Predicting age of acquisition for children's early vocabulary in five languages using language model surprisal

Eva Portelance, Yuguang Duan, Michael C. Frank, Gary Lupyan

March 2023

Abstract

What makes a word easy to learn? Early-learned words are frequent and tend to name concrete referents. But words typically do not occur in isolation. Some words are predictable from their contexts; others less so. Here, we investigate whether predictability relates to when children start producing different words (Age of Acquisition; AoA). We operationalized predictability in terms of a word's surprisal in child directed speech, computed using n-gram and long-short-term-memory (LSTM) language models. Predictability derived from LSTMs was generally a better predictor than predictability derived from n-gram models. Across five languages, average surprisal was positively correlated with the AoA of predicates and function words, but not nouns. Controlling for concreteness and word frequency, more predictable words were learned earlier. Differences in predictability between languages were associated with cross-linguistic differences in AoA: the same word was produced earlier in languages where the word was more predictable, confirming a prediction of the Linguistic Niche Hypothesis.

1 Introduction

In the first 2 years of life, children's grammatical knowledge and lexicon grow in tandem (Bates et al., 1994; Brinchmann, Braeken, & Lyster, 2019; Frank, Braginsky, Marchman, & Yurovsky, 2021). In addition, the order in which children acquire their first words show remarkable consistency (Clark, 1993; Tardif et al., 2008; Goodman, Dale, & Li, 2008; Braginsky, Yurovsky, Marchman, & Frank, 2019). For example, the words 'ball', 'car', and 'nose' are, on average, learned almost a year earlier than words like 'drawer', 'green' and 'animal'. This general pattern holds across multiple languages (Braginsky et al., 2019). Modeling when words are acquired can therefore help us understand what factors drive language learning more generally.

One way to study these ordering effects is by attempting to predict a word's age of acquisition (AoA) from lexical properties such as its part of speech, frequency, length, and concreteness (Goodman et al., 2008; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012; Braginsky et al., 2019). On average, nouns, tend to be learned before verbs; more frequent words before less frequent words, shorter words before longer words, and words with concrete referents before those referring to more abstract entities. Controlling for part of speech, frequency tends to be the most important predictors of AoA. Here, we go beyond using these word-level predictors, by considering the linguistic context in which words appear in speech directed at children.

One such contextual predictor, previously examined by Braginsky et al. (2019), is the mean length of utterance (MLU) in which the word occurs. MLU can be considered as a proxy for syntactic complexity. If word learning is constrained by a child's ability to parse longer utterances (which are often more complex), we should find longer MLUs to be associated with later AoAs. This is precisely what Braginsky et al. (2019) found, though MLU was a significant predictor only for predicates and function words (a result that turns out to be relevant to the current investigation). Others have used contextual diversity – a measure of semantic co-occurrence – as a predictor of AoA (Hills, Maouene, Riordan, & Smith, 2010). They found that words which appeared in more diverse contexts tended to be learnt earlier. This effect appeared to only hold for predicates and function words when controlling for the effect of word frequency. Though such a predictor can be considered a proxy for some semantic factors, it does not directly measure syntactic complexity.

Here, we examine another contextual factor – the predictability of words from their linguistic context. A word's predictability takes into account a word's linguistic context as a whole, including both syntactic and semantic information, and allow us to identify words that might be more or less difficult to learn from linguistic context.

There are several reasons for focusing on predictability. First, it is easy to compute, as we detail below. Second, it is psychologically real – we know that statistical factors such as transition probabilities play important roles in just about every aspect of early language learning (Saffran, 2020). Predictability (operationalized in terms of *surprisal*, see below) has also been shown to be a strong predictor of processing difficulty in adult psycholinguistic experiments (Levy, 2008; Demberg & Keller, 2008; Smith & Levy, 2013). More predictable words (i.e., those with lower surprisal) tend to be easier to process.[1] Finally, focusing on predictability may help unlock a puzzle concerning linguistic complexity. Namely, why aren't languages simpler than they are?

Thus, to help us understand the role of sequential predictability on the complexity and growth of children's vocabulary as well as on the complexity of languages as a whole, we propose to study the relation that exists between word predicability and AoA, treating AoA as a proxy for word learning difficulty. We frame our experiments around the following two sets of concrete questions:

1. Is a word's predictability in linguistic context a good predictor of how difficult a word is to learn beyond previously known predictors? If so, how much context matters? And, are these effects the same for all words or only some and why?

2. Second, are these effects observable across many languages and linguistic communities or are they isolated to only some? And, do differences in surprisal predict differences in AoA across languages?

2 Background

2.1 Operationalizing predictability

To determine a word's predictability, we use computational language models that consider a word's previous sequential linguistic context. Specifically, word predictability can be measured as the surprisal of a word – its negative log probability in a given context, averaged across all contexts in which it appears – given a *language model* as our probability model.

Language models are sequential predictive models trained to generate linguistic output which are commonly used in natural language processing (NLP). They do so by learning probability distributions over strings of words in a corpus, where the length of these word strings may vary.[2] The additional information contained in these substrings in the form of preceding words is what allows us to consider a word's predictability given previous syntactic and semantic context. In this paper, we consider two types of language models: n-gram models and LSTM language models (Sundermeyer, Schlüter, & Ney, 2012). N-gram models are simple conditional probability models while LSTM models are more sophisticated neural network models.

There has been growing interest in training or evaluating neural network language models on corpora that resemble the linguistic input of children more than the standard newspaper, Wikipedia, or web corpora used in NLP. This interest is motivated by two main goals: first, determining how neural network models come to learn languages, and second, using these models as tools to better understand language learning in children. In one model-centered investigation, Huebner, Sulem, Cynthia, and Roth (2021) found that training language models on corpora of child-directed utterances can actually help these models perform better on grammatical knowledge evaluations than when trained on similar sized or larger corpora of traditional NLP data, suggesting that child-directed language may help grammar learning. Chang and Bergen (2022) have also proposed to use the average surprisal of words as a proxy for the AoA of words for language models in order to develop a new model evaluation task. They evaluated whether previously known predictors of the AoA of words in children – like frequency, concreteness, MLU, lexical category, and number of characters – also predicted the 'AoA of words' in language models, and found that there are clear differences between the orders in which children and language models acquire words.

Language models have also been proposed as tools to evaluate children's language development. For example, Sagae (2021) suggest using LSTMs trained on children's utterances to quantify children's syntactic development, finding that they perform as well or better than previous metrics.

All of this work – both model-based and child-focused – has been limited to English data. Here, we expand our analyses to cross-linguistic data, considering models trained on five different languages: English, German, French, Spanish, and Mandarin.

[1] For an overview of empirical evidence supporting the validity of surprisal as a predictor of processing difficulty, see Hale, 2016.

[2] Context size can vary depending on the language model, ranging from a single previous token in n-gram models to a bounded ordered list of all previous or subsequent tokens in an LSTM model.

2.2 The origin of linguistic complexity

It has been long believed that all languages are equally complex, but this assumption has now been seriously challenged (e.g., see Newmeyer, 2002; Sampson, Gil, & Trudgill, 2009, for discussion). Much of the focus has stemmed from the observation that languages spoken by larger communities, those with many nonnative speakers, and those that have undergone substantial language contact (factors that are often, but not always positively correlated), tend to be morphologically simpler (Trudgill, 2011; Wray & Grace, 2007). In a large-scale correlational study, Lupyan and Dale (2010) showed that 'larger' languages tended to be morphologically simpler: they had simpler verb inflectional paradigms, simpler noun-verb agreement, and were more likely to use lexical rather than inflectional strategies for conveying information like tense, aspect, and evidentiality (see also Koplenig, 2019; Winters, Kirby, & Smith, 2015; Bentz, Dediu, Verkerk, & Jäger, 2018). Additionally, several lab experiments have found a causal link between group size and emergent linguistic complexity: larger groups of participants converged on simpler and more transparently-structured languages than smaller groups of participants (Raviv, Meyer, & Lev-Ari, 2019a, 2019b).

Despite this converging evidence of a link between social factors and language complexity, there is, at present, little consensus on the specific mechanisms that give rise to this relationship. Proposals range from simplification being an inevitable consequence of language contact (McWhorter, 2001), learning (and subsequent transmission) of the language by adults who fail to master the language in its full complexity (Trudgill, 2011; Lupyan & Dale, 2010; Winters et al., 2015), differences in input variability (Nettle, 2012; Atkinson, Kirby, & Smith, 2015), a pressure on language used by more heterogenuous populations to rely less on shared common ground (Wray & Grace, 2007), and differences in how linguistic innovations diffuse through social networks (Reali, Chater, & Christiansen, 2018). Regardless of which explanation proves correct, their common focus on linguistic simplification means they have little to say about why we find so much linguistic complexity in the first place. One may understand why the spread of a language is accompanied by simplification while having nothing to say concerning why these languages were so complex to begin with.

One possibility is that there may be a trade-off between optimizing languages for use by larger/more diverse groups, and optimizing languages for maximally efficient learning by young children. Consistent with this possibility is the observation that languages spoken by smaller groups are not only morphologically more complex, but are also more compressible than languages spoken by smaller groups (Lupyan & Dale, 2010; Lupyan, 2019; Koplenig, 2019). Greater compressibility is intimately linked to predictability: a more compressible language is one that contains utterances where one part is better predicted from other parts, making them informationally redundant. For example, in the utterance "I am going to play with the ball", the '-ing' of going is highly predictable from the preceding words, as is the 'to' and 'the'. And although one cannot be sure that the last word is 'ball', it is much more likely than other words like 'hamburger', 'patio', and 'spheroid', and so its occurrence conveys that much less novelty (i.e., information). An intriguing possibility is that the reason smaller languages are more compressible – i.e., have overall higher predictability / lower surprisal – is that predictability is especially important to young children learning their first language (Lupyan & Dale, 2010, 2016).

But does compressibility actually impact children's language learning? If it does, we may find that *within* a language, more predictable words are easier to learn. Of course there are all sorts of reasons why some words are both more predictable *and* easier to learn. We can strengthen the logic in two ways. First, by statistically controlling for known predictors of AoA such as frequency (which is closely related to a word's surprisal because – given no context – more frequent words are always more likely).

Second, we can examine whether cross-linguistic *differences* in AoA are predicted by cross-linguistic differences in word predictiveness. Despite AoA being similarly predicted by factors like frequency and concreteness across many languages, there *are* cross-linguistic differences in when children start producing words that mean roughly the same thing. For example, the AoA estimate for 'hot' in American English is 18 months — nearly 4 months earlier than it is for 'caliente' in Mexican Spanish. In contrast, the AoA for 'green' in American English is about 3 month later than for 'verde' in Mexican Spanish. What explains these differences? One candidate is differences word frequency. Perhaps 'hot' is simply more frequent in English than Spanish and 'green' is less frequent in English than Spanish. We will test whether this is indeed the case, and whether cross-linguistic differences in word predictability account for additional variance.

Table 1: Amount of data available in CHILDES and Wordbank by language

Language	**Child-directed tokens (CHILDES)**	**Dialect**	**Vocabulary reports (Wordbank)**
English	25,659,263	American	7,955
		British	23,129
		Australian	1,497
German	5,663,294		1,181
French	3,183,037	European	863
		Quebecois	1,364
Spanish	2,267,707	European	1,005
		Mexican	1,934
Mandarin	2,369,896	Beijing	1,938
		Taiwanese	2,654

2.3 Our approach

In the rest of this paper, we will take the following approach: for each language, we fit a set of language models on a corpus of child directed utterances and extract the average surprisal of words for which we have AoA estimates in children. We then compare regression models of children's AoA with average surprisal as one key predictor in concert with previous significant predictors, using cross-validation to estimate out-of-sample performance. We present two different modeling methods in two experiments. The first considers the effect of predictability on the AoA of words in each language individually, while the second considers its effect on all languages as a whole. We close this paper by considering how what we have learnt about the relation between word predictability and AoA can inform our understanding about the role predictability plays on both children's growth vocabulary and the complexity of languages as a whole.

3 Data

We relied on two types of data: (1) Corpora of child-directed utterances used to train the language models.[3]These were taken from the CHILDES database (MacWhinney, 2000); (2) Age of Acquisition estimates. These were based on parental reports of children's language use, taken from the Wordbank repository (Frank, Braginsky, Yurovsky, & Marchman, 2016). We go into further detail in the following subsections about both these resources and the data they contain. Importantly, in order for a language to be considered in this cross-linguistic study, there had to be sufficient data in this language in both of these resources. This criterion narrowed down the list of languages we could consider to English, German, French, Spanish, and Mandarin. English was by far the most represented language in CHILDES, but there were still enough data for the other languages to be able to fit our language models (see Table 1 for the amounts of data available in each language for both CHILDES and Wordbank).

3.1 CHILDES and child-directed utterances

The CHILDES database (MacWhinney, 2000) is a repository of child language data, containing text transcripts of child-caregiver interactions as well as video and sound recordings of some of these interactions. The data comes from many different studies conducted during the past 60 years, spanning multiple languages and countries. For the most part, the children in these studies range in age between nine months old and five and a half years old.

For this paper, we only considered text transcription data and no other modalities. For each of the five languages considered (English, German, French, Spanish, Mandarin), we collected all of the available transcripts across all corpora available through the childes-db API (Sanchez et al., 2018) in July 2021. We then removed all utterances spoken by the target child, leaving only the utterances said to the child or around the child. These utterances can

[3]All of the data, models, and experiment code presented in this paper are publicly available at `www.github.com/evaportelance/multilingual-aoa-prediction`.

be considered as an estimate of the linguistic input the children in these transcript have access to. We combined all of these child-directed utterances[4] into a corpus used to fit the language models presented in the next section and to calculate the relative frequency of words for the regression models presented in our experiments. The result was 5 corpora of child-directed utterances, one for each language. Unlike the Wordbank database (Frank et al., 2016) presented in the next subsection, childes-db does not explicitly distinguish data based on dialectal varieties of each language, so we use the same aggregated data for each language across all varieties when fitting our models.

3.2 Wordbank and age of acquisition estimates

The Wordbank database (Frank et al., 2016) is a repository of parental reports about their children's vocabularies – essentially, a checklist of words where parents can check off words their child produces or understands. Most of these reports are versions of the MacArthur-Bates Communicative Development Inventories (CDI) (Fenson et al., 1993). The database is a collection of reports originating from different studies that were conducted across the world. These studies and the vocabulary checklists they use are dialect-specific, so for each of five languages we consider in this study, we collected data from all available dialects. We did not combine the data from different dialects into single languages as each dialect contains different word lists on their reports, leaving fewer words at their intersection.

Our predictive target is the age at which a word is acquired. Since not all children learn a given word at the same time, we instead follow prior work in quantifying AoA as the age at which 50% of children are reported to produce a word on the CDI (Goodman et al., 2008).[5]. There are a number of methods to estimate this 50% point from a group of binary responses for children of different ages. The simplest method is to determine the youngest age group at which the empirical proportion of children producing the word is $> 50\%$, but this approach has several shortcomings. If words are very hard or very easy to learn, then it is possible that for the covered age range some words never reach the 50% point (e.g., *beside*), or have already surpassed the 50% point (e.g., *Mommy*) for even the youngest children. Such words would have to be discarded if we were to use this method. Another issue is that this approach is susceptible to bias AoA estimates towards ages for which more CDI instruments were available since the number of observations at each age is not equal (i.e., there may be more CDI instruments from 24-month-olds than with 20-month-olds in the dataset, but this density shouldn't lead to more words being acquired at exactly 24 months). For these reasons, we used Bayesian generalized linear models predicting acquisition as a function of age to estimate the AoA for each word, following the method suggested in Frank et al., 2021[6].

From the reports available in each language, we narrowed down the list of items used in our experiments to all single word items on the forms that were classified as either nouns, predicates (verbs and adjectives), or function words (closed class words like pronouns, prepositions, question words, connectives, determiners). We excluded items that were multi-word expressions (e.g. 'all gone') or that were classified as being part of the "other" lexical category, which included animal sounds, onomatopoeia, and other non-word expressions. Words were also excluded if they weren't in the five thousand most common words in each language in our corpora of child-directed utterances from CHILDES, as this was the vocabulary size used for the language models (described in the next section). Table 2 contains the exact number of items taken from Wordbank that we considered for each language, as well as their breakdown by lexical category.

4 Language models and predictability

To determine the predictability of words in the child-directed utterance corpora described above, we use language models as our probability models. Language models define probability distributions over subsequent words in a given context. Here, we consider the predictability of words solely based on linguistic contextual information. Specifically, we define the overall predictability of a word, w_i, as its average surprisal, or negative log probability, across all its contexts of use, C (eq. 1).

$$\sum_{C:w_i \in C} -\log P\left(w_i \mid w_1, \ldots, w_{i-1}\right) \times \frac{1}{|C|} \tag{1}$$

[4]We will use the term child-directed somewhat loosely here such that it may also refer to utterances that were directed to other adults or children present, but that the target child could still hear.

[5]We chose to use these AoA estimates from parental reports over those of Kuperman et al., 2012 which exist for a much larger vocabulary, because the latter are based on adult estimates of their own AoA, rather than timely reports of children's AoA.

[6]For a more detailed description of this method see Appendix E of Frank et al., 2021.

Table 2: Number of items and their breakdown by lexical category in each language

Language	**Number of items**	**Nouns**	**Predicates**	**Function words**
English (American)	563	301 (53%)	165 (29%)	97 (17%)
English (British)	333	196 (59%)	100 (30%)	37 (11%)
English (Australian)	368	230 (63%)	138 (37%)	0 (0%)
German	361	189 (52%)	108 (30%)	64 (18%)
French (European)	394	222 (56%)	119 (30%)	53 (13%)
French (Quebecois)	468	248 (53%)	159 (34%)	61 (13%)
Spanish (European)	430	212 (49%)	121 (28%)	97 (23%)
Spanish (Mexican)	339	181 (53%)	90 (27%)	68 (20%)
Mandarin (Beijing)	500	269 (54%)	185 (37%)	46 (9%)
Mandarin (Taiwanese)	412	257 (62%)	119 (29%)	36(8%)

where $w_1, \dots, w_{i-1}$ is a sequence of words of bounded length representing the preceding linguistic context.

There are many different types of language models. They vary in terms of the context sizes they consider, in how they represent words, and in how they come to calculate the overall probability of a word. As our definition in eq. 1 suggests, we only consider language models that take into account preceding linguistic context (and not following context) and do so in an incremental order. Specifically, the experiments that follow will contain average surprisal values obtained from two types of language models, n-gram models and LSTM models.

4.1 N-gram language models

N-gram models are basic language models that consider contexts of sequence length n, such that a bi-gram model keeps track of two word sequences and the surprisal of a word is based on a single preceding word, and a tri-gram keep track of 3 word sequences and surprisal of a word is based on the two preceding words. Thus, given our formula in eq. 1, we simply need to replace i by n to determine the average surprisal of a word for a given n-gram model.

In an n-gram probability model, the probability of a word w_n in a given context, $w_1, \dots, w_{n-1}$, or $P(w_n \mid w_1, \dots, w_{n-1})$, is simply its normalized count across all words that follow this context in the corpus. In this study, we used four different n-gram models: uni-grams, bi-grams, tri-grams, and four-grams. Note that uni-gram models only track single word contexts, or simply represent the normalized frequency counts of words, so the average surprisal of a word for a uni-gram model is equal to its negative log frequency.

One downside of n-gram models is that the context size is fixed across the whole probability model. This means that though some words may be better predicted from a single preceding word while others may be better predicted by two preceding words, we can only consider one of these context sizes at a time. LSTM models can help us get around this problem.

4.2 LSTM language models

Recurrent neural networks (RNNs) which use long-short term memory gating unit layers (Hochreiter & Schmidhuber, 1997), commonly known as LSTMs, are neural networks that can be trained on sequential data, such as sentences, up to some bounded maximum length n. These models can be used for language modeling (Sundermeyer et al., 2012) and have become a staple baseline that continue to be used in NLP because of there useful analytical properties, even though more recent model architectures outperform them (Vaswani et al., 2017). Furthermore, regular RNNs have previously been proposed as cognitive models for language learning (Elman, 1990, 1993; Christiansen, Allen, & Seidenberg, 1998), however, these earlier models were computationally limited and could be used only with small schematic datasets; in contrast, LSTMs can be applied to larger datasets. Thus, LSTM language models lend themselves well to our analyses.

LSTM language models process utterances incrementally and make use of nested layers of hidden units to learn abstract representations that can predict sequential dependencies between words across a range of dependency lengths (Linzen, Dupoux, & Goldberg, 2016). LSTM neural units use a gating system that allows them to 'forget' some of the

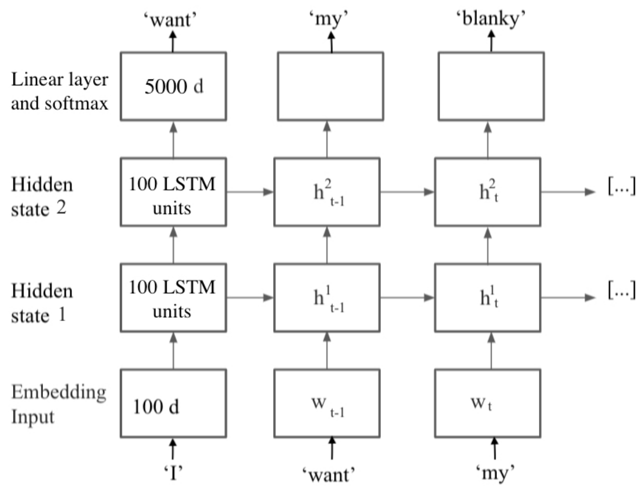

Figure 1: The LSTM model architecture incrementally processing the utterance 'I want my Blanky'.

previous states while 'remembering' others, thus learning to prioritize some dependencies in a sequence over others at each state. So, unlike n-gram models, when determining the average surprisal (eq. 1) of a word for these models, the preceding contexts, $w_1, \ldots, w_{i-1}$ in C, can vary in length for a given word w_i. Further, the probability of a word in context can weigh the importance of preceding words differently based on the information encoded in the model's different layers. The added richness of these representations may lead to a better probability model overall.

For the experiments that follow, we use a two-layered LSTM language model (Figure 1). The model has randomly initialized 100-dimensional word embeddings as its input layer, which are updated during learning. Hidden states encode information about the preceding context. At each time-step, the current word embedding w_t and the hidden state from the previous time-step h^1_{t-1} are passed through a transformation function, resulting in a new hidden state h^1_t. This hidden state h^1_t and the hidden state from the previous time-step in the second hidden layer h^2_{t-1} are then also passed through a transformation function, resulting in a new hidden state h^2_t. This final hidden state is then resized through a linear layer to the size of our vocabulary before going through a softmax transformation to produce the output – a distribution over the whole vocabulary representing a prediction about the upcoming word. We use a vocabulary size of 5,000, representing the most frequent words, because we found that including the 5,000 most common words usually resulted in the inclusion of almost all the words we had on our AoA word lists for all languages.

We performed cross-validation tests to find the parameter settings for the LSTM language models that best minimized overall surprisal. The parameters tested were the vocabulary size (2000, 4000, or 5000), the word embedding size (100, 150), the hidden dimension size for the LSTM layers (100, 150), the batch size (128, 256, 512), and the number of epochs (up to 50). We found that the optimal parameter combination was a vocabulary size of 5000, 100 dimensional word embeddings, a 100 hidden dimension size, a batch size of 256, and about 20 epochs of training.

We trained the models on all of the child-directed utterances for each language since the models were to be used as probability models and not predictive models. We were therefore not concerned with overfitting to the training data. Utterances were shuffled at each epoch of training. For further details on the model implementation see appendix A.

5 Experiment 1: The role of word predictability beyond log frequency

In this first experiment, we consider how word predictability beyond frequency predicts the AoA of words. Previous work (Goodman et al., 2008; Kuperman et al., 2012; Braginsky et al., 2019) has found that log frequency – or, equivalently, uni-gram surprisal – is an important predictor of the AoA of words; here, we evaluate the explanatory power of larger context sizes by using their residualised effect beyond uni-gram surprisal as predictors. We compare models with different versions of average surprisal, obtained using different language models: bi-gram, tri-gram, four-gram, or LSTM average surprisals. We do so using leave-one-out (LOO) cross-validation.

5.1 Predictors

There are two main types of predictors considered in our models. First, we consider several methods for computing average surprisal using language models conditioned on different sizes and types of previous linguistic context. Second, we include other predictors that have been found to be informative in previous work: concreteness and lexical category (Goodman et al., 2008; Kuperman et al., 2012; Braginsky et al., 2019). All predictors are scaled by centering their mean at zero and dividing by their standard deviation so that their magnitudes can be compared.

uni-gram average surprisal is computed as the negative logarithm of frequency. Log frequency has been found to explain substantial variance in AoA in previous work.

Residualised n-gram surprisal represents the residual variance left after fitting a linear model which predicts n-gram average surprisal as a function of uni-gram average surprisal, `n-gram average surprisal` $\sim$ `uni-gram average surprisal`. n-gram average surprisal is the predictability of a word given all contexts of size *n* in which it appears in the *n*-th position. We compare the average surprisal of bi-gram, tri-gram, and four-gram language models. If an item had multiple word forms associated to it on the parental report instrument (e.g. 'inside/in'), we used the weighted mean average surprisal across all forms – in other words, if one form was overall more frequent, then it was weighted it accordingly.[7]

Residualised LSTM surprisal represents the residual variance left after fitting a linear model which predicts LSTM average surprisal as a function of uni-gram average surprisal, `LSTM average surprisal` $\sim$ `uni-gram average surprisal`. LSTM average surprisal is calculated using the LSTM language models described above. We compute the average surprisal of each word across all of the child-directed utterances available in each language. We trained three LSTM language models in each language using different random seeds and then used the mean average surprisal across these three runs as our measure of word predictability in each language. As with n-gram surprisal, we used the weighted mean average surprisal across all word forms.

Concreteness is a rating score ranging between 1 and 5 for each word representing some measure along the abstract to concrete scale. These scores are taken from (Brysbaert, Warriner, & Kuperman, 2014). In order to obtain them for languages other than English, Wordbank data associates each item to a 'unilemma', which is an equivalent English concept across all languages for that item in the database. The concreteness score for the equivalent English concept was then used for each word. This practice follows previous work (Braginsky et al., 2019).

Lexical category is included via contrast coding of three lexical categories: nouns (common nouns), predicates (verbs and adjectives), and function words (closed-class words) following Bates et al., 1994. Word categories were derived from the categories on the CDI forms (e.g., verbs are listed as 'action words'). Lexical category serves as interacting variable with all predictors.

5.2 Regression models

Our approach involves a nested model comparison between the model containing previously known predictors, the uni-gram model, and augmented models that additionally contain residualised average surprisal from either LSTM or n-gram language models. This comparison allows us to assess whether adding information from more dynamic context sizes using LSTMs or from fixed n-gram context sizes beyond log frequency helps predict AoA. We consider the uni-gram model as our base model:

1. **The uni-gram model :** `AoA` $\sim$ `lexical category` $*$ `(uni-gram average surprisal` $+$ `concreteness)`

   We then compare it to the following augmented models:

2. **The residualised n-gram model :** `AoA` $\sim$ `lexical category` $*$`(uni-gram average surprisal` $+$ `n-gram residual average surprisal` $+$ `concreteness)`, where n-grams are either bi-grams, tri-grams, or four-grams.

[7] We also tried considering different forms as separate items; these results are available in the appendix C.

3. **The residualised LSTM model :** `AoA` $\sim$ `lexical category` $*$`(uni-gram average surprisal` $+$ `LSTM residual average surprisal` $+$ `concreteness)`

5.3 Results

We compare the uni-gram models in each language to their augmented versions which additionally contains residualised LSTM or n-gram average surprisals as predictors. We analyse the difference between the base uni-gram models and the augmented models using both Leave-one-out (LOO) cross-validation and an ANOVA nested model comparison across languages. We report the mean absolute deviation (MAD) across all LOO model fits in each language: the lower the MAD, the better the model fit.

5.3.1 Predictability overall

Table 3: Model comparison results augmenting unigram surprisal model with residualised surprisals by language and model. Numbers indicate leave-one-out cross validation mean absolute deviation (in months) with 95% CIs.

Language	unigram	2-gram	3-gram	4-gram	LSTM
English (American)	$\mathbf{2.0}_{[1.86,2.14]}$	$2.02_{[1.88,2.16]}$	$2.01_{[1.87,2.15]}$	$2.01_{[1.87,2.14]}$	$2.01_{[1.87,2.15]}*$
English (British)	$2.19_{[2.0,2.39]}$	$2.22_{[2.02,2.41]}$	$2.21_{[2.01,2.4]}$	$2.21_{[2.01,2.4]}$	$\mathbf{2.18}_{[1.99,2.37]}*$
English (Australian)	$\mathbf{1.93}_{[1.76,2.1]}$	$\mathbf{1.93}_{[1.76,2.1]}$	$1.94_{[1.77,2.11]}$	$1.94_{[1.77,2.11]}$	$\mathbf{1.93}_{[1.76,2.11]}$
German	$\mathbf{2.2}_{[1.99,2.41]}$	$2.22_{[2.02,2.43]}$	$2.22_{[2.01,2.43]}$	$2.21_{[2,2.42]}$	$2.3_{[2.02,2.58]}$
French (European)	$\mathbf{2.3}_{[2.1,2.5]}$	$2.33_{[2.12,2.53]}$	$2.31_{[2.11,2.52]}$	$\mathbf{2.3}_{[2.1,2.51]}$	$2.46_{[2.12,2.8]}$
French (Quebecois)	$\mathbf{2.54}_{[2.35,2.73]}$	$2.56_{[2.37,2.75]}$	$\mathbf{2.54}_{[2.31,2.73]}$	$2.55_{[2.36,2.74]}$	$\mathbf{2.54}_{[2.35,2.73]}$
Spanish (European)	$\mathbf{2.48}_{[2.29,2.67]}$	$2.49_{[2.3,2.69]}$	$2.5_{[2.31,2.69]}$	$2.51_{[2.31,2.7]}$	$2.49_{[2.3,2.68]}$
Spanish (Mexican)	$2.08_{[1.91,2.25]}$	$2.1_{[1.93,2.27]}$	$2.1_{[1.93,2.28]}$	$2.09_{[1.92,2.27]}$	$\mathbf{2.06}_{[1.89,2.23]}*$
Mandarin (Beijing)	$1.89_{[1.75,2.02]}$	$1.89_{[1.75,2.02]}*$	$\mathbf{1.86}_{[1.73,2.0]}**$	$1.87_{[1.74,2.01]}*$	$1.9_{[1.76,2.04]}$
Mandarin (Taiwanese)	$\mathbf{2.98}_{[2.75,3.21]}$	$3.0_{[2.77,3.23]}$	$3.0_{[2.78,3.23]}$	$3.0_{[2.77,3.22]}*$	$\mathbf{2.98}_{[2.75,3.2]}*$

* ($p < 0.05$) and ** ($p < 0.01$) indicate that the nested ANOVA is significant.

The overall results are available in table 3, where we report MAD and 95% confidence intervals across LOO cross-validation folds, as well as p values from our ANOVA nested model comparison. The models with the smallest MAD, or best fits, are bolded.

The nested ANOVA results suggest that adding residualised LSTM surprisal as a predictor significantly increases model fit in four of the ten datasets. However, given the cross-validation results which show that there is very little difference in MAD values between our base models and augmented models, we may want to be cautious about these results and suggest instead that the overall effects are likely small.

The large majority of items across all languages are nouns. Average surprisal using more linguistic context may not be such a good predictor of the AoA of nouns, at least not as much as simple frequency (Portelance, Degen, & Frank, 2020). For this reason, the fact that we do not see a large difference between our base and augmented models here may be expected. If we consider the interaction between lexical category and residualized LSTM surprisal, we find that the interaction terms for predicates are significant ($p < 0.05$) in five of the languages (*English (American), English (British), English (Australian), French (Quebecois), Spanish (Mexican)*).[8]

[8] We used contrast coding for our lexical categories, such that interaction terms between a lexical category and residualized surprisal can be interpreted as main effects, indicating a difference from the overall mean.

5.3.2 Predictability by lexical category

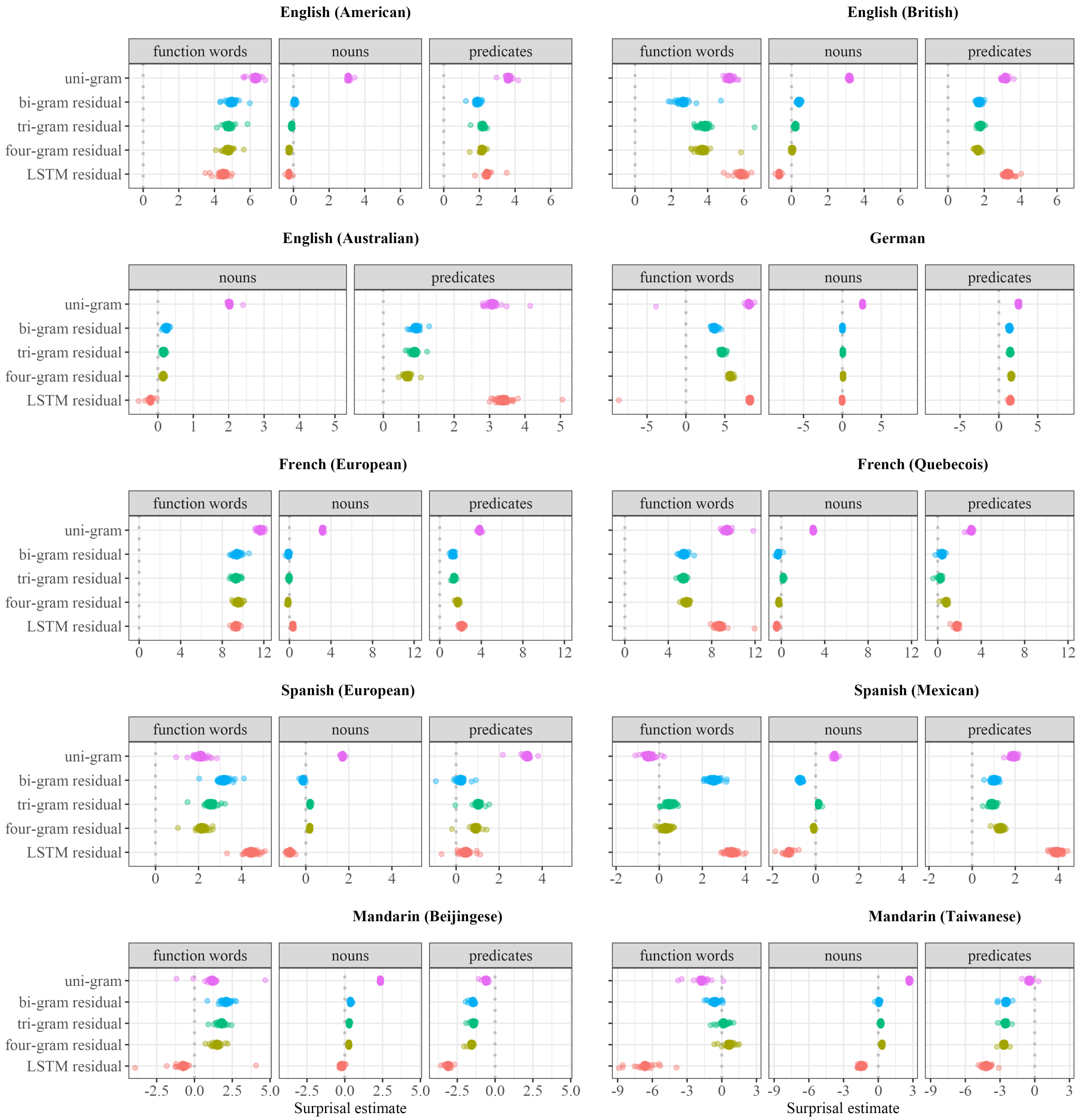

Figure 2: Coefficient estimates for surprisal values by lexical category in each language

Adding residualised surprisal beyond uni-gram surprisal as a predictor of AoA does not substantially improve overall prediction, but it may make a difference for predicting the AoA of specific words. We next consider how effect sizes may differ by lexical category.[9] Here, we do so by plotting the estimated coefficients by lexical category for the different surprisal predictors.[10]

[9]Since lexical categories differ significantly in their concreteness ratings, we also explored whether the by-lexical-category differences in surprisal effects seen in this subsection can be better explained by an interaction between residualised surprisal and concreteness in Appendix E.

[10]In appendix D, we also fit models to items in each lexical category individually and consider their MAD; these results show that nouns are generally best predicted by simply using the uni-gram base model, while predicates and function words often benefit from the augmented versions which include some form of higher order residualised surprisal, LSTM and tri-gram residualised surprisal generally doing best.

The plots in figure 2 show the estimated coefficients for variables across LOO folds by lexical category. In each of these graphs, a point represents the estimated coefficient of a predictor for one fitted model run from the LOO cross-validation, so for example in *English (American)* there are 563 items and therefore 563 folds all of which have been plotted here.

For each language in figure 2, we see that residualised bi-gram, tri-gram, four-gram, and LSTM surprisals generally have little to no effect for nouns, but have a positive effect for function words and predicates in most languages. In other words, the harder function words and predicates are to predict in their linguistic contexts, the later they are acquired. The exception to this rule is Mandarin, where both function words and predicates show a negative effect for all types of residual surprisal, meaning predicates with higher average surprisal, or less predictability, seem to be learned earlier and have a lower AoA.

Several explanations for the disparate results in Mandarin are possible. First, Mandarin has been known to pattern differently to other languages in other early word learning studies. For example, it has been said that Mandarin learners do not show the same 'noun bias' – the observations that learners tend to initially learn to produce more nouns before increasing their productions of verbs – that English learners seem to have (Tardif, Gelman, & Xu, 1999), an observation that has been reproduced using the Wordbank parental report data (Frank et al., 2021; Yee, 2020). Instead, Mandarin learners have been found to produce more predicates early on during learning, even though both English and Mandarin speaking parents produce relatively more predicates than nouns (Tardif et al., 1999). Another second possible explanation for this difference may however lie in some of the Wordbank data itself. As Frank et al. (ch.11; 2019) note, there seem to be some discrepancies with some of the Mandarin data in the repository, specifically forms collected for *Mandarin (Beijing)* from the Tardif, Fletcher, Liang, & Kaciroti, 2009 study seem to show a much stronger predicate bias than other forms available for *Mandarin (Beijing)* or other languages. This data imbalance may also be contributing to this effect for predicates. On the other hand, the *Mandarin (Taiwanese)* data is not known to have this same issue, yet it shows a similarly negative effect for average surprisal with predicates albeit weaker than that of *Mandarin (Beijing)*.[11]

Finally, uni-gram surprisal has a positive effect for all lexical categories across almost all language: words that are more predictable based on their frequency are generally learned earlier, here again the exception being Mandarin function words and predicates.

5.4 Interim conclusions

In this first experiment, we addressed our first set of research questions: Is a word's predictability in linguistic context a good predictor of how difficult a word is to learn beyond previously known predictors? If so, how much context matters? And, are these effects the same for all words or only some and why? Since log frequency has been found to be the most important predictor to date in previous work and since average surprisal is dependent on word frequency, we elected to remove any variance explained by frequency and used residualised surprisal values as our predictor. We found that predictability beyond log frequency was only useful when trying to predict the AoA of predicates and function words, such that less predictable words in linguistic context were generally acquired later. The effects of predictability on word learning were different for different lexical categories. As for how much contexts matters when determining word predictability, based on our plots in figure 2, residual LSTM surprisal generally has a larger effect size then all other n-gram surprisals, suggesting that dynamic context sizes which vary from word to word may be most useful in measuring predictability.

6 Experiment 2: The role of word predictability across languages

In the previous section, we fit models for different languages using different word lists (table 2). It was therefore hard to determine whether the differences in effect sizes we observed across languages were caused by variation within each individual language word lists or by real distinctions in effect sizes. To remedy this issue, we needed to use the same word list for all languages. We achieved this by unifying words with the same concept across languages using their unilemmas, and then taking their intersection.

[11] It is also possible that these differences are simply due to strength and nature of the resources we had at our disposition for Mandarin. For example, it is possible that different corpora in CHILDES data used different character systems or word segmentation norms, since there exists different standards for this language, which in turn could have led to noisier data.

First, we aggregated our data across languages to find the intersection of all unilemmas. There were a total of 89 unilemmas for which we have AoA estimates in all languages. These included 64 nouns and 25 predicates; no function words are left because one language, *English (Australian)*, did not have function words in its word lists. Although we were left with very few unilemmas overall, since we have 10 language groups, we still had 640 noun items and 250 predicate items. Since our previous results suggested that the effects of residualised surprisal differed by lexical category, we split our data into nouns and predicates, testing each category separately. We then fitted a mixed-effects models on each category with by-language random effects to compare effect sizes across languages.

6.1 Regression models

The models in this section differ from the those used in the previous section as they include an additional random effect term: `(1 + uni-gram average surprisal + residual average surprisal | language)`. This term means that the models consider by-language differences in coefficient estimates for intercepts, the effects of uni-gram average surprisal and LSTM or n-gram residual average surprisal, taking those differences as a source of variance in the data. We did not include a separate random effect by language for concreteness ratings, since these were based on the unilemmas for words and were therefore identical in all languages.

1. **The residualised n-gram mixed-effects model :** `AoA` $\sim$ `uni-gram average surprisal` + `n-gram residual average surprisal` + `concreteness` + `(1 + uni-gram average surprisal + n-gram residual average surprisal | language)`, where n-grams are either bi-grams, tri-grams, or four-grams.

2. **The residualised LSTM mixed-effects model :** `AoA` $\sim$ `uni-gram average surprisal` + `LSTM residual average surprisal` + `concreteness` + `(1 + uni-gram average surprisal + LSTM residual average surprisal | language)`

6.2 Results

All predictors were scaled by language, centering predictors at zero and setting the standard deviation to one[12].

6.2.1 Predictability overall

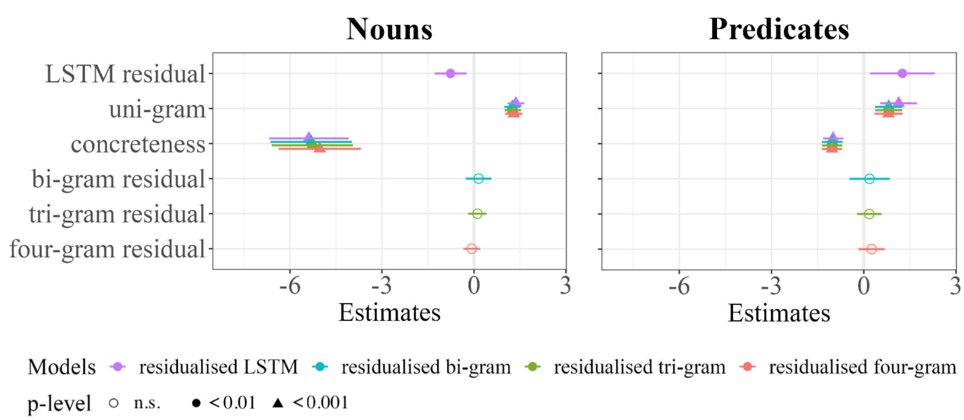

Figure 3: Mean coefficient estimates for surprisal values by lexical category for mixed-effects model using the intersection of items in all languages

As shown in figure 3, when we examine words occurring in all languages, we find that the effects of frequency (uni-gram surprisal) and concreteness are the same as before. Words with higher uni-gram surprisal are learned later, i.e., more frequent words are learned earlier. Controlling for frequency and concreteness, none of the residual

[12]Doing so for each language individually allows us to compare the effects across languages without worrying that any differences in variation are due to our predictive n-gram or LSTM models having different surprisal distributions in different languages.

n-gram surprisal predictors showed significant effects overall, but LSTM surprisal did. The reason may be that n-grams only include information from a fixed length of context, while LSTM surprisal contains information from more dynamic context sizes in the course of word learning. For predicates, higher LSTM surprisal predicts later AoA, so less predictable predicates in linguistic context are harder to learn and therefore learned later. However, for nouns, residual LSTM surprisal instead shows a negative effect, meaning that less predictable nouns tend to be learned earlier.

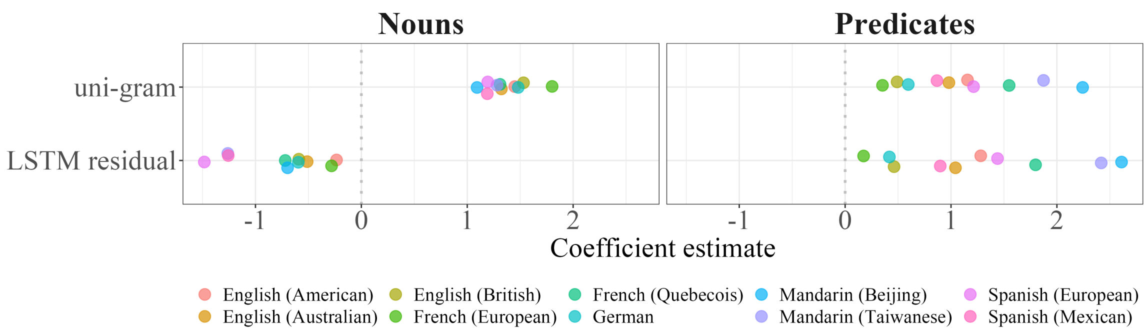

Figure 4: Coefficient estimates for surprisal values for nouns and predicates in each language. Only words with data for all languages are included in this analysis.

A possible explanation for the case of nouns is that cross-situational learning for nouns across diverse social and visual contexts in theory may also lead to more diverse linguistic contexts and therefore higher surprisal (Roy, Frank, DeCamp, Miller, & Roy, 2015). For predicates, however, children rely heavily on linguistic context to learn these words. Therefore, higher predictability in linguistic context may indicate that there are more linguistic cues helping children learn certain predicates earlier than others.

6.2.2 Surprisal and AoA across languages

Figure 4 plots effect sizes of uni-gram surprisal and residual LSTM surprisal for each individual language by lexical category [13]. Comparing the predictive power of surprisal across languages, we find that uni-gram surprisal (word frequency) is equally, and positively predictive of noun AoA. LSTM surprisal for nouns shows some variation, being somewhat more negative – greater surprisal is associated with earlier AoA – in Mandarin and Spanish than in other tested languages. In the case of predicates, we see more variation. This result may simply be due to there being fewer items used to fit the predicate mixed-effects model, with only 250 words that were predicates compared to 640 words in the nouns mixed-effects model. The positive effect of both uni-gram surprisal and LSTM surprisal is strongest for Mandarin (Taiwanese and Beijing). This is an interesting result because our analyses from our first experiment using separate models for each language (section 5) found that Mandarin was the only language to show a negative effect for surprisal when predicting the AoA of predicates. It is likely that this difference in effect polarity is due to items included the mixed-effects models being a small sub-sample of those included in the previous experiment, suggesting that the specific words used to fit the models may be introducing a bias towards one or the other effect direction.

Although children learning different languages follow broadly similar learning trajectories (e.g., learning frequent concrete nouns before less common abstract verbs), there *are* differences between the AoA of words that mean roughly the same things. Taking just the words available in all the languages, we find that for 45% of the word/language pairs, the mean difference in AoA is greater than 2 months. For 16% of the pairs, it is greater than 4 months. Are these differences predictable by differences in surprisal for the same word across languages?

To answer this question, we first computed differences in AoA, uni-gram surprisal, and LSTM residual surprisal (i.e., LSTM surprisal controlling for uni-gram surprisal) for each lemma attested in each pair of languages (E.g., American English and Spanish). Uni-gram surprisal and LSTM residual average surprisal values were scaled within each language before computing the differences. We then used a mixed-effects model to predict cross-linguistic differences in AoA from differences in uni-gram surprisal and differences in LSTM residual surprisal. Since some pairs of languages belong to the same base language, e.g., English (American) and English (Australian), we also added a binary variable to the model indicating whether the two languages have the same base language. Moreover, we also

[13]The effect sizes by language for residualised n-gram surprisal values are available in appendix F.

included by-unilemma, by-the-first-base-language, and by-the-second-base-language random intercepts in the model, i.e., `(1 | unilemma)` + `(1 | base language 1)` + `(1 | base language 2)` because those random effects could also be a source of variance in the data. Concreteness was not included as a predictor because we did not have separate concreteness estimates for each language.

- **The AoA differences mixed-effects model :** `AoA differences` ∼ `differences in uni-gram average surprisal` + `differences in LSTM residual average surprisal` + `whether the base languages are the same` + `(1 | unilemma)` + `(1 | base language 1)` + `(1 | base language 2)`

From our final analysis (Figure 5), we see that for both nouns and predicates, if a word occurs more frequently in language 1 than its equivalent[14] in language 2, then its AoA is likely to be earlier in language 1. The same pattern is true with higher-order LSTM surprisal for predicates, though for nouns we see an opposite effect, which aligns with our previous finding that LSTM surprisal has opposite effects on predicates and nouns. A greater difference in surprisal between two languages is associated with earlier AoAs for nouns and later AoAs for predicates.

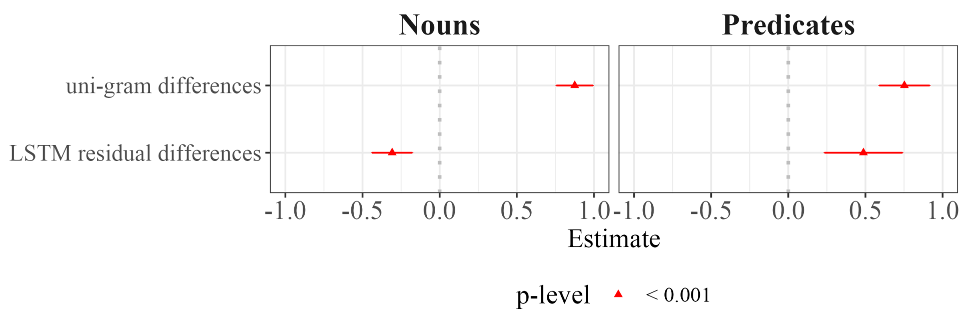

Figure 5: Differences in frequency (uni-gram surprisal) and LSTM-based word surprisal are associated with differences in AoA. A positive coefficient indicates that if the same word has higher surprisal in language 1 than language 2, it is learned later in language 1 than language 2.

6.3 Interim conclusions

In this second experiment, we tried to answer our second set of questions: Are these effects observable across many languages and linguistic communities or are they isolated to only some? And, do differences in surprisal predict differences in AoA across languages? We found that word predictability in context was a good predictor across language, showing little variation in effect size. The effects for nouns and predicates had polarity, however, such that less predictable nouns in linguistic contexts were learnt earlier, while less predictable predicates were learnt later. Additionally, in our secondary analysis we found that for predicates, there was a positive relationship between differences in surprisal and differences in AoA: predicate lemmas that have a greater surprisal in language 1 than language 2 tend to have a later AoA in language 1.

7 Discussion

Are words that are more predictable given their linguistic context (i.e., words with lower surprisal), easier to learn by young children? For predicates and fucntion words, the answer seems to be yes. Although uni-gram surprisal (i.e., word frequency) was always a better predictor of AoA than higher-order surprisal, surprisal derived from an LSTM neural network predicted AoA beyond uni-gram surprisal and the words' frequency. This effect was largely restricted to predicates and function words – words whose meaning is especially dependent on their context.

One route by which linguistic context is known to affect learning is via *syntactic bootstrapping*. Introduced by Brown (1957) and coined by Gleitman (1990), syntactic bootstrapping refers to the use of syntactic context to help determine meaning, especially for predicates. A wealth of evidence from individual experiments and computational models supports the ability of children to use linguistic information to make inferences about meaning (?, ?; Fisher,

[14]Equivalent words refer to word with the same unilemma in Wordbank.

Jin, & Scott, 2020). Although our experiments do not directly test this idea, our results are certainly congruent with syntactic bootstrapping accounts.

A somewhat puzzling finding is that when predicting the AoA of nouns, greater LSTM surprisal was associated with *earlier* age of acquisition – the opposite of what we observed for predicates and function words. One possible explanation for this pattern is that noun learning is more dependent on the perception of concrete referents, the salience of which is often obvious from the extra-linguistic context (Gillette, Gleitman, Gleitman, & Lederer, 1999). Frequent nouns may appear in a broader set of contexts, which would *increase* their higher-order surprisal while at the same time making it more semantically interconnected and improve learning (Hills et al., 2010; Roy et al., 2015).

We found the effects of surprisal to be mostly consistent across the tested languages. The one exception was Mandarin. In our first experiment using separate models for each language, Mandarin showed the opposite effects for surprisal on predicates and function words compared to other languages. However, in experiment 2 using a single-mixed effects model over common unilemmas, Mandarin showed the largest effect of LSTM-based surprisal on AoA.

Finally, our results shed light on a puzzle of linguistic complexity. It is possible to predict certain aspects of a language's structure from knowing the environment in which it is learned and used (Lupyan & Dale, 2016; Bentz et al., 2018). Smaller language are more complex and less compressible. But why? One hypothesis is that this redundancy may provide a benefit by facilitating child language learning. We found initial support for this idea: at least for predicates, a decrease in surprisal is associated with faster word learning suggesting that redundancy (lower surprisal) may provide more effective learning opportunities for young children. Further tests of this hypothesis would require comparison across a much wider range of languages, however.

8 Conclusion and Limitations

We investigated the relationship between word predictability – formulated as average surprisal in linguistic contexts and Age of Acquisition – a proxy for how difficult learning a given word is for children. Word predictability seems to be especially important for the acquisition of predicates and function words, rather than nouns. Less predictable verbs and adjective in linguistic contexts tend to be acquired later by children and are therefore harder to learn. We found this effect to be true is multiple languages, and that differences in surprisal predict differences in AoA. This finding is broadly in line with the prediction that decreasing surprisal (i.e., increasing redundancy) may – all else being equal – facilitate child language learning, particularly of more relational words.

Theories of language learning must explain not just words like 'ball' and 'dog' but also words whose meaning in context is almost entirely dependent on other words. Sequential models like the LSTM we used here may be a promising avenue to help explain the acquisition of these 'hard words'.

A major limitation of our analyses is that it is based on the data available in CHILDES. These corpora of child-directed utterances were – by necessity – assembled from many sub-corpora from different children and studies, and may not be the best representations of the regularity, idiosyncrasies, and contextual diversity found in the language targeted to a single child. Furthermore, our corpora contained utterances directed at children who were older in some cases than those surveyed for the AoA estimates. Ideally, these sentences would span the exact same developmental stages. A second limitation is that the words and AoA estimates we use are based on parents' reports of their children's language. Although validated and highly reliable, these data cannot capture the full richness of individual children's vocabularies and language use. Finally, our sample of language is restricted to languages with a large digital footprint. We hope future studies can begin to overcome these limitations.

References

Atkinson, M., Kirby, S., & Smith, K. (2015). Speaker input variability does not explain why larger populations have simpler languages. *PLOS ONE*, *10*(6), e0129463. Retrieved 2016-05-24, from `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0129463` doi: 10.1371/journal.pone.0129463

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., ... Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*(1), 85–123.

Bentz, C., Dediu, D., Verkerk, A., & Jäger, G. (2018). The evolution of language families is shaped by the environment beyond neutral drift. *Nature Human Behaviour*, *2*(11), 816–821. doi: 10.1038/s41562-018-0457-6

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, 1–16.

Brinchmann, E. I., Braeken, J., & Lyster, S.-A. H. (2019). Is there a direct relation between the development of vocabulary and grammar? *Developmental Science*, *22*(1), e12709.

Brown, R. W. (1957). Linguistic determinism and the part of speech. *The Journal of Abnormal and Social Psychology*, *55*, 1–5.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Chang, T. A., & Bergen, B. K. (2022). Word acquisition in neural language models. *Transactions of the Association of Computational Linguistics*. Retrieved from https://arxiv.org/abs/2110.02406

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and cognitive processes*, *13*(2-3), 221–268.

Clark, E. V. (1993). Early lexical development. In *The lexicon in acquisition* (p. 21–42). Cambridge University Press.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

Fenson, L., Dale, P., Reznick, J. S., Thal, D., Bates, E., Hartung, J., ... Reilly, J. (1993). MacArthur Communicative Inventories: User's guide and technical manual. *San Diego*.

Fisher, C., Jin, K.-s., & Scott, R. M. (2020). The developmental origins of syntactic bootstrapping. *Topics in Cognitive Science*, *12*(1), 48–77.

Frank, M. C., Braginsky, M., Marchman, V., & Yurovsky, D. (2021). *Variability and Consistency in Early Language Learning: The Wordbank Project*. MIT Press.

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.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135–176. doi: 10.1016/s0010-0277(99)00036-0

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3–55.

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.

Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, *10*(9), 397–412.

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

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th conference on computational natural language learning* (pp. 624–646).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. Retrieved from https://arxiv.org/pdf/1412.6980.pdf

Koplenig, A. (2019). Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers. *Royal Society Open Science*, *6*(2), 181274. Retrieved 2022-11-08, from https://royalsocietypublishing.org/doi/full/10.1098/rsos.181274 (Publisher: Royal Society) doi: 10.1098/rsos.181274

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990.

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

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *TACL*, *4*, 521–535.

Lupyan, G. (2019). Larger languages have higher entropy.. Retrieved from https://cle.ppls.ed.ac.uk/index.php/ielc2019/

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE*, *5*(1), e8559. Retrieved from http://dx.doi.org/10.1371/journal.pone.0008559 doi: 10.1371/journal.pone.0008559

Lupyan, G., & Dale, R. (2016). Why are there different languages? the role of adaptation in linguistic diversity. *Trends in Cognitive Sciences*, *20*(9), 649–660. doi: http://dx.doi.org/10.1016/j.tics.2016.07.005

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition.* Erlbaum.

McWhorter, J. (2001). The world's simplest grammars are creole grammars. *Linguistic Typology*, *5*(2), 125–166.

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1597), 1829–1836. Retrieved 2022-11-08, from https://royalsocietypublishing.org/doi/10.1098/rstb.2011.0216 (Publisher: Royal Society) doi: 10.1098/rstb.2011.0216

Newmeyer, F. (2002). Uniformitarian assumptions and language evolution research. In A. Wray (Ed.), *The transitions to language* (pp. 359–375). Oxford University Press. Retrieved from http://groups.lis.illinois.edu/amag/langev/paper/newmeyer02uniformitarianAssumptions.html

Portelance, E., Degen, J., & Frank, M. C. (2020). Predicting Age of Acquisition in Early Word Learning Using Recurrent Neural Networks. In *Cogsci.*

Raviv, L., Meyer, A., & Lev-Ari, S. (2019a). Compositional structure can emerge without generational transmission. *Cognition*, *182*, 151–164. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010027718302464 doi: 10.1016/j.cognition.2018.09.010

Raviv, L., Meyer, A., & Lev-Ari, S. (2019b). Larger communities create more systematic languages. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1907), 20191262. Retrieved from http://royalsocietypublishing.org/doi/10.1098/rspb.2019.1262 doi: 10.1098/rspb.2019.1262

Reali, F., Chater, N., & Christiansen, M. H. (2018). Simpler grammar, larger vocabulary: How population size affects language. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1871), 20172586. Retrieved 2019-03-01, from https://royalsocietypublishing.org/doi/full/10.1098/rspb.2017.2586 doi: 10.1098/rspb.2017.2586

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663–12668.

Saffran, J. R. (2020). Statistical language learning in infancy. *Child development perspectives*, *14*(1), 49–54. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8078161/ doi: 10.1111/cdep.12355

Sagae, K. (2021). Tracking child language development with neural network language models. *Frontiers in Psychology*, *12*.

Sampson, G., Gil, D., & Trudgill, P. (2009). *Language complexity as an evolving variable*. Oxford University Press, USA.

Sanchez, A., Meylan, S., Braginsky, M., MacDonald, K., Yurovsky, D., & Frank, M. C. (2018). childes-db: a flexible and reproducible interface to the Child Language Data Exchange System.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–319.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In *13th annual conference of the international speech communication association.*

Tardif, T., Fletcher, P., Liang, W., & Kaciroti, N. (2009). Early vocabulary development in Mandarin (Putonghua) and Cantonese. *Journal of child language*, *36*, 1115–1144.

Tardif, T., Fletcher, P., Liang, W., Zhang, Z., Kaciroti, N., & Marchman, V. A. (2008). Baby's first 10 words. *Developmental Psychology*, *44*(4), 929.

Tardif, T., Gelman, S. A., & Xu, F. (1999). Putting the "noun bias" in context: A comparison of English and Mandarin. *Child development*, *70*, 620–635.

Trudgill, P. (2011). *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, *7*(3), 415–449. Retrieved from https://www.cambridge.org/core/journals/language-and-cognition/article/abs/languages-adapt-to-their-contextual-niche/83E9F516875C340E0A9263B4A7C38F43 (Publisher: Cambridge University Press) doi: 10.1017/langcog.2014.35

Wray, A., & Grace, G. (2007). The consequences of talking to strangers : Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, *117*(3), 543–578. Retrieved from http://cat.inist.fr/?aModele=afficheN&cpsidt=18462025

Yee, S. (2020). Is noun bias universal? Evidence from Chinese and Korean compared with French and English. In *Studies in the linguistic sciences: Illinois working papers* (pp. 32–44).

A LSTM model details

Models were implemented in Python 3.8 using the Pytorch 1.8.1 and trained on a single Nvidia RTX 3080 GPU. They are composed of an embedding layer, two sequential and fully connected LSTM layers, and a linear layer. We used an Adam optimizer across all models during training (Kingma & Ba, 2014) and simple cross-entropy across the outputs of all states as our loss function.

A.1 Hyperparameters

Here are the hyperparameters tested, bolded ones are the settings we used for the experiments reported in the main paper.
number of epochs: up to 50, **20** used
learning rate: **0.0001**
batch size: 128 **/ 256 /** 512
hidden size: **100 /** 150
embedding size: **100 /** 150
vocabulary size: 2000 / 4000 **/ 5000**
random seed: **[0:2]**

B Relationship between uni-gram and higher order surprisals

Table 4: Pearson correlation between LSTM, bi-gram, tri-gram, four-gram average surprisal and uni-gram average surprisal in each language

Language	**1gm-LSTM**	**1gm-2gm**	**1gm-3gm**	**1gm-4gm**
English (American)	0.65	0.91	0.80	0.56
English (British)	0.68	0.89	0.77	0.5
English (Australian)	0.72	0.85	0.68	0.34
German	0.63	0.81	0.51	0.18
French (European)	0.65	0.89	0.59	0.23
French (Quebecois)	0.69	0.87	0.55	0.21
Spanish (European)	0.97	0.87	0.42	0.16
Spanish (Mexican)	0.97	0.83	0.44	0.12
Mandarin (Beijing)	0.96	0.81	0.55	0.43
Mandarin (Taiwanese)	0.97	0.81	0.55	0.41

C Approach 1 - All words as separate items

The following section reproduces the analyses from our first approach, but instead of combining items with multiple lexical forms (e.g. in/inside), it considers each form as a separate item when fitting models. This method may be problematic if one form is much more frequent then another since it will give them equal importance when fitting regression models, as each word form would be considered independently from its related forms.

Table 5 shows the results overall of our model comparison using LOO cross-validation and reporting MAD. The results are similar to those reported in the main paper.

Table 5: Approach 1 model comparison results augmenting uni-gram surprisal model with residualised surprisal by language

	LOO MAD$_{[95\% CI]}$				
Language	**1gm-base**	**2gm-rd**	**3gm-rd**	**4gm-rd**	**lstm-rd**
English (American)	$\mathbf{2.01}_{[1.87,2.15]}$	$2.03_{[1.89,2.17]}$	$\mathbf{2.01}_{[1.88,2.15]}$	$\mathbf{2.01}_{[1.8,2.15]}$	$2.02_{[1.88,2.16]}$
English (British)	$2.19_{[2.0,2.38]}$	$2.21_{[2.02,2.41]}$	$2.21_{[2.01,2.4]}$	$2.20_{[2.01,2.4]}$	$\mathbf{2.17}_{[1.87,2.15]}*$
English (Australian)	$1.94_{[1.77,2.11]}$	$\mathbf{1.93}_{[1.77,2.1]}$	$1.94_{[1.77,2.11]}$	$1.94_{[1.77,2.11]}$	$\mathbf{1.93}_{[1.76,2.1]}$
German	$\mathbf{2.2}_{[1.99,2.41]}$	$2.22_{[2.02,2.43]}$	$2.22_{[2.01,2.43}$	$2.21_{[2,2.42]}$	$2.3_{[2.02,2.58]}$
French (European)	$\mathbf{2.32}_{[2.13,2.52]}$	$2.35_{[2.15,2.54]}$	$\mathbf{2.32}_{[2.13,2.53]}$	$\mathbf{2.32}_{[2.12,2.51]}$	$2.34_{[2.15,2.54]}$
French (Quebecois)	$\mathbf{2.56}_{[2.37,2.74]}$	$2.57_{[2.38,2.76]}$	$\mathbf{2.56}_{[2.37,2.75]}$	$\mathbf{2.56}_{[2.38,2.75]}$	$2.57_{[2.38,2.76]}$
Spanish (European)	$2.53_{[2.34,2.71]}$	$\mathbf{2.52}_{[2.34,2.71]}*$	$2.54_{[2.36,2.73]}$	$2.55_{[2.37,2.73]}$	$2.53_{[2.35,2.71]}$
Spanish (Mexican)	$2.09_{[1.92,2.26]}$	$2.1_{[1.93,2.27]}$	$2.11_{[1.93,2.28]}$	$2.10_{[1.92,2.27]}$	$\mathbf{2.06}_{[1.89,2.34]}**$
Mandarin (Beijing)	$1.88_{[1.75,2.01]}$	$1.88_{[1.74,2.01]}*$	$\mathbf{1.86}_{[1.73,2.99]}**$	$1.87_{[1.73,2]}*$	$1.89_{[1.75,2.02]}$
Mandarin (Taiwanese)	$3.01_{[2.78,3.23]}$	$3.03_{[2.8,3.26]}$	$3.03_{[2.8,3.26]}$	$3.03_{[2.8,3.25]}$	$\mathbf{3}_{[2.78,3.22]}**$

* (p <0.05) and ** (p <0.01) indicate that the nested ANOVA is significant.

Figure 6 shows the estimated coefficients for surprisal and residualised surprisal by language and lexical category. These results are similar to those in the main paper.

D Approach 1 - By lexical category MAD

Table tab:loo-resid-augment-lex shows the LOO MAD results for models fit on each lexical category separately.

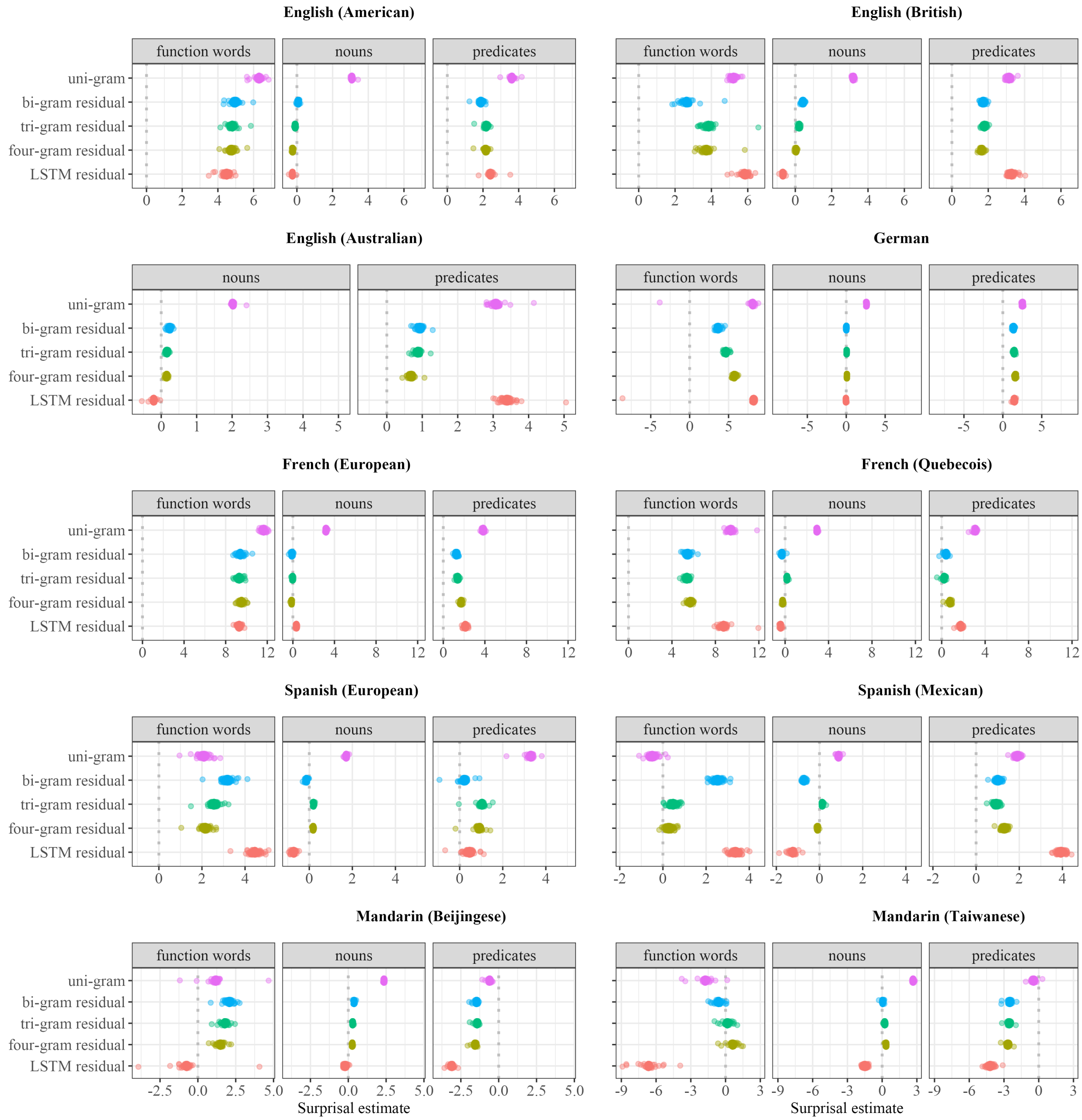

Figure 6: Coefficient estimates by lexical category in each language using first approach with all words as separate items.

Table 6: Approach 1 model comparison results augmenting uni-gram surprisal model with residualised surprisal by language and lexical category

		LOO MAD$_{[95\% CI]}$				
Language	**Category**	**1gm-base**	**2gm-rd**	**3gm-rd**	**4gm-rd**	**lstm-rd**
English (American)	Nouns	**1.95**$_{[1.76,2.14]}$	1.96$_{[1.77,2.15]}$	1.96$_{[1.77,2.15]}$	1.96$_{[1.77,2.15]}$	1.96$_{[1.77,2.15]}$
	Predicates	1.79$_{[1.58,2.01]}$	1.81$_{[1.59,2.03]}$	1.78$_{[1.55,2.0]}$	**1.77**$_{[1.55,2.0]}$	1.81$_{[1.58,2.04]}$
	Function words	**2.52**$_{[2.13,2.91]}$	2.56$_{[2.16,2.97]}$	2.55$_{[2.15,2.94]}$	2.54$_{[2.15,2.94]}$	2.54$_{[2.15,2.93]}$
English (British)	Nouns	2.21$_{[1.96,2.47]}$	2.22$_{[1.97,2.48]}$	2.23$_{[1.97,2.48]}$	2.23$_{[1.97,2.49]}$	**2.2**$_{[1.94,2.45]}$
	Predicates	1.92$_{[1.62,2.22]}$	1.93$_{[1.62,2.23]}$	**1.91**$_{[1.61,2.21]}$	1.92$_{[1.62,2.22]}$	**1.91**$_{[1.6,2.21]}$
	Function words	2.83$_{[2.17,3.49]}$	2.97$_{[2.29,3.66]}$	2.92$_{[2.24,3.6]}$	2.88$_{[2.21,3.55]}$	**2.17**$_{[1.87,2.15]}$
English (Australian)	Nouns	**2.06**$_{[1.83,2.29]}$	**2.06**$_{[1.83,2.29]}$	**2.06**$_{[1.83,2.29]}$	2.06$_{[1.83,2.29]}$	2.07$_{[1.83,2.3]}$
	Predicates	1.73$_{[1.49,1.96]}$	1.72$_{[1.48,1.96]}$	1.72$_{[1.49,1.96]}$	1.74$_{[1.5,2.97]}$	**1.71**$_{[1.48,1.95]}$
German	Nouns	**1.86**$_{[1.66,2.05]}$	**1.86**$_{[1.67,2.06]}$	**1.86**$_{[1.67,2.06]}$	**1.86**$_{[1.67,2.06]}$	1.87$_{[1.67,2.06]}$
	Predicates	**1.86**$_{[1.58,2.14]}$	1.88$_{[1.6,2.17]}$	1.88$_{[1.59,2.16]}$	1.87$_{[1.59,2.15]}$	1.89$_{[1.6,2.18]}$
	Function words	**3.77**$_{[2.94,4.6]}$	3.86$_{[3.06,4.65]}$	3.84$_{[3.02,4.66]}$	3.81$_{[2.97,4.65]}$	4.27$_{[2.99,5.55]}$
French (European)	Nouns	**2.03**$_{[1.76,2.3]}$	2.05$_{[1.77,2.32]}$	2.05$_{[1.77,2.32]}$	2.04$_{[1.77,2.31]}$	2.05$_{[1.78,2.32]}$
	Predicates	2.45$_{[2.12,2.79]}$	2.48$_{[2.15,2.81]}$	**2.43**$_{[2.1,2.76]}$	2.44$_{[2.11,2.78]}$	2.48$_{[2.14,2.82]}$
	Function words	**3.06**$_{[2.46,3.67]}$	3.16$_{[2.51,3.82]}$	3.16$_{[2.52,3.81]}$	3.07$_{[2.43,3.71]}$	4.14$_{[2.06,6.21]}$
French (Quebecois)	Nouns	**2.47**$_{[2.21,2.72]}$	2.49$_{[2.23,2.75]}$	**2.47**$_{[2.21,2.74]}$	2.48$_{[2.22,2.74]}$	2.48$_{[2.22,2.73]}$
	Predicates	2.48$_{[2.17,2.79]}$	2.48$_{[2.16,2.79]}$	**2.46**$_{[2.14,2.77]}$	2.49$_{[2.17,2.8]}$	2.45$_{[2.13,2.76]}$
	Function words	3.01$_{[2.43,3.6]}$	3.07$_{[2.46,3.67]}$	3.05$_{[2.46,3.64]}$	**3.0**$_{[2.42,3.59]}$	3.04$_{[2.46,3.62]}$
Spanish (European)	Nouns	**2.11**$_{[1.86,2.37]}$	2.12$_{[1.86,2.37]}$	2.13$_{[1.88,2.38]}$	2.13$_{[1.88,2.38]}$	**2.11**$_{[1.85,2.36]}$
	Predicates	**2.16**$_{[1.81,2.5]}$	2.17$_{[1.83,2.52]}$	2.18$_{[1.83,2.53]}$	2.19$_{[1.84,2.54]}$	2.18$_{[1.83,2.52]}$
	Function words	**3.69**$_{[3.29,4.09]}$	3.71$_{[3.3,4.13]}$	3.71$_{[3.3,4.13]}$	3.74$_{[3.32,4.15]}$	3.71$_{[3.31,4.1]}$
Spanish (Mexican)	Nouns	**2.04**$_{[1.79,2.29]}$	**2.04**$_{[1.79,2.29]}$	2.05$_{[1.79,2.3]}$	2.05$_{[1.8,2.3]}$	2.05$_{[1.8,2.3]}$
	Predicates	1.79$_{[1.56,2.03]}$	1.83$_{[1.59,2.07]}$	1.82$_{[1.58,2.06]}$	1.8$_{[1.56,2.05]}$	**1.72**$_{[1.47,1.96]}$
	Function words	2.58$_{[2.15,3.0]}$	2.63$_{[2.23,3.04]}$	2.62$_{[2.19,3.06]}$	2.59$_{[2.16,3.03]}$	**2.54**$_{[2.13,2.95]}$
Mandarin (Beijing)	Nouns	1.91$_{[1.74,2.09]}$	**1.9**$_{[1.72,2.08]}$	**1.9**$_{[1.72,2.08]}$	**1.9**$_{[1.73,2.08]}$	1.92$_{[1.74,2.1]}$
	Predicates	1.67$_{[1.47,1.87]}$	1.69$_{[1.48,1.89]}$	1.66$_{[1.46,1.87]}$	1.67$_{[1.46,1.87]}$	**1.65**$_{[1.45,1.85]}$
	Function words	2.59$_{[1.94,3.25]}$	2.6$_{[1.97,3.34]}$	**2.46**$_{[1.82,3.09]}$	2.51$_{[1.87,3.15]}$	2.76$_{[2.02,3.49]}$
Mandarin (Taiwanese)	Nouns	**2.88**$_{[2.6,3.17]}$	2.9$_{[2.61,3.18]}$	2.9$_{[2.61,3.18]}$	2.89$_{[2.61,3.18]}$	2.9$_{[2.61,3.18]}$
	Predicates	**3.07**$_{[2.65,3.48]}$	3.09$_{[2.65,3.52]}$	3.08$_{[2.67,3.49]}$	3.08$_{[2.68,3.48]}$*	**3.07**$_{[2.66,3.47]}$
	Function words	3.39$_{[2.56,4.22]}$	3.48$_{[2.64,4.33]}$	3.5$_{[2.65,4.34]}$	3.47$_{[2.63,4.3]}$	**3.25**$_{[2.46,4.04]}$

E Approach 1 - Interactions with concreteness

Since lexical categories differ significantly in their concreteness ratings (figure 7), it is possible that the differences in effect sizes for surprisal by lexical categories are caused by concreteness. In this set of model comparisons, we replace the interactions between surprisal and lexical category with interactions between surprisal and concreteness. We want to know whether interactions with concreteness may better predict AoA than interactions with lexical category.

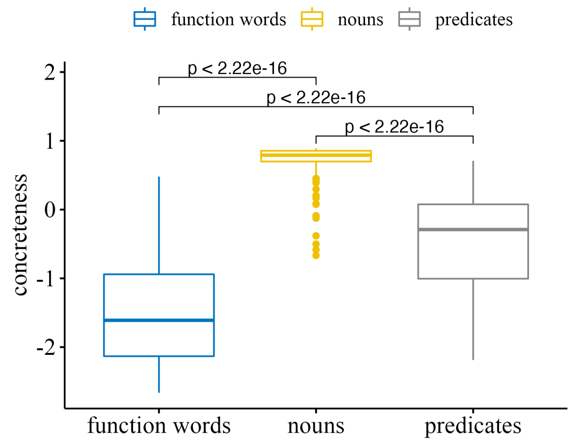

Figure 7: Concreteness distribution for different lexical categories

E.1 Regression models

The regression models used here resemble those from the second approach in the main paper, though this time the interaction terms are concreteness, while lexical category is just another predictor.

1. **The uni-gram concreteness model :** `AoA` $\sim$ `concreteness` $*$`(uni-gram average surprisal` $+$ `lexical category)`
2. **The residualised n-gram concreteness model :** `AoA` $\sim$ `concreteness` $*$`(uni-gram average surprisal` $+$ `n-gram residual average surprisal` $+$ `lexical category)`, where n-grams are either bi-grams, tri-grams, or four-grams
3. **The residualised LSTM concreteness model :** `AoA` $\sim$ `concreteness` $*$`(uni-gram average surprisal` $+$ `LSTM residual average surprisal` $+$ `lexical category)`

E.2 Results

Table 7 shows the LOO MAD results for our approach 1 model comparison using interactions with concreteness ratings instead of lexical category. Residual surprisal seems to have less of an effect overall with these interaction terms than with those reported in the main paper.

Table 7: Approach 1 model comparison results using interaction with concreteness by language

	LOO MAD$_{[95\%\ CI]}$				
Language	**1gm-base**	**2gm-rd**	**3gm-rd**	**4gm-rd**	**lstm-rd**
English (American)	**2.0**$_{[1.87,2.14]}$	2.01$_{[1.87,2.14]}$	2.01$_{[1.87,2.14]}$	2.01$_{[1.88,2.15]}$	**2.0**$_{[1.86,2.14]}$
English (British)	**2.23**$_{[2.04,2.42]}$	2.24$_{[2.05,2.43]}$	**2.23**$_{[2.04,2.42]}$	**2.23**$_{[2.04,2.42]}$	2.24$_{[2.05,2.43]}$
English (Australian)	**1.92**$_{[1.76,2.09]}$	1.93$_{[1.76,2.09]}$	1.93$_{[1.77,2.1]}$	1.94$_{[1.77,2.1]}$	1.93$_{[1.77,2.1]}$
German	**2.2**$_{[1.99,2.42]}$	2.23$_{[2.01,2.44]}$	2.22$_{[2.01,2.43]}$	**2.2**$_{[1.99,2.41]}$	2.23$_{[2.02,2.44]}$
French (European)	**2.36**$_{[2.17,2.56]}$	2.38$_{[2.19,2.58]}$	2.38$_{[2.18,2.57]}$	2.38$_{[2.18,2.57]}$	**2.36**$_{[2.16,2.55]}$
French (Quebecois)	**2.55**$_{[2.36,2.74]}$	2.57$_{[2.38,2.75]}$	2.56$_{[2.37,2.75]}$	2.56$_{[2.37,2.75]}$	2.57$_{[2.38,2.76]}$
Spanish (European)	**2.52**$_{[2.34,2.71]}$	2.53$_{[2.34,2.71]}$	2.53$_{[2.35,2.72]}$	2.54$_{[2.36,2.72]}$	**2.52**$_{[2.34,2.71]}$
Spanish (Mexican)	**2.09**$_{[1.92,2.26]}$	2.1$_{[1.92,2.27]}$	2.1$_{[1.93,2.27]}$	2.10$_{[1.93,2.28]}$	2.1$_{[1.93,2.27]}$
Mandarin (Beijing)	1.91$_{[1.78,2.04]}$	**1.9**$_{[1.76,2.03]}$	**1.9**$_{[1.76,2.03]}$	**1.9**$_{[1.77,2.04]}$	**1.9**$_{[1.77,2.04]}$
Mandarin (Taiwanese)	2.99$_{[2.77,3.22]}$	3.01$_{[2.78,3.23]}$	3.01$_{[2.78,3.23]}$	3.01$_{[2.78,3.23]}$	**2.95**$_{[2.74,3.16]}$

F Approach 2 - N-gram predictability by language

Figure 8 shows the by language coefficient estimates for the mixed-effects models of approach 2 for residualised bi-gram, tri-gram, and four-gram models.

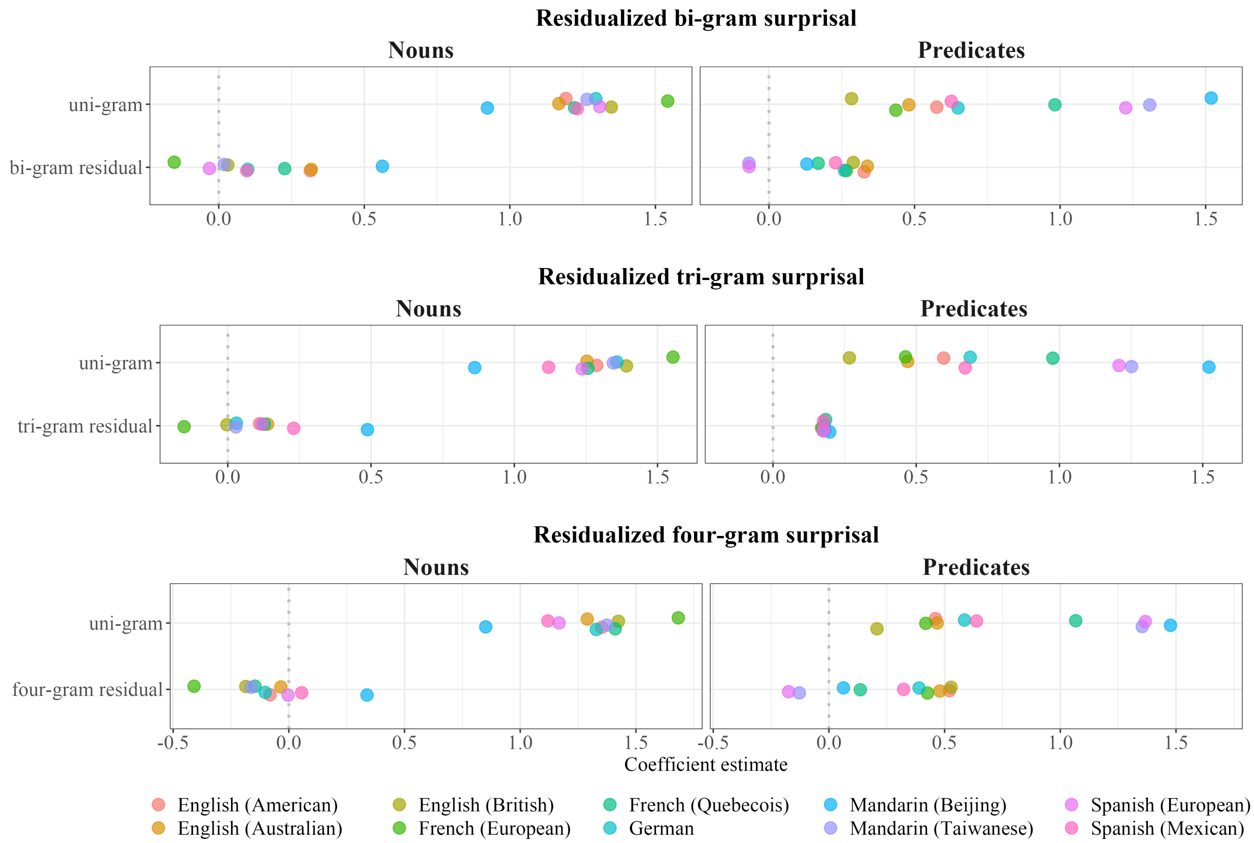

Figure 8: By-language random effects in Residualised n-gram mixed-effects models for second approach