lead to better architectures and task designs for future models, while also providing insights into distributional learning mechanisms in people.

2 Related work

Our work draws on methodologies from word acquisition studies in children and psycholinguistic evaluations of language models. In this section, we briefly outline both lines of research.

2.1 Child word acquisition

Child development researchers have previously studied word acquisition in children, identifying variables that help predict words' ages of acquisition in children. In Wordbank, Frank et al. (2017) compiled reports from parents reporting when their child produced each word on the MacArthur-Bates Communicative Development Inventory (CDI; Fenson et al., 2007). For each word w, Braginsky et al. (2016) fitted a logistic curve predicting the proportion of children that produce w at different ages; they defined a word's age of acquisition as the age at which 50% of children produce w. Variables such as word frequency, word length, lexical class, and concreteness were found to influence words' ages of acquisition in children across languages. Recently, it was shown that fully-trained LSTM language model surprisals are also predictive of words' ages of acquisition in children (Portelance et al., 2020). However, no studies have evaluated ages of acqui-

sition in language models themselves.

2.2 Evaluating language models

Recently, there has been substantial research evaluating language models using psycholinguistic approaches, reflecting a broader goal of interpreting language models (BERTology; Rogers et al., 2020). For instance, Ettinger (2020) used the output token probabilities from BERT in carefully constructed sentences, finding that BERT learns commonsense and semantic relations to some degree, although it struggles with negation. Gulordava et al. (2018) found that LSTM language models recognize long distance syntactic dependencies; however, they still struggle with more complicated constructions (Marvin and Linzen, 2018).

These psycholinguistic methodologies do not rely on specific language model architectures or fine-tuning on a probe task. Notably, because these approaches rely only on output token probabilities from a given language model, they are well-suited to evaluations early in training, when fine-tuning on downstream tasks is unfruitful. That said, previous language model evaluation studies have focused on fully-trained models, progressing largely independently from human language acquisition literature. Our work seeks to bridge this gap.

3 Method

We trained unidirectional and bidirectional language models with LSTM and Transformer architectures. We quantified each language model's age of acquisition for each word in the CDI (Fenson et al., 2007). Similar to word acquisition studies in children, we identified predictors for words' ages of acquisition in language models.[1]

3.1 Language models

Datasets and training Language models were trained on a combined corpus containing the BookCorpus (Zhu et al., 2015) and WikiText-103 datasets (Merity et al., 2017). Following Devlin et al. (2019), each input sequence was a sentence pair; the training dataset consisted of 25.6M sentence pairs. The remaining sentences (5.8M pairs) were used for evaluation and to generate word

---

[1]Code and data are available at https://github.com/tylerachang/word-acquisition-language-models.

	# Parameters	Perplexity
LSTM	37M	54.8
GPT-2	108M	30.2
BiLSTM	51M	9.0
BERT	109M	7.2

Table 1: Parameter counts and evaluation perplexities for the trained language models. For reference, the pre-trained BERT base model from Huggingface reached a perplexity of 9.4 on our evaluation set. Additional perplexity comparisons with comparable models are included in Appendix A.1.

learning curves. Sentences were tokenized using the unigram language model tokenizer implemented in SentencePiece (Kudo and Richardson, 2018). Models were trained for 1M steps, with batch size 128 and learning rate 0.0001. As a metric for overall language model performance, we report evaluation perplexity scores in Table 1. We include evaluation loss curves, full training details, and hyperparameters in Appendix A.1.

Transformers The two Transformer models followed the designs of GPT-2 (Radford et al., 2019) and BERT (Devlin et al., 2019), allowing us to evaluate both a unidirectional and bidirectional Transformer language model. GPT-2 was trained with the causal language modeling objective, where each token representation is used to predict the next token; the masked self-attention mechanism allows tokens to attend only to previous tokens in the input sequence. In contrast, BERT used the masked language modeling objective, where masked tokens are predicted from surrounding tokens in both directions.

Our BERT model used the base size model from Devlin et al. (2019). Our GPT-2 model used the similar-sized model from Radford et al. (2019), equal in size to the original GPT model. Parameter counts are listed in Table 1. Transformer models were trained using the Huggingface Transformers library (Wolf et al., 2020).

LSTMs We also trained both a unidirectional and bidirectional LSTM language model, each with three stacked LSTM layers. Similar to GPT-2, the unidirectional LSTM predicted the token at time t from the hidden state at time $t-1$. The bidirectional LSTM (BiLSTM) predicted the token at time t from the sum of the hidden states at times $t-1$ and $t+1$ (Aina et al., 2019).

3.2 Learning curves and ages of acquisition

We sought to quantify each language model's ability to predict individual words over the course of training. We considered all words in the CDI that were considered one token by the language models (611 out of 651 words).

For each such token w, we identified up to 512 occurrences of w in the held-out portion of the language modeling dataset.[2] To evaluate a language model at training step s, we fed each sentence pair into the model, attempting to predict the masked token w. We computed the surprisal: $-\log_2(P(w))$ averaged over all occurrences of w to quantify the quality of the models' predictions for word w at step s (Levy, 2008; Goodkind and Bicknell, 2018).

We computed this average surprisal for each target word at approximately 200 different steps during language model training, sampling more heavily from earlier training steps, prior to model convergence. The selected steps are listed in Appendix A.1. By plotting surprisals over the course of training, we obtained a learning curve for each word, generally moving from high surprisal to low surprisal. The surprisal axis in our plots is reversed to reflect increased understanding over the course of training, consistent with plots showing increased proportions of children producing a given word over time (Frank et al., 2017).

For each learning curve (4 language model architectures × 611 words), we fitted a sigmoid function to model the smoothed acquisition of word w. Sample learning curves are shown in Figure 1 and Figure 2.

Age of acquisition To extract age of acquisition from a learning curve, we established a cutoff surprisal where we considered a given word "learned." In child word acquisition studies, an analogous cutoff is established when 50% of children produce a word (Braginsky et al., 2016).

Following this precedent, we determined our cutoff to be 50% between a baseline surprisal (predicting words based on random chance) and the minimum surprisal attained by the model for word w. We selected the random chance baseline to

---

[2] We only selected sentence pairs with at least eight tokens of context, unidirectionally or bidirectionally depending on model architecture. Thus, the unidirectional and bidirectional samples differed slightly. Most tokens (92.3%) had the maximum of 512 samples both unidirectionally and bidirectionally, and all tokens had at least 100 samples in both cases.
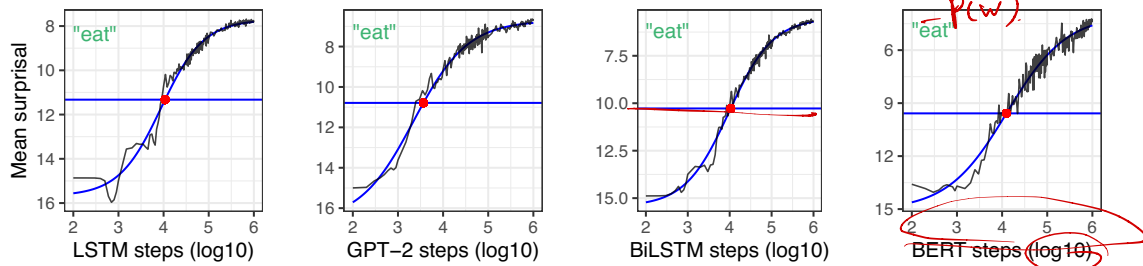
Figure 2: Learning curves for the word "eat" for all four language model architectures. Blue horizontal lines indicate age of acquisition cutoffs, and blue curves represent fitted sigmoid functions.

best reflect a language model's ability to predict a word with no access to any training data, similar to an infant's language-specific knowledge prior to any linguistic exposure. We selected minimum surprisal as our other bound to reflect how well a particular word can eventually be learned by a particular language model, analogous to an adult's understanding of a given word.

For each learning curve, we found the intersection between the fitted sigmoid and the cutoff surprisal value. We defined age of acquisition for a language model as the corresponding training step, on a log10 scale. Sample cutoffs and ages of acquisition are shown in Figure 1 and Figure 2.

3.3 Predictors for age of acquisition

As potential predictors for words' ages of acquisition in language models, we selected variables that are predictive of age of acquisition in children (Braginsky et al., 2016). When predicting ages of acquisition in language models, we computed word frequencies and utterance lengths over the language model training corpus. Our five selected predictors were:

- Log-frequency: the natural log of the word's per-1000 token frequency.

- MLU: we computed the mean length of utterance as the mean length of sequences containing a given word.[3] MLU has been used as a metric for the complexity of syntactic contexts in which a word appears (Roy et al., 2015).

- n-chars: as in Braginsky et al. (2016), we used the number of characters in a word as a coarse proxy for the length of a word.

- Concreteness: we used human-generated concreteness norms from Brysbaert et al. (2014), rated on a five-point scale. We imputed missing values (3% of words) using the mean concreteness score.

- Lexical class: we used the lexical classes annotated in Wordbank. Possible lexical classes were Noun, Verb, Adjective, Function Word, and Other.

We ran linear regressions with linear terms for each predictor. To determine statistical significance for each predictor, we ran likelihood ratio tests, comparing the overall regression (including the target predictor) with a regression including all predictors except the target. To determine the direction of effect for each continuous predictor, we used the sign of the coefficient in the overall regression.

As a potential concern for interpreting regression coefficient signs, we assessed collinearities between predictors by computing the variance inflation factor (VIF) for each predictor. No VIF exceeded 5.0,[4] although we did observe moderate correlations between log-frequency and n-chars ($r = -0.49$), and between log-frequency and concreteness ($r = -0.64$). These correlations are consistent with those identified for child-directed speech in Braginsky et al. (2016). To ease collinearity concerns, we considered single-predictor regressions for each predictor, using adjusted predictor values after accounting for log-frequency (residuals after regressing the predictor over log-frequency). In all cases, the coefficient sign in the adjusted single predictor regression was consistent with the sign of the coefficient in the overall regression.

When lexical class (the sole categorical predictor) reached significance based on the likelihood

---

[3]We also considered a unidirectional MLU metric (counting only previous tokens) for the unidirectional models, finding that it produced similar results.

[4]Common VIF cutoff values are 5.0 and 10.0.

	LSTM	GPT-2	BiLSTM	BERT	Children
Log-frequency	*** (−)	*** (−)	*** (−)	*** (−)	*** (−)
MLU		** (+)	*** (+)	*** (+)	*** (+)
n-chars	*** (−)	*** (−)	*** (−)	*** (−)	** (+)
Concreteness					*** (−)
Lexical class	***	***			***
R^2	0.93	0.92	0.95	0.94	0.43

Table 2: Significant predictors for a word's age of acquisition are marked by asterisks ($p < 0.05$*; $p < 0.01$**; $p < 0.001$***). Signs of coefficients are notated in parentheses. The R^2 denotes the adjusted R^2 in a regression using all five predictors.

ratio test, we ran a one-way analysis of covariance (ANCOVA) with log-frequency as a covariate. The ANCOVA ran a standard ANOVA on the age of acquisition residuals after regressing over log-frequency. We used Tukey's honestly significant difference (HSD) test to assess pairwise differences between lexical classes.

3.4 Age of acquisition in children

For comparison, we used the same variables to predict words' ages of acquisition in children, as in Braginsky et al. (2016). We obtained smoothed ages of acquisition for children from the Wordbank dataset (Frank et al., 2017). When predicting ages of acquisition in children, we computed word frequencies and utterance lengths over the North American English CHILDES corpus of child-directed speech (MacWhinney, 2000).

Notably, CHILDES contained much shorter sentences on average than the language model training corpus (mean sentence length 4.50 tokens compared to 15.14 tokens). CDI word log-frequencies were only moderately correlated between the two corpora ($r = 0.78$). This aligns with previous work finding that child-directed speech contains on average fewer words per utterance, smaller vocabularies, and simpler syntactic structures than adult-directed speech (Soderstrom, 2007). These differences were likely compounded by differences between spoken language in the CHILDES corpus and written language in the language model corpus. We computed word frequencies and MLUs separately over the two corpora to ensure that our predictors accurately reflected the learning environments of children and the language models.

We also note that the language model training corpus was much larger overall than the CHILDES corpus. CHILDES contained 7.5M tokens, while the language model corpus contained 852.1M to-
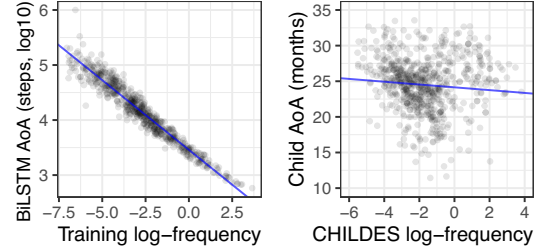

Figure 3: Effects of log-frequency on words' ages of acquisition (AoA) in the BiLSTM and children. The BiLSTM was the language model architecture with the largest effect of log-frequency (adjusted $R^2 = 0.94$).

kens. Children are estimated to hear approximately 13K words per day (Gilkerson et al., 2017), for a total of roughly 19.0M words during their first four years of life. Because contemporary language models require much more data than children hear, the models do not necessarily reflect how children would learn if restricted solely to linguistic input. Instead, the models serve as examples of relatively successful distributional learners, establishing how one might expect word acquisition to progress according to effective distributional mechanisms.

4 Results

Significant predictors of age of acquisition are shown in Table 2, comparing children and each of the four language model architectures.

Log-frequency In children and all four language models, more frequent words were learned earlier (a negative effect on age of acquisition). As shown in Figure 3, this effect was much more pronounced in language models (adjusted $R^2 = 0.91$ to 0.94) than in children (adjusted $R^2 = 0.01$).[5]

---

[5]Because function words are frequent but acquired later by children, a quadratic model of log-frequency on age of ac-

The sizeable difference in log-frequency predictivity emphasizes the fact that language models learn exclusively from distributional statistics over words, while children have access to additional social and sensorimotor cues.

MLU Except in unidirectional LSTMs, MLU had a positive effect on a word's age of acquisition in language models. Interestingly, we might have expected the opposite effect (particularly in Transformers) if additional context (longer utterances) facilitated word learning. Instead, our results are consistent with effects of MLU in children; words in longer utterances are learned later, even after accounting for other variables. The lack of effect in unidirectional LSTMs could simply be due to LSTMs being the least sensitive to contextual information of the models under consideration. The positive effect of MLU in other models suggests that complex syntactic contexts may be more difficult to learn through distributional learning alone, which might partly explain why children learn words in longer utterances more slowly.

n-chars There was a negative effect of n-chars on age of acquisition in all four language models; longer words were learned earlier. This contrasts with children, who acquire shorter words earlier. This result is particularly interesting because the language models we used have no information about word length. We hypothesize that the effect of n-chars in language models may be driven by polysemy, which is not accounted for in our regressions. Shorter words tend to be more polysemous (a greater diversity of meanings; Casas et al., 2019), which could lead to slower learning in language models. In children, this effect may be overpowered by the fact that shorter words are easier to parse and produce.

Concreteness Although children overall learn more concrete words earlier, the language models showed no significant effects of concreteness on age of acquisition. This entails that the effects in children cannot be explained by correlations between concrete words and easier distributional learning contexts. Again, this highlights the importance of sensorimotor experience and conceptual development in explaining the course of child language acquisition.

Lexical class The bidirectional language models showed no significant effects of lexical class on age of acquisition. In other words, the differences between lexical classes were sufficiently accounted for by the other predictors for BERT and the BiLSTM. However, in the unidirectional language models (GPT-2 and the LSTM), nouns and function words were acquired later than adjectives and verbs.[6] This contrasts with children learning English, who on average acquired nouns earlier than adjectives and verbs, acquiring function words last.[7]

Thus, children's early acquisition of nouns cannot be explained by distributional properties of English nouns, which are acquired later by unidirectional language models. This result is compatible with the hypothesis that nouns are acquired earlier because they often map to real world objects; function words might be acquired later because their meanings are less grounded in sensorimotor experience. It has also been argued that children might have an innate bias to learn objects earlier than relations and traits (Markman, 1994). Lastly, it is possible that the increased salience of sentence-final positions (which are more likely to contain nouns in English and related languages) facilitates early acquisition of nouns in children (Caselli et al., 1995). Consistent with these hypotheses, our results suggest that English verbs and adjectives may be easier to learn from a purely distributional perspective, but children acquire nouns earlier based on sensorimotor, social, or cognitive factors.

4.1 First and last learned words

As a qualitative analysis, we compared the first and last words acquired by the language models and children, as shown in Table 3. In line with our previous results, the first and last words learned by

---

quisition in children provided a slightly better fit ($R^2 = 0.03$) if not accounting for lexical class. A quadratic model of log-frequency also provided a slightly better fit for unidirectional language models ($R^2 = 0.93$ to 0.94), particularly for high-frequency words; in language models, this could be due either to a floor effect on age of acquisition for high-frequency words or to slower learning of function words. Regardless, significant effects of other predictors remained the same when using a quadratic model for log-frequency.

[6]Significant pairwise comparisons between lexical classes are listed in Appendix A.2.

[7]There is ongoing debate around the existence of a universal "noun bias" in early word acquisition. For instance, Korean and Mandarin-speaking children have been found to acquire verbs earlier than nouns, although this effect appears sensitive to context and the measure of vocabulary acquisition (Choi and Gopnik, 1995; Tardif et al., 1999).

	Language models	Children
First	*a, and, for, he, her, his, I, it, my, of, on, she, that, the, to, was, with, you*	*baby, ball, bye, daddy, dog, hi, mommy, moo, no, shoe, uh, woof, yum*
Last	*bee, bib, choo, cracker, crayon, giraffe, glue, kitty, moose, pancake, popsicle, quack, rooster, slipper, tuna, yum, zebra*	*above, basement, beside, country, downtown, each, hate, if, poor, walker, which, would, yourself*

Table 3: First and last words acquired by the language models and children. For language models, we identified words that were in the top or bottom 5% of ages of acquisition for all models. For children, we identified words in the top or bottom 2% of ages of acquisition.

the language models were largely determined by word frequencies. The first words acquired by the models were all in the top 3% of frequent words, and the last acquired words were all in the bottom 15%. Driven by this effect, many of the first words learned by the language models were function words or pronouns. In contrast, many of the first words produced by children were single-word expressions, such as greetings, exclamations, and sounds. Children acquired several highly frequent words late, such as "if," which is in the 90th frequency percentile of the CHILDES corpus. Of course, direct comparisons between the first and last words acquired by the children and language models are confounded by differing datasets and learning environments, as detailed in Section 3.4.

4.2 Age of acquisition vs. minimum surprisal

Next, we assessed whether a word's age of acquisition in a language model could be predicted from how well that word was learned in the fully-trained model. To do this, we considered the minimum surprisal attained by each language model for each word. We found a significant effect of minimum surprisal on age of acquisition in all four language models, even after accounting for all five other predictors (using likelihood ratio tests; $p < 0.001$). In part, this is likely because the acquisition cutoff for each word's fitted sigmoid was dependent on the word's minimum surprisal.

It could then be tempting to treat minimum surprisal as a substitute for age of acquisition in language models; this approach would require only

publicly-available fully-trained language models. Indeed, the correlation between minimum surprisal and age of acquisition was substantial (Pearson's $r = 0.88$ to 0.92). However, this correlation was driven largely by effects of log-frequency, which had a large negative effect on both metrics. When adjusting minimum surprisal and age of acquisition for log-frequency (using residuals after linear regressions), the correlation decreased dramatically (Pearson's $r = 0.22$ to 0.46). While minimum surprisal accounts for a significant amount of variance in words' ages of acquisition, the two metrics are not interchangeable.

4.3 Alternative age of acquisition definitions

Finally, we considered alternative operationalizations of words' ages of acquisition in language models. For instance, instead of defining an acquisition cutoff at 50% between random chance and the minimum surprisal for each word, we could consider the midpoint of each fitted sigmoid curve. This method would be equivalent to defining upper and lower surprisal baselines at the upper and lower asymptotes of the fitted sigmoid, relying on the assumption that these asymptotes roughly approximate surprisal values before and after training. However, this assumption fails in cases where the majority of a word's learning curve is modeled by only a sub-portion of the fitted sigmoid. For example, for the word "for" in Figure 4, the high surprisal asymptote is at 156753.5, compared to a random chance surprisal of 14.9 and a minimum surprisal of 4.4. Using the midpoint age of acquisition in this case would result in an age of acquisition of -9.6 steps (log10).

We also considered alternative cutoff proportions (replacing 50%) in our original age of acquisition definition. We considered cutoffs at each possible increment of 10%. The signs of nearly all significant coefficients in the overall regressions (see Table 2) remained the same for all language models regardless of cutoff proportion.[8]

5 Language model learning curves

The previous sections identified factors that predict words' ages of acquisition in language models. We now proceed with a qualitative analysis of the learning curves themselves. We found

---

[8]The only exception was a non-significant positive coefficient for n-chars in BERT with a 90% acquisition cutoff.
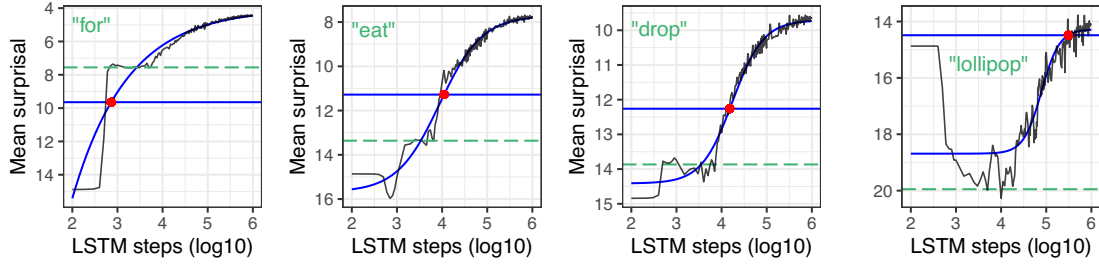
Figure 4: LSTM learning curves for the words "for," "eat," "drop," and "lollipop." Blue horizontal lines indicate age of acquisition cutoffs, and blue curves represent fitted sigmoid functions. Green dashed lines indicate the surprisal if predicting solely based on unigram probabilities (raw token frequencies). Early in training, language model surprisals tended to shift towards the unigram frequency-based surprisals.

that language models learn traditional distributional statistics in a systematic way.

5.1 Unigram probabilities

First, we observed a common pattern in word learning curves across model architectures. As expected, each curve began at the surprisal value corresponding to random chance predictions. Then, as shown in Figure 4, many curves shifted towards the surprisal value corresponding to raw unigram probabilities (i.e. based on raw token frequencies). This pattern was particularly pronounced in LSTM-based language models, although it appeared in all architectures. Interestingly, the shift occurred even if the unigram surprisal was higher (or "worse") than random-chance surprisal, as demonstrated by the word "lollipop" in Figure 4. Thus, we posited that language models pass through an early stage of training where they approximate unigram probabilities.

To test this hypothesis, we aggregated each model's predictions for randomly masked tokens in the evaluation dataset (16K sequences), including tokens not on the CDI. For each saved training step, we computed the average Kullback-Leibler (KL) divergence between the model predictions and the unigram frequency distribution. For comparison, we also computed the KL divergence with a uniform distribution (random chance) and with the one-hot true token distribution. We note that the KL divergence with the one-hot true token distribution is equivalent to the cross-entropy loss function using log base two.[9]

As shown in Figure 5, we plotted the KL diver-

gences between each reference distribution and the model predictions over the course of training. As expected, all four language models converged towards the true token distribution (minimizing the loss function) throughout training, diverging from the uniform distribution. Divergence from the uniform distribution could also reflect that the models became more confident in their predictions during training, leading to lower entropy predictions.

As hypothesized, we also found that all four language models exhibited an early stage of training in which their predictions approached the unigram distribution, before diverging to reflect other information. This suggests that the models overfitted to raw token frequencies early in training, an effect which was particularly pronounced in the LSTM-based models. Importantly, because the models eventually diverged from the unigram distribution, the initial unigram phase cannot be explained solely by mutual information between the true token distribution and unigram frequencies.

5.2 Bigram probabilities

We then ran a similar analysis using bigram probabilities, where each token probability was dependent only on the previous token. A bigram distribution P_b was computed for each masked token in the evaluation dataset, based on bigram counts in the training corpus. As dictated by the bigram model definition, we defined $P_b(w_i) = P(w_i|w_{i-1})$ for unidirectional models, and $P_b(w_i) = P_b(w_i|w_{i-1}, w_{i+1}) \propto P(w_i|w_{i-1})P(w_{i+1}|w_i)$ for bidirectional models. We computed the average KL divergence between the bigram probability distributions and the language model predictions.

As shown in Figure 5, during the unigram learning phase, the bigram KL divergence de-

---

[9]All KL divergences were computed using log base two. KL divergences were computed as $\mathrm{KL}(\mathbf{y}_{ref}, \hat{\mathbf{y}})$, where $\hat{\mathbf{y}}$ was the model's predicted probability distribution and $\mathbf{y}_{ref}$ was the reference distribution.
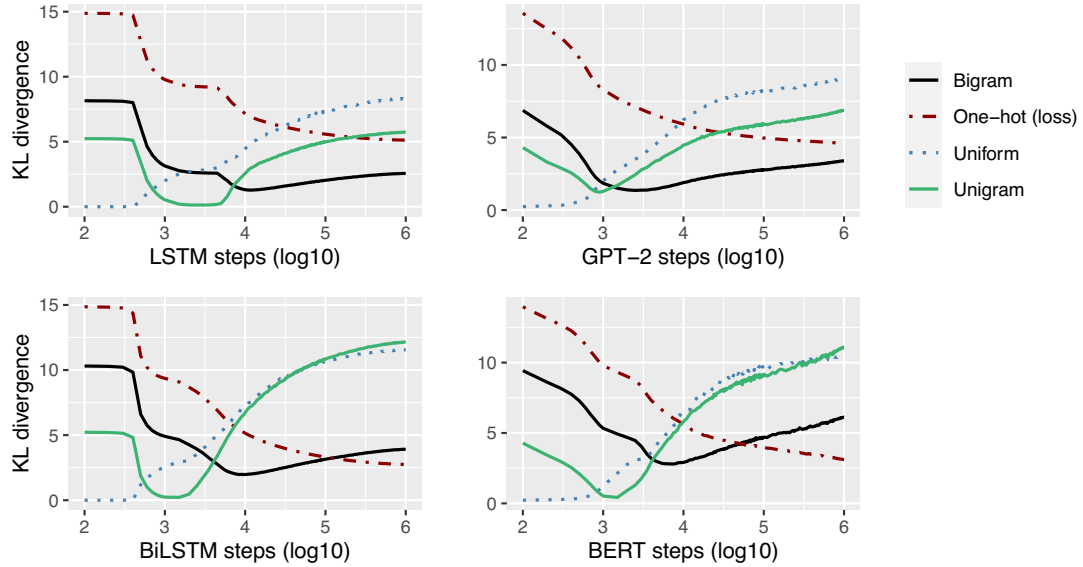
Figure 5: KL divergences between reference distributions and model predictions over the course of training. The KL divergence with the one-hot true token distribution is equivalent to the base two cross-entropy loss. Early in training, the models temporarily overfitted to unigram then bigram probabilities.

creased for all language models. This is likely caused by mutual information between the unigram and bigram distributions; as the models approached the unigram distribution, their divergences with the bigram distributions roughly approximated the average KL divergence between the bigram and unigram distributions themselves (average KL = 3.86 between unidirectional bigrams and unigrams; average KL = 5.88 between bidirectional bigrams and unigrams). In other words, the models' initial decreases in bigram KL divergences can be explained predominantly by unigram frequency learning.

However, when the models began to diverge from the unigram distribution, they continued to approach the bigram distributions. Each model then hit a local minimum in average bigram KL divergence before diverging from the bigram distributions. This suggests that the models overfitted to bigram probabilities after the unigram learning phase. Thus, it appears that early in training, language models make predictions based on unigram frequencies, then bigram probabilities, eventually learning to make more nuanced predictions.

Of course, this result may not be surprising for LSTM-based language models. Because tokens are fed into LSTMs sequentially, it is intuitive that they would make use of bigram probabilities. Our results confirm this intuition, and they further show that Transformer language models fol-

low a similar pattern. Because BERT and GPT-2 only encode token position information through learned absolute position embeddings before the first self-attention layer, they have no architectural reason to overfit to bigram probabilities based on adjacent tokens.[10] Instead, unigram and bigram learning may be a natural consequence of the language modeling task, or even distributional learning more generally.

6 Discussion

We found that language models are highly sensitive to basic statistics such as frequency and bigram probabilities during training. Their acquisition of words is also sensitive to features such as sentence length and (for unidirectional models) lexical class. Importantly, the language models exhibited notable differences with children in the effects of lexical class, word lengths, and concreteness, highlighting the importance of social, cognitive, and sensorimotor experience in child language development.

6.1 Distributional learning, language modeling, and NLU

In this section, we address the broader relationship between distributional language acquisition and contemporary language models.

---

[10] Absolute position embeddings in the Transformers were randomly initialized at the beginning of training.

Distributional learning in people There is ongoing work assessing distributional mechanisms in human language learning (Aslin and Newport, 2014). For instance, adults can learn syntactic categories using distributional information alone (Reeder et al., 2017). Adults also show effects of distributional probabilities in reading times (Goodkind and Bicknell, 2018) and neural responses (Frank et al., 2015). In early language acquisition, there is evidence that children are sensitive to transition (bigram) probabilities between phonemes and between words (Romberg and Saffran, 2010), but it remains an open question to what extent distributional mechanisms can explain effects of other factors (e.g. utterance lengths and lexical classes) known to influence naturalistic language learning.

To shed light on this question, we considered neural language models as distributional language learners. If analogous distributional learning mechanisms were involved in children and language models, then we would observe similar word acquisition patterns in children and the models. Our results demonstrate that a purely distributional learner would be far more reliant on frequency than children are. Furthermore, while the effects of utterance length on words' ages of acquisition in children can potentially be explained by distributional mechanisms, the effects of word length, concreteness, and lexical class cannot.

Distributional models Studying language acquisition in distributional models also has implications for core NLP research. Pre-trained language models trained only on text data have become central to state-of-the-art NLP systems. Language models even outperform humans on some tasks (He et al., 2021), making it difficult to pinpoint why they perform poorly in other areas. In this work, we isolated ways that language models differ from children in how they acquire words, emphasizing the importance of sensorimotor experience and cognitive development for human-like language acquisition. Future work could investigate the acquisition of syntactic structures or semantic information in language models.

Non-distributional learning We showed that distributional language models acquire words in very different ways from children. Notably, children's linguistic experience is grounded in sensorimotor and cognitive experience. Children

as young as ten months old learn word-object pairings, mapping novel words onto perceptually salient objects (Pruden et al., 2006). By the age of two, they are able to integrate social cues such as eye gaze, pointing, and joint attention (Çetinçelik et al., 2021). Neural network models of one-word child utterances exhibit vocabulary acquisition trajectories similar to children when only using features from conceptual categories and relations (Nyamapfene and Ahmad, 2007). Our work shows that these grounded and interactive features impact child word acquisition in ways that cannot be explained solely by intra-linguistic signals.

That said, there is a growing body of work grounding language models using multimodal information and world knowledge. Language models trained on visual and linguistic inputs have achieved state-of-the-art performance on visual question answering tasks (Antol et al., 2015; Lu et al., 2019; Zellers et al., 2021b), and models equipped with physical dynamics modules are more accurate than standard language models at modeling world dynamics (Zellers et al., 2021a). There has also been work building models directly for non-distributional tasks; reinforcement learning can be used for navigation and multi-agent communication tasks involving language (Chevalier-Boisvert et al., 2019; Lazaridou et al., 2017; Zhu et al., 2020). These models highlight the grounded, interactive, and communicative nature of language. Indeed, these non-distributional properties may be essential to more human-like natural language understanding (Bender and Koller, 2020; Emerson, 2020). Based on our results for word acquisition in language models, it is possible that these multimodal and non-distributional models could also exhibit more human-like language acquisition.

7 Conclusion

In this work, we identified factors that predict words' ages of acquisition in contemporary language models. We found contrasting effects of lexical class, word length, and concreteness in children and language models, and we observed much larger effects of frequency in the models than in children. Furthermore, we identified ways that language models aquire unigram and bigram statistics early in training. This work paves the way for future research integrating language acquisition and natural language understanding.

Acknowledgements

References

Laura Aina, Kristina Gulordava, and Gemma Boleda. 2019. Putting words in context: LSTM language models and lexical ambiguity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3348, Florence, Italy. Association for Computational Linguistics.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *International Conference on Computer Vision*.

Richard Aslin and Elissa Newport. 2014. Distributional language learning: Mechanisms and models of category formation. *Language Learning*, 64:86–105.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Conference on Neural Information Processing Systems*.

Marc Brysbaert, Amy Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46.

Bernardino Casas, Antoni Hernández-Fernández, Neus Català, Ramon Ferrer-i-Cancho, and Jaume Baixeries. 2019. Polysemy and brevity versus frequency in language. *Computer Speech & Language*, 58:19–50.

Maria Cristina Caselli, Elizabeth Bates, Paola Casadio, Judi Fenson, Larry Fenson, Lisa Sanderl, and Judy Weir. 1995. A cross-linguistic study of early lexical development. *Cognitive Development*, 10(2):159–199.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. 2019. BabyAI: A platform to study the sample efficiency of grounded language learning. In *International Conference on Learning Representations*.

Soonja Choi and Alison Gopnik. 1995. Early acquisition of verbs in Korean: A cross-linguistic study. *Journal of Child Language*, 22(3):497–529.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Guy Emerson. 2020. What are the goals of distributional semantics? In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 7436–7453, Online. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Larry Fenson, Virginia Marchman, Donna Thal, Phillip Dale, Steven Reznick, and Elizabeth Bates. 2007. *MacArthur-Bates communicative development inventories*. Paul H. Brookes Publishing Company, Baltimore, MD.

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

Stefan Frank, Leun Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. 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 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations*.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annual Review of Linguistics*, 4(1):151–171.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Conference on Neural Information Processing Systems*.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk.* Lawrence Erlbaum Associates, Mahwah, NJ.

Ellen Markman. 1994. Constraints on word meaning in early language acquisition. *Lingua*, 92:199–227.

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, Brussels, Belgium. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *Proceedings of the Fifth International Conference on Learning Representations*.

Abel Nyamapfene and Khurshid Ahmad. 2007. A multimodal model of child language acquisition at the one-word stage. In *International Joint Conference on Neural Networks*, pages 783–788.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Eva Portelance, Judith Degen, and Michael Frank. 2020. Predicting age of acquisition in early word learning using recurrent neural networks. In *Proceedings of CogSci 2020*.

Shannon M. Pruden, Kathy Hirsh-Pasek, Roberta Michnick Golinkoff, and Elizabeth Hennon. 2006. The birth of words: Ten-month-olds learn words through perceptual salience. *Child Development*, 77(2):266–280.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Patricia Reeder, Elissa Newport, and Richard Aslin. 2017. Distributional learning of subcategories in an artificial grammar: Category generalization and subcategory restrictions. *Journal of Memory and Language*, 97:17–29.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Alexa Romberg and Jenny Saffran. 2010. Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews in Cognitive Science*, 1(6):906–914.

Brandon Roy, Michael Frank, Philip DeCamp, Matthew Miller, and Deb Roy. 2015. Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, 112(41):12663–12668.

David Rumelhart and James McClelland. 1986. On learning the past tenses of English verbs. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 2.

Melanie Soderstrom. 2007. Beyond babytalk: Reevaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27(4):501–532.

Twila Tardif, Susan Gelman, and Fan Xu. 1999. Putting the "noun bias" in context: A comparison of English and Mandarin. *Child Development*, 70(3):620–635.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Matthew Peters, Roozbeh Mottaghi, Aniruddha Kembhavi, Ali

Farhadi, and Yejin Choi. 2021a. PIGLeT: Language grounding through neuro-symbolic interaction in a 3D world. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2040–2050, Online. Association for Computational Linguistics.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021b. MERLOT: Multimodal neural script knowledge models. *arXiv preprint arXiv:2106.02636v2*.

Wang Zhu, Hexiang Hu, Jiacheng Chen, Zhiwei Deng, Vihan Jain, Eugene Ie, and Fei Sha. 2020. BabyWalk: Going farther in vision-and-language navigation by taking baby steps. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2539–2556, Online. Association for Computational Linguistics.

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.

Melis Çetinçelik, Caroline Rowland, and Tineke Snijders. 2021. Do the eyes have it? A systematic review on the role of eye gaze in infant language development. *Frontiers in Psychology*, 11.

A Appendix

A.1 Language model training details

Language model training hyperparameters are listed in Table 4. Input and output token embeddings were tied in all models. Each model was trained using four Titan Xp GPUs. The LSTM, BiLSTM, BERT, and GPT-2 models took four, five, seven, and eleven days to train respectively.

To verify language model convergence, we plotted evaluation loss curves, as in Figure 6. To ensure that our language models reached performance levels comparable to contemporary language models, in Table 6 we report perplexity comparisons between our trained models and

Hyperparameter	Value
Hidden size	768
Embedding size	768
Vocab size	30004
Max sequence length	128
Batch size	128
Train steps	1M
Learning rate decay	Linear
Warmup steps	10000
Learning rate	1e-4
Adam ϵ	1e-6
Adam β_1	0.9
Adam β_2	0.999
Dropout	0.1
Transformer hyperparameter	**Value**
Transformer layers	12
Intermediate hidden size	3072
Attention heads	12
Attention head size	64
Attention dropout	0.1
BERT mask proportion	0.15
LSTM hyperparameter	**Value**
LSTM layers	3
Context size	768

Table 4: Language model training hyperparameters.

models with the same architectures in previous work. For BERT, we evaluated the perplexity of Huggingface's pre-trained BERT base uncased model on our evaluation dataset (Wolf et al., 2020). For the remaining models, we used the evaluation perplexities reported in the original papers: Gulordava et al. (2018) for the LSTM,[11] Radford et al. (2019) for GPT-2 (using the comparably-sized model evaluated on the WikiText-103 dataset), and Aina et al. (2019) for the BiLSTM. Because these last three models were cased, we could not evaluate them directly on our uncased evaluation set. Due to differing vocabularies, hyperparameters, and datasets, our perplexity comparisons are not definitive; however, they show that our models perform similarly to contemporary language models.

Finally, we evaluated each of our models for word acquisition at 208 checkpoint steps during training, sampling more heavily from earlier steps.

---

[11]The large parameter count for the LSTM in Gulordava et al. (2018) is primarily due to its large vocabulary without a decreased embedding size.

LSTM	GPT-2	Children
Adj < Function ***	Adj < Function ***	Nouns < Adj ***
Adj < Nouns **	Adj < Other *	Nouns < Verbs ***
Adj < Other **	Verbs < Function ***	Nouns < Function ***
Verbs < Function ***	Verbs < Nouns *	Function > Adj ***
Verbs < Nouns **	Verbs < Other **	Function > Verbs ***
Verbs < Other *	Nouns < Function ***	Function > Other ***
		Other < Adj **
		Other < Verbs **

Table 5: Significant pairwise differences between lexical classes when predicting words' ages of acquisition in language models and children (adjusted $p < 0.05^*$; $p < 0.01^{**}$; $p < 0.001^{***}$). A higher value indicates that a lexical class is acquired later on average. The five possible lexical classes were Noun, Verb, Adjective (Adj), Function Word (Function), and Other.
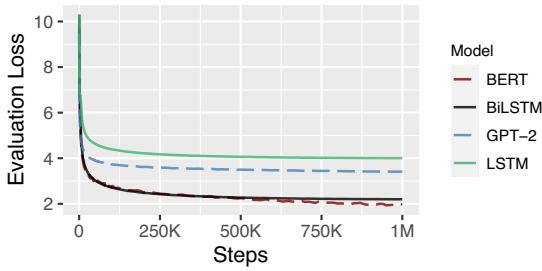

Figure 6: Evaluation loss during training for all four language models. Note that perplexity is equal to $\exp(loss)$.

	Ours		Previous work	
	# Params	Perplexity	# Params	Perplexity
LSTM	37M	54.8	72M [a]	52.1
GPT-2	108M	30.2	117M [b]	37.5
BiLSTM	51M	9.0	42M [c]	18.1
BERT	109M	7.2	110M [d]	9.4

Table 6: Rough perplexity comparisons between our trained language models and models with the same architectures in previous work ([a]Gulordava et al., 2018; [b]Radford et al., 2019; [c]Aina et al., 2019; [d]Wolf et al., 2020).

We evaluated checkpoints at the following steps:

- Every 100 steps during the first 1000 steps.
- Every 500 steps during the first 10,000 steps.
- Every 1000 steps during the first 100,000 steps.
- Every 10,000 steps for the remainder of training (ending at 1M steps).

A.2 Lexical class comparisons

We assessed the effect of lexical class on age of acquisition in children and each language model. As described in the text, when lexical class reached significance based on the likelihood ratio test (accounting for log-frequency, MLU, n-chars, and concreteness), we ran a one-way analysis of covariance (ANCOVA) with log-frequency as a covariate. There was a significant effect of lexical class in children and the unidirectional language models (the LSTM and GPT-2; $p < 0.001$).

Pairwise differences between lexical classes were assessed using Tukey's honestly significant difference (HSD) test. Significant pairwise differences are listed in Table 5.