Chapter 2

Language models and the age of acquisition of words

In this second chapter, I determine whether language models that learn word representations from linguistic context present similar word learning ordering effects to children. If similarities are found, then we can hypothesise that the input information used by models to learn word representations may also be used by children to learn new words. This can help us understand what information within children's linguistic input is useful to them when learning the meaning of new words.

Specifically, I consider whether the surprisal of words for language models can predict the age of acquisition (AoA) of those same words in children. I ask two questions: (1) does surprisal help explain when words are acquired by children beyond previously studied variables like frequency? and (2) are words which are difficult for the model to learn acquired later by children?

2.1 Background

Children's lexicon and syntactic abilities grow in tandem, resulting in a tight correlation between vocabulary size and grammatical complexity (Bates et al., 1994; Brinchmann, Braeken, & Lyster, 2019; Frank et al., 2019). Additionally, children are remarkably consistent in the order in which the acquire their first words (Clark, 1993a; Tardif et al., 2008; Goodman et al., 2008). This ordering consistency presents an opportunity: modeling when words are acquired can help us understand what drives language learning more generally.

One approach which has been used to study these ordering effects is to create quantitative models that predict the children's age of acquisition (AoA) for words. These analyses typically combine large-scale survey data about word acquisition with corpus estimates of language input from different children to make aggregate-level predictions. Such studies have investigated the effects of word properties such as frequency, number of phonemes, and concreteness, across a range of languages (Goodman et al., 2008; Kuperman et al., 2012; Braginsky et al., 2019), with simple frequency typically being the most important factor: the more frequent a word is in the child's linguistic input, the earlier it is acquired. However, these analyses have generally not considered the linguistic contexts in which words appear. Some have considered the length of utterances as a weak proxy for sentence complexity (Braginsky et al., 2019). Others have suggested using contextual diversity – a measure of semantic co-occurrence – as a possible predictor of AoA beyond frequency (Hills et al., 2010). Though such a predictor can be considered a proxy for some semantic factors, it does not directly measure syntactic complexity. Given the strong connection between vocabulary growth and syntactic ability in children's output, it seems only fair to hypothesize that the syntactic contexts in which words appear in children's language input should also be an important. I propose to use a word's overall predictability across its contexts of use in child-directed utterances as a new predictor which takes into account a word's linguistic context as a whole. To determine a word's predictability, I use computational language models that consider previous sequential linguistic context. Specifically, word predictability can be measured as the average surprisal of a word – its the negative log probability in a given context[1], averaged across all contexts in which it appears –, given a language model as our probability model. Surprisal has previously been shown to be a strong predictor of human processing difficulty in psycholinguistic experiments (Levy, 2008; Demberg & Keller, 2008; Smith & Levy, 2013), adding support for its use as a predictor of AoA.[2]

Language models are sequential predictive models trained to generate linguistic output which are commonly used in NLP. They do so by learning probability distributions over substrings in a corpus, where the size of these substrings may vary. 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. I consider

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

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

two types of language models: n-gram models and LSTM language models (Sundermeyer et al., 2012), where n-gram models are simple conditional probability models while LSTM models are more sophisticated neural network based models.

In addition to determining whether word predictability conditioned on previous linguistic context helps predict the AoA of words, this chapter also seeks to determine the answer to a related, yet different question: whether neural network based language models, here LSTMs, show similar word learning ordering effects to those of children. Since neural language models' learning objective is to minimize the surprisal of words in context, the average surprisal of a word constitutes a measure of how difficult that word is for the model to learn, in addition to being a measure of overall word predictability. Thus, average surprisal can also be used to establish a linking hypothesis between language model word learning difficulty and children's word learning ordering effects.

In the last year, there has been growing interest in training or evaluating neural network language models on corpora that resemble the linguistic input of children more than standard NLP corpora. 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. 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 then when trained on similar sized or larger corpora of traditional NLP data, like Wikipedia data, suggesting that child-directed language may help grammar learning. Following my own work (Portelance, Degen, & Frank, 2020) on which this chapter is partially based, Chang and Bergen (to appear) have 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 testing task. They evaluated whether previously known predictors of the AoA of words in children – like frequency, concreteness, mean utterance length, 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 tool to evaluate children's language development. Sagae (2021) suggest using LSTMs trained on children's utterances to quantify children's syntactic development finding that they do as well or better than previously used metrics. It should be noted that all of the work mentioned in this paragraph has however limited itself to using language models trained only on English data. In this chapter, I expand

my analyses to cross-linguistic data, considering models trained on five different languages: English, German, French, Spanish, and Mandarin.

My approach is as follows. For each of the previously mentioned languages, I train 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. I then compare regression models of children's AoA using average surprisal as one key predictor in concert with previous predictor sets using cross-validation to estimate out-of-sample performance, an additional advance on previous approaches to predicting AoA. I am then be able to determine the predictive power of this new measure and investigate how well sequential prediction difficulty in language models relates to patterns of acquisition for children.

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

2.2 Data

There are two main types of data required for the experiments in this chapter: corpora of child-directed utterances used to train the language models, and parental reports of children's vocabulary development to estimate the AoA of words for the regression models. The first type was acquired from the CHILDES database (MacWhinney, 2000) and the second from the Wordbank project repository (Frank et al., 2016). I 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 criteria narrowed down the list of languages considered in this chapter to five: English, German, French, Spanish, and Mandarin. English was by far the most represented language in CHILDES, but the other chosen languages still presented enough data to train our language models, though a rather small amount for today's standard model sizes (see table 2.1 for the amounts of data available in each language for both CHILDES and Wordbank).

2.2.1 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

Table 2.1: Amount of data available in CHILDES and Wordbank by language

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

of some of these interactions. The data comes from many different studies, some longitudinal, that were conducted in the past 60 years. These studies span 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 study, I only considered text transcription data and no other modalities. For each of the five languages considered (English, German, French, Spanish, Mandarin), I collected all of the available transcripts across all corpora available through the childes-db API (Sanchez et al., 2018) in July 2021. I then removed any utterance that was marked to have been spoken by the target child, leaving me with only the utterances said to the child or around the child. These utterances can be considered as the linguistic input the children in these transcript have access to. I combined all of these 'child-directed' utterances[3] into a corpus I use to train the language models presented in the next section and to calculate the relative frequency of words for the regression models presented in my experiments. Thus, I have 5 corpora of child-directed utterances, one for each language. For the total number of tokens (words) in each of these corpora, see table 2.1. 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

[3]I will use the term child-directed somewhat loosely here such that it also contains utterances that may have been directed to other adults or children present, but that the target child could still hear.

language across all varieties to train our models.

2.2.2 Wordbank and age of acquisition estimates

The Wordbank database (Frank et al., 2016) is a repository of parental reports of child 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 often dialect specific, so for each of five languages I consider in this study, I collected data from all available dialects. I 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, and furthermore, these reports where not always administered in the same conditions, making it hard to control for these differences.

Our predictive target is the age at which a word is acquired. We assume that AoA correlates with ease of acquisition. Since not all children learn a given word at the same time, we instead quantified AoA as the age at which 50% of children are reported to produce a word (Goodman et al., 2008). Following previous work (Braginsky et al., 2019), I used these parental reports to estimate the AoA of the words they included.[4]. There are a number of methods to estimate the 50% point. 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*). 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 filled with 24-month-olds than with 20-month-olds in the dataset). For these reasons, I use Bayesian generalized linear models fitted to the reports available

[4]A reviewer for my CogSci proceedings paper on which this chapter is based (Portelance et al., 2020) asked why we chose to use these AoA estimates over those of Kuperman et al., 2012. Though Kuperman et al., 2012 have estimates for a much larger vocabulary, they are based on adult estimates of their own AoA, rather than timely reports of children's AoA. Thus, we favored using AoA estimates collected from CDI instruments.

Table 2.2: Number of items and their breakdown by lexical category in each language

Language	**Number of items**	**Nouns**	**Predicates**	**Function words**
English (American)	560	300 (54%)	165 (29%)	95 (17%)
English (British)	332	195 (59%)	100 (30%)	37 (11%)
English (Australian)	362	227 (63%)	135 (37%)	0 (0%)
German	339	186 (55%)	102 (30%)	51 (15%)
French (European)	377	214 (57%)	114 (30%)	49 (13%)
French (Quebecois)	443	239 (54%)	151 (34%)	53 (12%)
Spanish (European)	383	199 (52%)	114 (30%)	70 (18%)
Spanish (Mexican)	298	172 (58%)	83 (28%)	43 (14%)
Mandarin (Beijingese)	429	245 (57%)	146 (34%)	38 (9%)
Mandarin (Taiwanese)	392	250 (64%)	109 (28%)	33 (8%)

for each language to estimate the AoA of words, following the method suggested in Frank et al., 2019[5].

From the reports available in each language, I narrowed down the list of items I consider in my 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); in other words I excluded items that were multi-word expressions or that were classified as of the 'other' lexical category, which included animal sounds, onomatopoeia, and other non-word expressions. The final list of words was further reduced by the fact that words had to be contained in the child-directed utterance corpora I constructed for each language (as described in the previous subsection). So words were also excluded if they weren't present in the child-directed utterance corpora built from CHILDES, and furthermore, if they weren't in the five thousand most common words in each language present in the corpora, as this was the vocabulary size used for the language models described in the next section. Table 2.2 contains the exact number of items taken from Wordbank that I considered for each language, as well as their breakdown by lexical category.

[5]For a more detailed description of this method see Appendix E of Frank et al., 2019.

2.3 Language models and predictability

For the experiments that follow, I use language models as probability models. These models allow me to determine the predictability of words in the child-directed utterance corpora I described above. Here, I consider the predictability of words solely based on linguistic contextual information. Specifically, I 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. 2.1).

$$\sum_{C:w_i \in C} -\log P\left(w_i \mid w_1, ..., w_{i-1}\right) \times \frac{1}{|C|} \tag{2.1}$$

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

In addition to being a measure of the predictability of words for language models, surprisal has been shown to be a strong predictor of human processing difficulty in psycholinguistic experiments (Levy, 2008; Demberg & Keller, 2008). These facts make this measure an ideal candidate for representing the predictability of words in children's input.

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. 2.1 suggests, I only consider language models that consider 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.

2.3.1 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. 2.1, 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, $w_1, ..., w_{n-1}$, or $P\left(w_n \mid w_1, ..., w_{n-1}\right)$, is simply its normalized count across all words that follow this context in the corpus. In this study, I considered 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.

2.3.2 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 et al., 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 this project.

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 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. 2.1) of a word for these models, the preceding contexts, $w_1, ..., w_{i-1}$ 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.

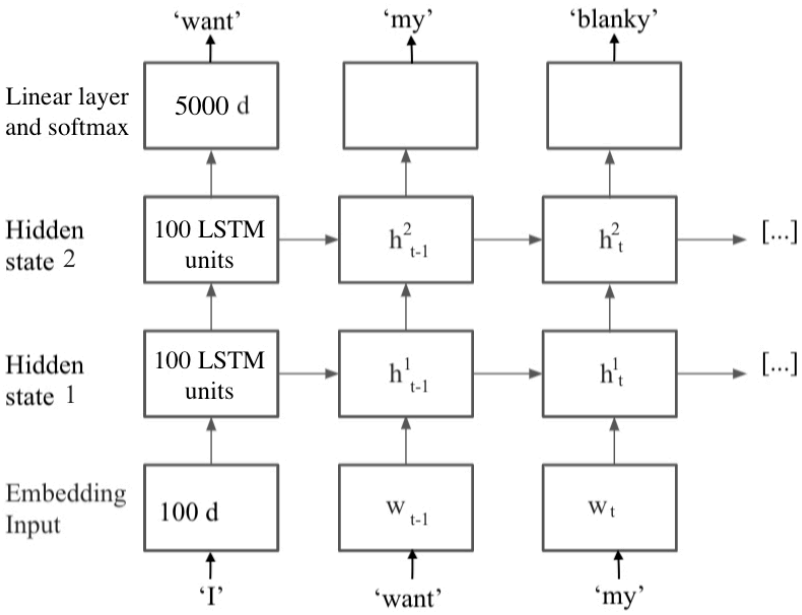


Figure 2.1: The LSTM model architecture incrementally processing the utterance 'I want my Blanky'.

For the experiments that follow, I use a two-layered LSTM language model, its architecture is illustrated in figure 2.1. 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 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. I use a vocabulary size of 5,000, representing the most frequent words. I found that including the 5,000 most common words usually resulted in the inclusion of almost all the words I had on my AoA word lists for all languages.

I performed cross-validation tests to find the parameter settings for the LSTM language models that best minimized overall surprisal. The parameters I 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). I 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.

I trained the models on all of the child-directed utterances I had for each language since the models were to be used as probability models and not predictive models. I was 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.1.

Conveniently, the LSTM language model's learning objective is to minimize the surprisal of words given their preceding context in an utterance. So, the average surprisal of a word across all the utterances in which it appears is also a measure of how difficult it is for the model to converge on a representation for that word. This observation can be leveraged in the analyses that follow to consider whether language model word learning difficulty in addition to word predictability are good predictors of the AoA of words in children.

2.4 Experiment 1 : Predicting AoA using word predictability in context

The goal of this experiment is to understand whether word predictability in children's input – measured as average surprisal for an LSTM language model – is related to difficulty of acquisition for children. I chose to use LSTM language models over n-gram models as my probability models for this experiment to explore whether model learning word difficulty is related to difficulty of acquisition for children – given that average surprisal is also a measure of model learning difficulty specifically in the case of LSTMs. Additionally, using LSTM average surprisal allows me to replicate and extend my results from Portelance et al., 2020. The two main research questions explored in this section are: (1) does average surprisal predict AoA beyond previously known predictors like frequency and word concreteness across all languages? (2) What is the relation between average surprisal and AoA for different lexical category across all languages? To answer these questions, I compare a series of regression models that include our predictors of interest using cross validation.

2.4.1 Predictors

In addition to average surprisal, I also include other predictors that have been found to be informative in previous work in my models (Goodman et al., 2008; Kuperman et al.,

2012; Braginsky et al., 2019). All predictors are scaled by centering their mean at zero and dividing them by their standard deviation so that they can be better compared.

Average surprisal is calculated using the LSTM language models described above. I compute the average surprisal of each word across all of the child-directed utterances available in each language. I trained three LSTM language models in each language using different random seeds and then used the mean average surprisal across these three runs as my measure of word predictability in each language. If an item had multiple word forms associated to it on the parental report instrument (e.g. 'inside/in'), I once again used the mean average surprisal across all word forms.

Frequency consists of the frequency counts of words taken from the same corpora of child-directed utterances used for training our language models. These counts were normalized by the size of the corpus in each language. I did not aggregate counts across inflected forms (e.g. 'give' and 'gave') since the LSTM language model considered inflected forms separately during learning. As previously mentioned, items that had a zero count were excluded. Like with average surprisal, if an item had multiple word forms associated to it on the parental report instrument (e.g. 'inside/in'), I used the mean normalized frequency 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. Thus, the concreteness score for the equivalent English concept was used for each word. This practice follows previous work (Braginsky et al., 2019).

Lexical category interactions are used with all previously listed predictors. We considered 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').

Table 2.3: Variance inflation factor (VIF) for all predictors by language.

Language	VIF score by predictor			
	Average surprisal	**Frequency**	**Concreteness**	**Lexical category**
English (American)	1.12	1.31	3.68	4.2
English (British)	1.13	1.32	3.38	4.11
English (Australian)	1.11	1.17	2.91	2.78
German	1.07	1.38	3.4	3.76
French (European)	1.05	1.46	3.6	4.53
French (Quebecois)	1.07	1.54	3.46	4.56
Spanish (European)	2.67	2.01	4.14	4.23
Spanish (Mexican)	2.65	2.24	4.06	4.76
Mandarin (Beijingese)	1.35	1.3	3.09	3.42
Mandarin (Taiwanese)	1.39	1.28	4.13	4.06

I performed a collinearity analysis to ensure that all the predictors listed above were not correlated. The Pearson correlation coefficients between average surprisal and frequency are generally very low for most languages, ranging between $r = -0.16$ and $r = -0.21$, except for Mandarin (both Beijingese and Taiwanese) were $r = -0.39$ and Spanish (both European and Mexican) were $r \approx -0.72$. As for concreteness, most scores were low, ranging from $r = -0.32$ to $r = -0.42$ with frequency, and from $r = 0.23$ to $r = 0.43$ with average surprisal. The only case that was a little higher was once again Spanish (both European and Mexican), where the correlation between concreteness and average surprisal is $r \approx -0.55$. I also report the variance inflation factor (VIF) for all predictors in all languages in table 2.3. VIFs were relatively low, with the exception of concreteness and lexical category. Lexical category is a categorical variable with relatively few levels (3 levels respectively), while concreteness ratings vary very little within each lexical category, explaining these findings. Since additionally they do not affect average surprisal or frequency, these slightly higher VIFs can be safely ignored. Overall, I did not find any worrying evidence of strong correlations between predictors in most languages nor of multicollinearity.

2.4.2 Regression models

To determine if average surprisal from LSTM language models increased the accuracy of AoA predictions, I compared linear regression models with different predictor sets using

Table 2.4: Model comparison results by language

	LOO MAD$_{[95\% \text{ CI}]}$			Nested ANOVA	
Language	**Null model**	**Base model**	**Augmented model**	$F_{(\text{Dfs})}$ **statistic**	p **value**
English (American)	$2.82_{[0.31,5.33]}$	$2.46_{[0.0,6.28]}$	$2.29_{[0.0,5.63]}$	$31.57_{(3,554)}$	$<0.0001^{***}$
English (British)	$2.74_{[0.31,5.16]}$	$2.30_{[0.24,4.36]}$	$2.27_{[0.25,4.29]}$	$4.17_{(3,333)}$	$<0.01^{**}$
English (Australian)	$2.48_{[0.31,4.66]}$	$2.38_{[0.0,6.78]}$	$2.23_{[0.0,6.06]}$	$29.1_{(2,371)}$	$<0.0001^{***}$
German	$2.75_{[0.15,5.36]}$	$2.26_{[0.0,4.61]}$	$2.25_{[0.0,4.71]}$	$7.88_{(3,350)}$	$<0.0001^{***}$
French (European)	$3.1_{[0.31,5.89]}$	$2.62_{[0.08,5.17]}$	$2.44_{[0.0,4.89]}$	$19.41_{(3,420)}$	$<0.0001^{***}$
French (Quebecois)	$3.11_{[0.34,5.87]}$	$2.73_{[0.2,5.27]}$	$2.64_{[0.06,5.24]}$	$13.45_{(3,473)}$	$<0.0001^{***}$
Spanish (European)	$2.76_{[0.3,5.23]}$	$2.65_{[0.33,4.98]}$	$2.58_{[0.28,4.87]}$	$8.97_{(3,461)}$	$<0.0001^{***}$
Spanish (Mexican)	$2.22_{[0.22,4.23]}$	$2.09_{[0.24,3.93]}$	$2.09_{[0.25,3.95]}$	$0.96_{(3,331)}$	n.s.
Mandarin (Beijingese)	$2.33_{[0.3,4.37]}$	$2.08_{[0.21,3.95]}$	$1.9_{[0.1,3.7]}$	$31.21_{(3,505)}$	$<0.0001^{***}$
Mandarin (Taiwanese)	$3.46_{[0.71,6.21]}$	$3.19_{[0.62,5.77]}$	$3.05_{[0.36,5.75]}$	$9.72_{(3,407)}$	$<0.0001^{***}$

leave-one-out (LOO) cross-validation. I evaluate the mean absolute deviation (MAD) of these models' predictions across all words, since each word represents one instance of a LOO model fit. The absolute deviation of a word is the absolute difference in months between the actual AoA estimate and the predicted AoA estimate. I compared the following models:

1. **The null model :** `AoA` ~ 1 , to ensure that all models were in fact different from null.

2. **The base model :** `AoA` $\sim$ `lexical category` $*$`(frequency` $+$ `concreteness)` , which contains both base predictors, frequency and concreteness, and their interactions with lexical category.

3. **The augmented model :** `AoA` $\sim$ `lexical category` $*$`(average surprisal` $+$ `frequency` $+$ `concreteness)` , where I add our new predictor, average surprisal, and its interaction with lexical category to those in the base model.

In addition to reporting MAD across LOO model fits, I also ran a nested model comparison between the base model and augmented model fitted on all of the data in each language and report the F-statistic and p-value to help with the reader's interpretation of results.

2.4.3 Results

The first question asked in this section was whether average surprisal can predict AoA beyond previously known predictors like frequency and word concreteness. As the results in table 2.4 show, adding average surprisal as a predictor in the augmented model leads to better fitting models in all of the languages tested except one. This observation is true both in the case of the cross-validation analyses and the nested model comparison analyses across languages. Thus, it seems fair to say that average surprisal as a measure of word predictability in context can account for a significant amount of the variance and is a good predictor of AoA in children cross-linguistically.

The one case where adding average surprisal did not make a significant difference for predicting AoA beyond frequency and concreteness was *Spanish (Mexican)*. In this case, no significant difference was found between our base and augmented models. Some possible reasons for this are that this is the language for which I had the least number of items to consider in the regression model, as well as the least amount of corpus data for training our language model from which average surprisal was taken. However, it should be noted that average surprisal was in fact a significant predictor of AoA in *Spanish (European)*. Since average surprisal and frequency values for these two models came from the same probability model and corpus, the issue most likely lies in the specific items that we have AoA for in *Spanish (Mexican)*. Based on the breakdown of items by lexical category in table 2.2, the composition of the item lists between these two varietals of Spanish are very similar. Additionally, the average count in the Spanish child-directed utterance corpus for the items in these languages is also similar, about 439 (SD = 1426) instances on average for each item in *Spanish (Mexican)* and 392 (SD = 1388) for *Spanish (European)*. This leaves only one final possible explanation, as noted at the beginning of this section, the average surprisal values from the Spanish LSTM probability model were correlated with item frequency for both dialects. However, the specific items that we have AoA estimates for in *Spanish (Mexican)* lead to slightly higher overall correlation between average surprisal and frequency (r=-0.74) then the items in *Spanish (European)* (r=-0.71); this difference may have been enough to lead to a non-significant result in one dialect, versus a significant one in the other. Either way, the high correlation between predictability in our language model and frequency suggest that the Spanish LSTMs generally use very little linguistic context to predict words, most likely learning uni-grams above all else, while other language models may favor more context. I explore the question of context size and predictability of words

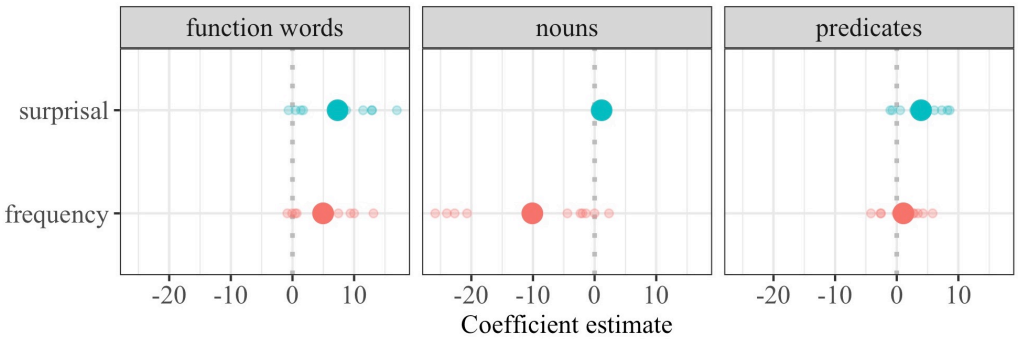


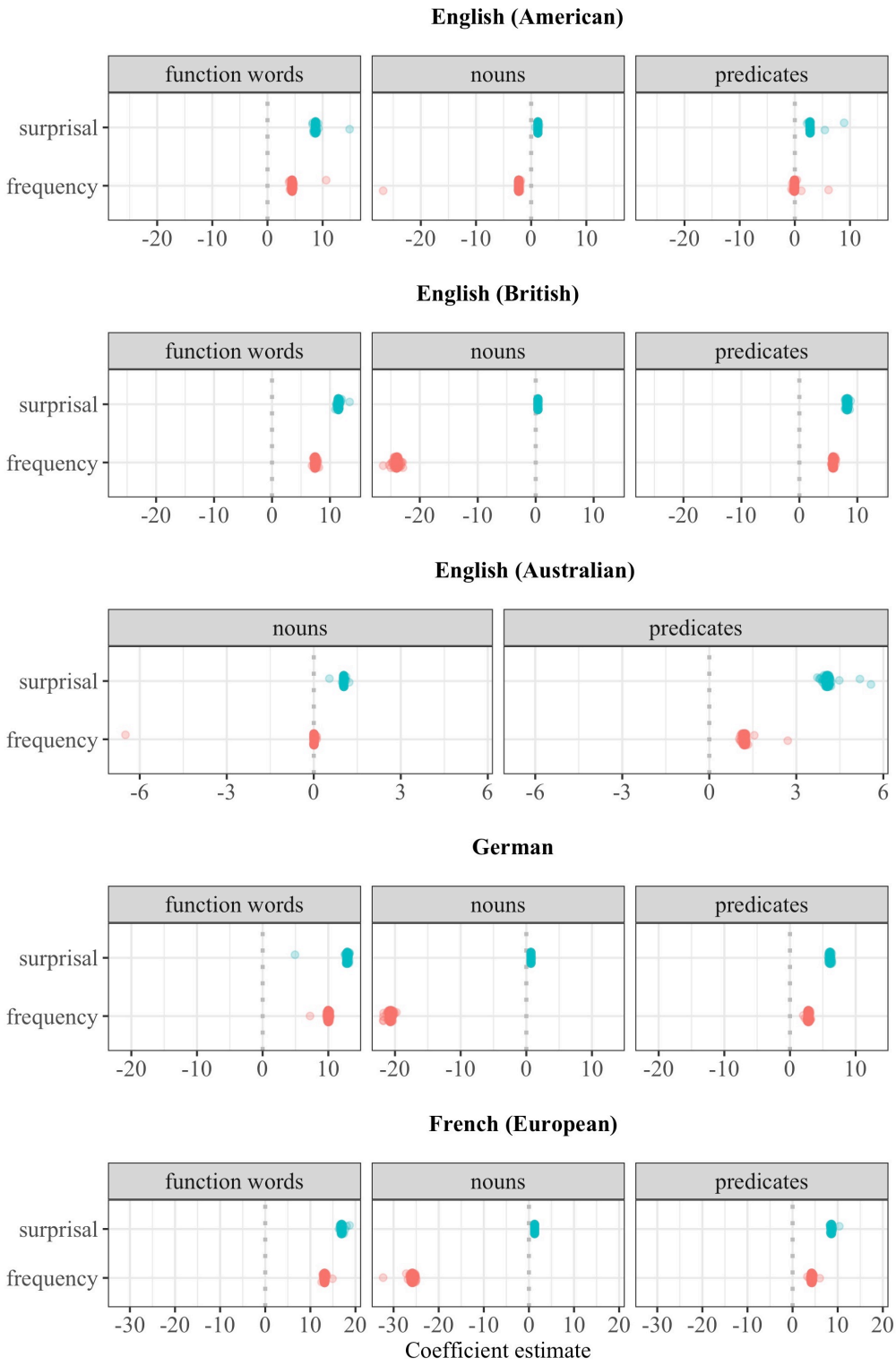
Figure 2.2: Mean coefficient estimate by lexical category across languages for experiment 1

in relation to AoA in the next experiment.

Clearly, given these results, word predictability is important to predicting the order in which words are acquired. The second question asked in this section was about the relation between average surprisal and AoA across lexical categories. As we might expect, words with higher average surprisal – or lower predictability – tend to be acquired later and therefore have higher AoA. I explore this relation in further detail in the next subsection.

2.4.4 The relation between predictability and AoA

I consider the relation between predictability and AoA in each language by looking at the coefficient estimates for average surprisal and comparing them to the coefficients for frequency in the fitted augmented models from this experiment. Since the model definition includes an interaction term with lexical category, I plot the estimates by lexical category. Figure 2.2 plots the mean overall estimates across languages, where each point is the mean coefficient in one language and the large point is the overall mean, showing the general trend across all the data. Figure 2.3 shows each language individually in more detail in individual graphs. 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 560 items and therefore 560 folds all of which have been plotted here. Thus, if a coefficient estimate varied very little across all folds and is very certain then points appear densely on top of one another, but if there is more uncertainty and variation across folds, then the points are more scattered. If estimates fall along the zero dotted line or cross it, then it is safe to say that they are no different from zero and do not have a real effect on AoA prediction. Since all predictors are scaled by centering their mean at zero



(a) Part 1: Coefficient estimates by lexical category in each language for experiment 1

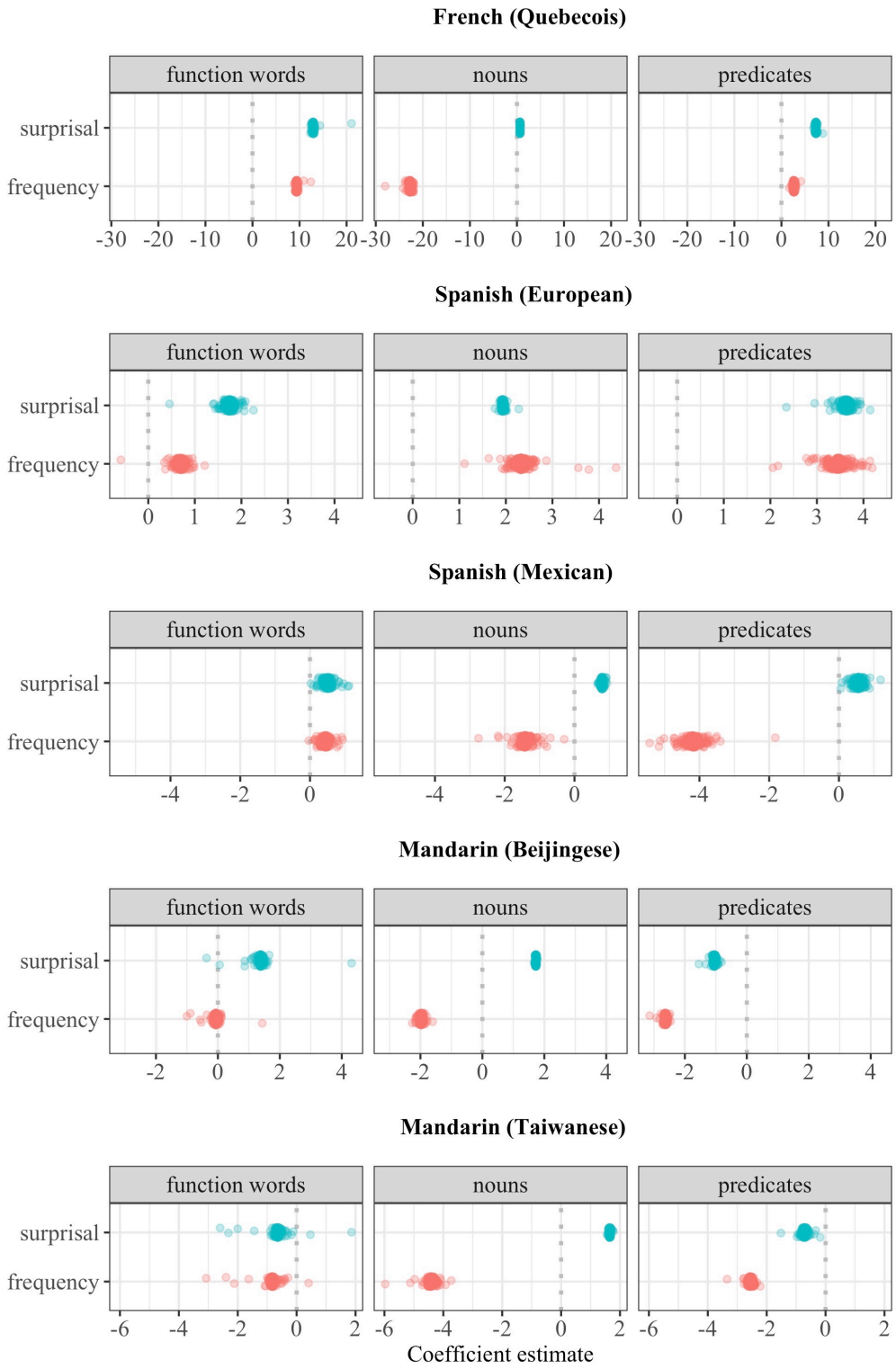


Figure 2.3: Coefficient estimates by lexical category in each language for experiment 1

and setting their standard deviation to one, it is also possible to compare the magnitude of estimated coefficients across predictors.

First, I consider the case of average surprisal. Figure 2.2 shows the overall mean effect of surprisal across all languages. As we can see, The overall trend seems to be that average surprisal shows a positive effect for both function words and predicates, but no or little effect when predicting the AoA of nouns. This trend reproduces my earlier findings (Portelance et al., 2020) which were only for *English (American)*, and indeed, if we look more closely at the results in each language in 2.3, we see this exact pattern *English (American)* shows this exact pattern. It means that in the case of function words and predicates, words that are harder to our models to predict given their previous linguistic context are also harder for children and acquired later, but this is not always the case for nouns. As 2.3 shows, the exact same pattern can be found in *English (British)*, *English (Australian)*, *German*, *French (European)*, and *French (Quebecois)*; the only exceptions are in the cases of Spanish and Mandarin.

Given that adding average surprisal as a predictor beyond frequency did not significantly help explain any more variance in our data for *Spanish (Mexican)* and that the correlation between frequency and average surprisal was high for both Spanish dialects, it is not surprising that we should see much noisier coeffient estimates fits across cross-validation folds in these dialects. These facts may also explain why they seem to pattern differently. Like the rest of the languages, *Spanish (European)* function words and predicates with higher average surprisal are learnt later and have higher AoA. However, unlike other languages this effect also seems to hold true for nouns. In *Spanish (Mexican)*, the estimated coefficient for average surprisal is no different from zero for either function words or predicates, and is just above zero for nouns.

The patterns for Mandarin are an interesting case. Average surprisal does not seem to be a good predictor of AoA for function words in this language. Instead, it seems to be a better predictor of the AoA of nouns, where less predicable nouns in context are learnt later. It also still seems to be a predictor of AoA for predicates though less strongly. However, interestingly, Mandarin seems to have the opposite pattern to all the other languages in their case: predicates with higher average surprisal, or less predictability, seem to be learnt earlier and have a lower AoA. Mandarin has been known to pattern differently to other language 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 initial learn to

produce more nouns before increasing their productions of verbs – that English learners seem to (Tardif, Gelman, & Xu, 1999), an observation that has since been reproduced using the Wordbank parental report data (Frank et al., 2019; Yee, 2020). Instead, Mandarin learners have been found to produce more predicates early on during learning, even both English and Mandarin speaking parents produce relatively more predicates than nouns (Tardif et al., 1999). Another possible explanation for this difference may however lie in some of the Wordbank data itself. As Frank et al. (chapter 11; 2019) note, there seem to be some discrepancies with some of the Mandarin data in the repository, specifically forms collected for *Mandarin (Beijingese)* from the Tardif, Fletcher, Liang, & Kaciroti, 2009 study seem to show a much stronger predicate bias than other forms available for *Mandarin (Beijingese)* or other languages. This data imbalance may also be contributing to this effect for predicates. Now, 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 (Beijingese)*. So, though children's input may be similarly distributed in terms of lexical category productions, which in turn should lead to similar average word surprisal distributions in children's input, they do not necessarily always learn words in the same order. Here, we find that Mandarin learners seems to differ from other Indo-European language learners. These observations suggests that there may be more important cultural factors that affect the order in which words are acquired beyond linguistic context.

Second, let's consider the case of frequency. Across all languages tested there is a negative relation between frequency and AoA in the case of nouns, in other words, more frequent nouns are learnt earlier, this overall effect is visible in figure 2.2. There are two exceptions to this rule are *English (Australian)*, where the frequency coeffient for nouns is no different from zero, and *Spanish (European)*, where again, the pattern seems to differ, see 2.3. Interestingly, unlike nouns, we actually see an opposite effect in the case of function words and, for some languages, predicates (figure 2.2), where more frequent words in these categories are learnt later. Possible explanations for these effects are that very frequent function words tend to be more polysemous or have referential meaning, making learning their sometimes ambiguous meaning more difficult. Additionally, in many languages, more common predicates are often also predicates with irregular grammatical forms, since their ubiquity can stave off language evolution towards more regular forms; this irregularity may make them harder to learn to produce. Some exceptions to this pattern between frequency and AoA for function words and predicates are *English (American)*, where there is no

effect for predicates, *Spanish (Mexican)*, where there is no effect for function words and a negative one for predicates, and Mandarin, where like with average surprisal we see the opposite pattern for predicates and no effect for function words. The reasons for Mandarin's opposite from the pattern of other language in the case of predicates for both frequency and average surprisal may be the same.

Overall, the relation between the predictability of words in their linguistic context and their AoA seems to be pretty stable across languages, mattering mostly for the acquisition of function words and predicates, both lexical categories which depend on linguistic context to express meaning. Mandarin is a notable exception to this rule, however, since average surprisal did not seem to matter for the acquisition of function words, but did so for nouns, and though it also mattered for predicates, this language showed the opposite effect to other languages in their case.

2.5 Experiment 2 : What context matters ?

In this second experiment, my goal is to understand how much previous linguistic context seems to matter for predicting the AoA of words. I explore this question with the sets of model comparisons. First, I compare different context sizes using the average surprisal of n-gram models, specifically, I compare predictive models of word AoA using uni-grams, bi-gram, tri-gram, or four-gram average surprisal as predictors. Like in experiment 1, I do so using LOO cross-validation. Second, using the best n-gram context size found in the first model comparison, I consider whether adding more dynamic context sizes using LSTM average surprisal helps predict AoA beyond this initial n-gram surprisal. I do so by adding residualised LSTM average surprisals to my predictive model. For this second part I compare predictive models with and without residualized LSTM average surprisal as a predictor using both LOO cross-validation and nested model comparison. Finally, I consider whether the benefits of this additional predictor are once again seen more in some lexical categories versus others.

2.5.1 Predictors

In this experiment I compare the effect of average surprisal from different language models, representing the predictability of words conditioned on different sizes and types of previous linguistic context. Like in the previous experiments, all predictors are scaled by centering

their mean at zero and dividing by their standard deviation.

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. In subsection 2.5.3, I compare the average surprisal of uni-gram, 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'), I used the mean average surprisal across all forms. Since uni-gram surprisal is included in the model comparison, which is equivalent to $-\log(frequency)$, there is no need to include frequency as an additional predictor.

Residualised LSTM surprisal represents the residual variance left after fitting a linear model which predicts LSTM average surprisal as a function of n-gram average surprisal, `LSTM average surprisal` $\sim$ `n-gram average surprisal`, where the LSTM average surprisal values are those described in experiment 1. In other words, this predictor represents the average surprisal of LSTM language models which is not already explained by the average surprisal of an n-gram model.

Concreteness and lexical category interactions are also included in all of the regression models used in this experiment. They are defined in experiment 1.

2.5.2 Regression models

As previously mentioned, the first part of this experiment in subsection 2.5.3 involves comparing regression models which use the average surprisal of words given different n-gram language models, to assess how much previous linguistic context is important for predicting AoA. Thus, I compare the following models:

1. **The null model :** `AoA` ~ 1 , to ensure that all models are in fact different from null.
2. **The uni-gram model :** `AoA` $\sim$ `lexical category` $*$`(uni-gram average surprisal + concreteness)` , which considers word predictability based on no previous context.
3. **The bi-gram model :** `AoA` $\sim$ `lexical category` $*$`(bi-gram average surprisal + concreteness)` , which considers word predictability based on one previous word of context.

4. **The tri-gram model :** `AoA` $\sim$ `lexical category` $*$`(tri-gram average surprisal` $+$ `concreteness)` , which considers word predictability based on two previous words of context.

5. **The four-gram model :** `AoA` $\sim$ `lexical category` $*$`(four-gram average surprisal` $+$ `concreteness)` , which considers word predictability based three words of previous context.

The second part of this experiment in subsection 2.5.4 involves a nested model comparison between the best of these regression models and an augmented model that additionally contains residualized average surprisal from LSTM language models. This second comparison allows us to assess whether adding more dynamic context sizes using LSTMs beyond fixed n-gram context sizes helps predict AoA. Here, we add the following regression model, where the predictor `n-gram average surprisal` refers to the best n-gram language model average surprisal values from the set above.

The augmented model : `AoA` $\sim$ `lexical category` $*$`(residualized LSTM surprisal` $+$ `n-gram average surprisal` $+$ `concreteness)`

For the first model comparison, I report MAD across LOO model fits in each language in subsection 2.5.3. In the second model comparison reported in subsection 2.5.4, I present both MAD across LOO model fits and also run a nested model comparison between the base and the augmented models fitted on all of the data in each language. For the nested model comparison I report the F-statistic and p-value.

2.5.3 A comparison of context size using n-gram models

The results for the first model comparison are presented in table 2.5. They show the average model fits across LOO cross-validation for regression models containing average surprisal of different n-gram language models as predictors.

There is a consistent pattern across all languages: using less context to determine word predictability is better overall. Thus, the models with the smallest MAD in all cases are those which use uni-gram average surprisal as a predictor. These are then followed by those using bi-grams, then tri-grams, and finally four-grams. As a reminder, uni-gram average surprisals are determined using no previous context and are in fact just the average negative log frequencies of words. Though this result may seem surprising at first glance, when we

Table 2.5: Model comparison results using n-gram surprisal by language

	LOO MAD$_{[95\% \text{ CI}]}$				
Language	**uni-gram**	**bi-gram**	**tri-gram**	**four-gram**	**null**
English (American)	**2.01**$_{[0.68,3.34]}$	2.09$_{[0.69,3.49]}$	2.22$_{[0.78,3.67]}$	2.35$_{[0.84,3.86]}$	2.82$_{[1.04,4.6]}$
English (British)	**2.19**$_{[0.76,3.62]}$	2.25$_{[0.81,3.71]}$	2.33$_{[0.87,3.79]}$	2.42$_{[0.91,3.94]}$	2.73$_{[1.02,4.45]}$
English (Australian)	**1.94**$_{[0.6,3.27]}$	2.0$_{[0.66,3.35]}$	2.10$_{[0.72,3.47]}$	2.19$_{[0.75,3.62]}$	2.48$_{[0.94,4.02]}$
German	**2.2**$_{[0.57,3.82]}$	2.3$_{[0.60,4.0]}$	2.41$_{[0.7,4.12}$	2.42$_{[0.71,4.13]}$	2.75$_{[0.91,4.6]}$
French (European)	**2.32**$_{[0.67,3.98]}$	2.47$_{[0.75,4.19]}$	2.66$_{[0.81,4.51]}$	2.78$_{[0.91,4.65]}$	3.1$_{[1.13,5.07]}$
French (Quebecois)	**2.56**$_{[0.87,4.24]}$	2.64$_{[0.85,4.43]}$	2.75$_{[0.91,4.59]}$	2.84$_{[0.99,4.7]}$	3.11$_{[1.15,5.06]}$
Spanish (European)	**2.53**$_{[0.91,4.14]}$	2.59$_{[0.94,4.24]}$	2.67$_{[0.98,4.36]}$	2.72$_{[1.03,4.4]}$	2.76$_{[1.02,4.51]}$
Spanish (Mexican)	**2.09**$_{[0.79,3.39]}$	2.11$_{[0.8,3.42]}$	2.12$_{[0.8,3.44]}$	2.13$_{[0.82,3.42]}$	2.22$_{[0.81,3.64]}$
Mandarin (Beijingese)	**1.88**$_{[0.64,3.12]}$	1.98$_{[0.7,3.26]}$	2.08$_{[0.73,3.43]}$	2.13$_{[0.74,3.51]}$	2.33$_{[0.9,3.77]}$
Mandarin (Taiwanese)	**3.01**$_{[1.12,4.9]}$	3.18$_{[1.21,5.14]}$	3.3$_{[1.32,5.27]}$	3.34$_{[1.38,5.3]}$	3.46$_{[1.52,5.41]}$

consider the composition of the word lists for which we have AoA in all languages, as well as our results from experiment 1, the result makes sense. The large majority of items across all languages are nouns and, as we saw in experiment 1, average surprisal isn't such a good predictor of the AoA of nouns, but frequency is. Thus, if only choosing a single context size, no context and uni-gram surprisal will necessarily dominate as the best choice when most of our items are best predicted from their frequency. In the next subsections, I consider whether adding information from dynamic context sizes using LSTM average surprisals beyond uni-gram average surprisals helps predict AoA overall and for specific types of words.

2.5.4 Including more dynamic context sizes using LSTM models

Here, I compare the best regression models from the previous subsection, the uni-gram models, to an augmented version of these models which additionally contains residualized LSTM average surprisal as a predictor. I analyse the difference between the base uni-gram models and the augmented models using both LOO cross-validation and an ANOVA nested model comparison across languages. The results are available in table 2.6.

The nested ANOVA results suggest that adding residualized LSTM average surprisal as a predictor significantly increases model fit in three of the languages, but based on the cross-validation results, I would like to suggest that these significance results should be

Table 2.6: Model comparison results by language using LSTM average surprisal beyond uni-gram surprisal

	LOO MAD$_{[95\% \text{ CI}]}$		**Nested ANOVA**	
Language	**uni-gram model**	**Augmented model**	$F_{(\text{Dfs})}$ **statistic**	p **value**
English (American)	$2.01_{[0.13,3.9]}$	$2.02_{[0.1,3.94]}$	$1.32_{(3,554)}$	n.s.
English (British)	$2.19_{[0.16,4.22]}$	$2.17_{[0.16,4.18]}$	$3.33_{(3,333)}$	<0.05*
English (Australian)	$1.94_{[0.05,3.82]}$	$1.93_{[0.01,3.86]}$	$2.23_{(2,371)}$	n.s.
German	$2.2_{[0.0,4.5]}$	$2.3_{[0.0,5.35]}$	$1.37_{(3,350)}$	n.s.
French (European)	$2.32_{[0.0,4.67]}$	$2.34_{[0.0,4.69]}$	$0.29_{(3,420)}$	n.s.
French (Quebecois)	$2.56_{[0.18,4.94]}$	$2.57_{[0.15,4.98]}$	$2.01_{(3,473)}$	n.s.
Spanish (European)	$2.53_{[0.24,4.81]}$	$2.53_{[0.24,4.82]}$	$1.3_{(3,461)}$	n.s.
Spanish (Mexican)	$2.09_{[0.25,3.93]}$	$2.06_{[0.24,3.89]}$	$3.95_{(3,331)}$	<0.01**
Mandarin (Beijingese)	$1.88_{[0.12,3.63]}$	$1.89_{[0.09,3.69]}$	$1.68_{(3,505)}$	n.s.
Mandarin (Taiwanese)	$3.01_{[0.33,5.68]}$	$3.0_{[0.39,5.61]}$	$4.42_{(3,407)}$	<0.01**

Table 2.7: Correlation between LSTM average surprisal and uni-gram average surprisal in each language

Language	**Pearson** r
English (American)	0.65
English (British)	0.67
English (Australian)	0.72
German	0.63
French (European)	0.65
French (Quebecois)	0.69
Spanish (European)	0.97
Spanish (Mexican)	0.97
Mandarin (Beijingese)	0.96
Mandarin (Taiwanese)	0.96

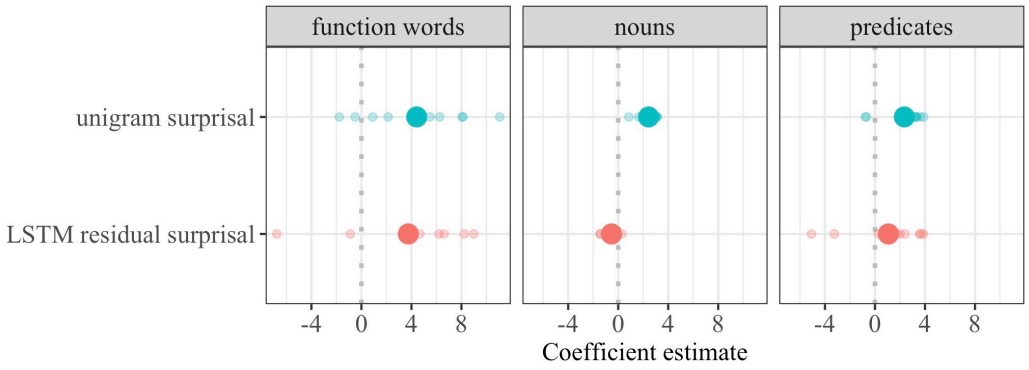
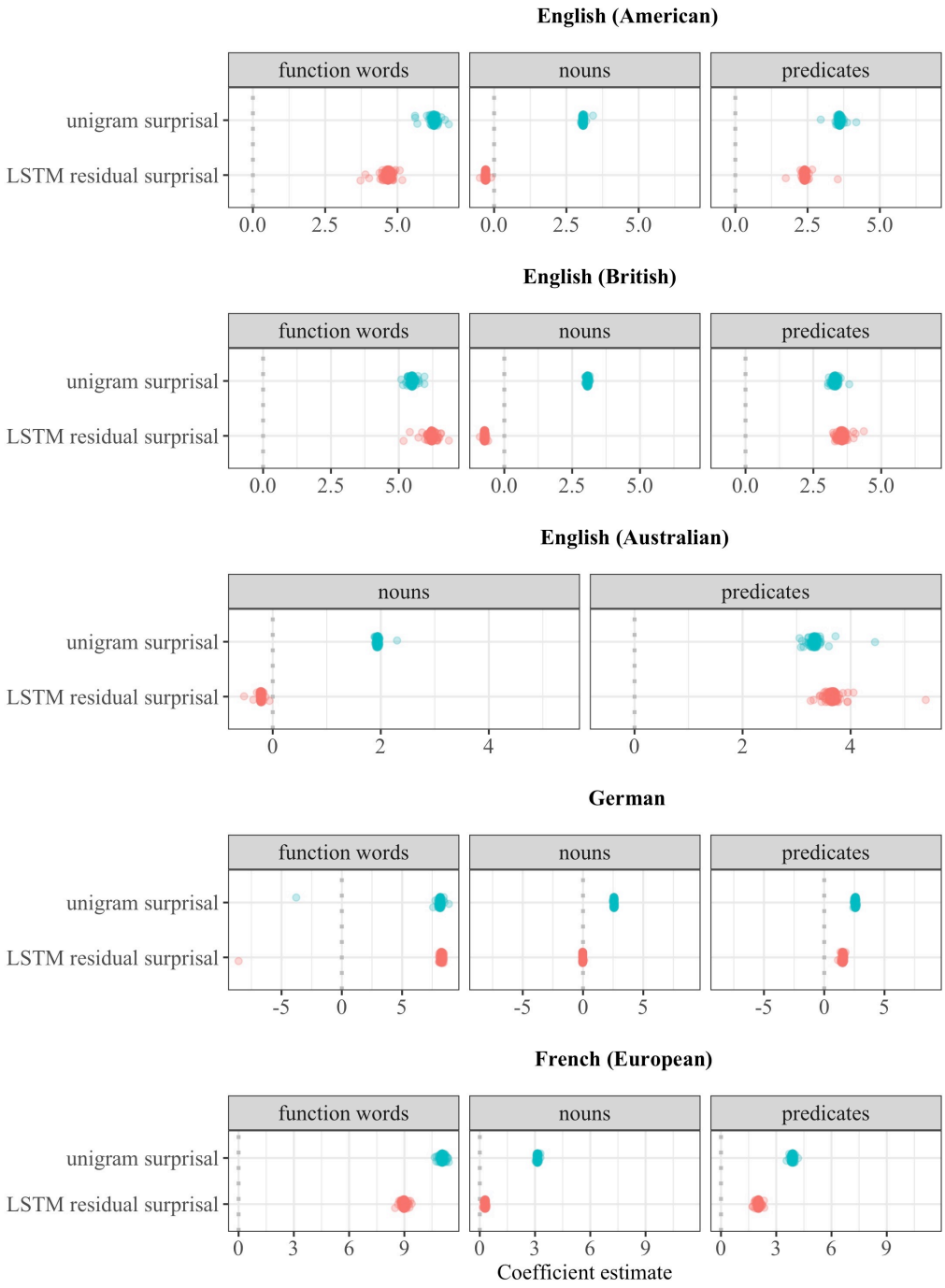


Figure 2.4: Mean coefficient estimate by lexical category across languages for experiment 2

ignored since the difference is very small and not that different from other languages or dialects where no such significant results were found. Thus, overall adding more dynamic average surprisal values beyond uni-gram ones does not make a noticeable difference in AoA predictability. Again, most of the items are nouns and experiment 1 results suggest that LSTM average surprisal best predicts AoA for predicates and function words. Additionally, If we consider the correlations between LSTM average surprisals and uni-gram average surprisals across languages in table 2.7, we can see that they are high. LSTMs in many languages seem to mostly track uni-gram frequencies to determine the probability of many words, ignoring much of the preceding context. There are some exceptions however. In the next subsection I look at the items for which using LSTM average surprisal beyond uni-gram average surprisal truly does seem to make a difference.

2.5.5 The relation between lexical category and context

Though adding residualized LSTM average surprisal beyond uni-gram surprisal as a predictor of AoA did not make a significant difference overall, it can make a difference for predicting the AoA of specific words, and those words once again tend to be predicates and function words. If we look at the mean coefficient estimates by lexical category across languages for the augmented model in figure 2.4, we see that residualized LSTM surprisal generally has no effect for nouns, but still has a positive effect for function words and predicates in most languages, showing much of the same pattern as experiment 1. Looking more closely at the estimates for each language in figure 2.5, we see that the exception to this rule is once again Mandarin, where in this case both function words and predicates show



(a) Part 1: Coefficient estimates by lexical category in each language for experiment 2

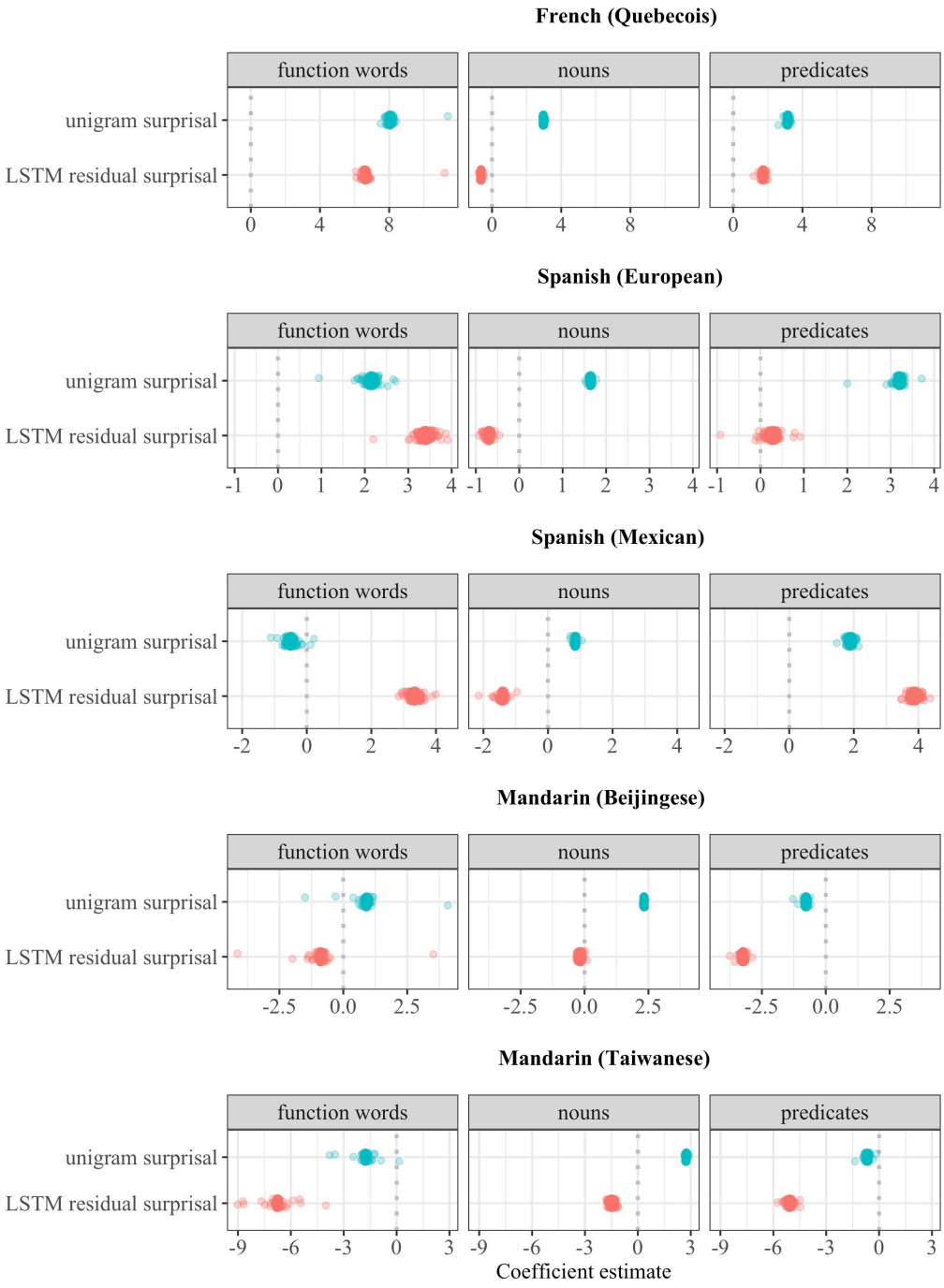


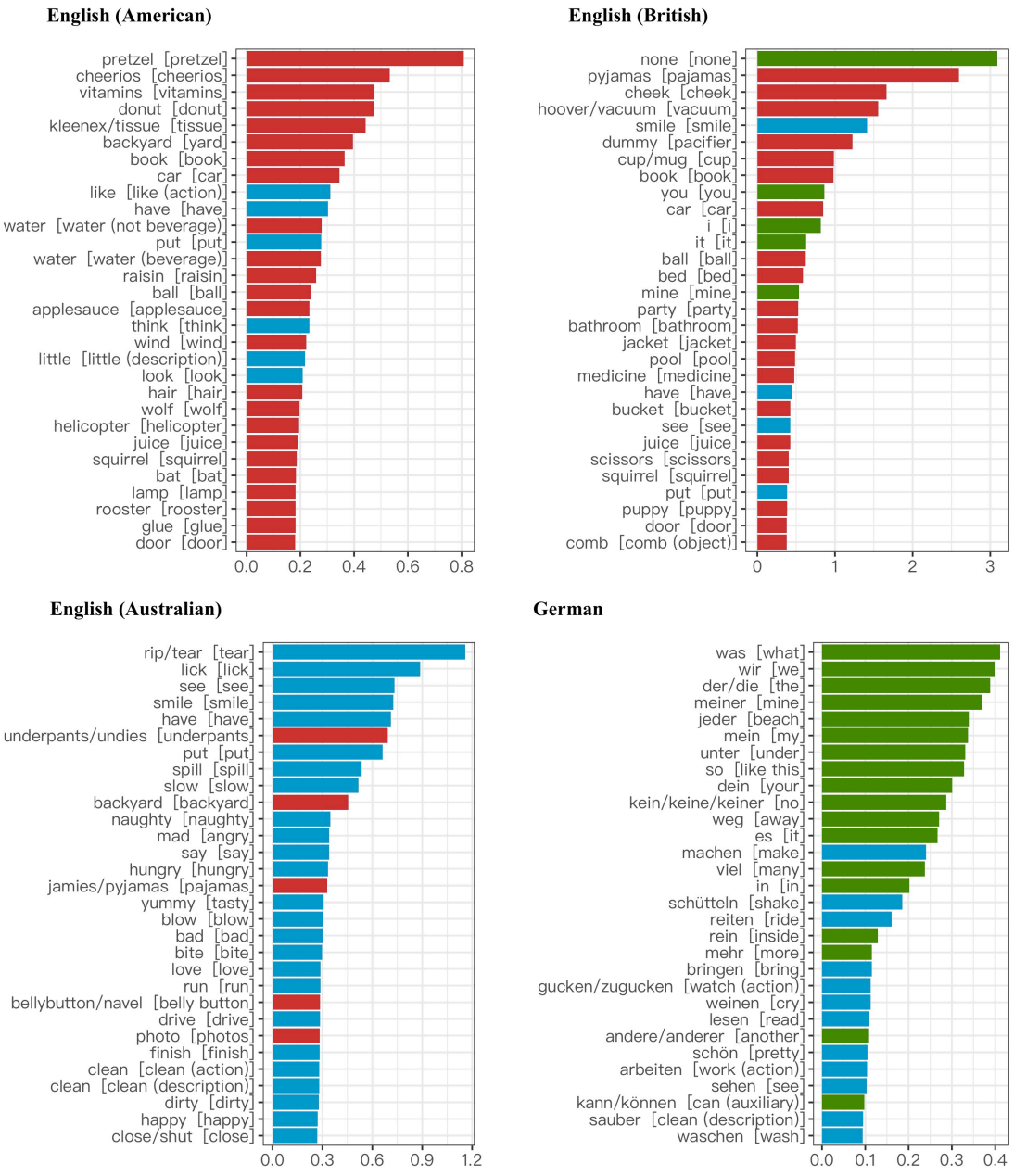
Figure 2.5: Coefficient estimates by lexical category in each language for experiment 2

a negative effect for LSTM residual surprisal as a predictor. As for uni-gram surprisal, it has a positive effect in all categories across almost all language – words that are more predictable based on their frequency are generally learnt earlier –, here again the exception being Mandarin function words and predicates.

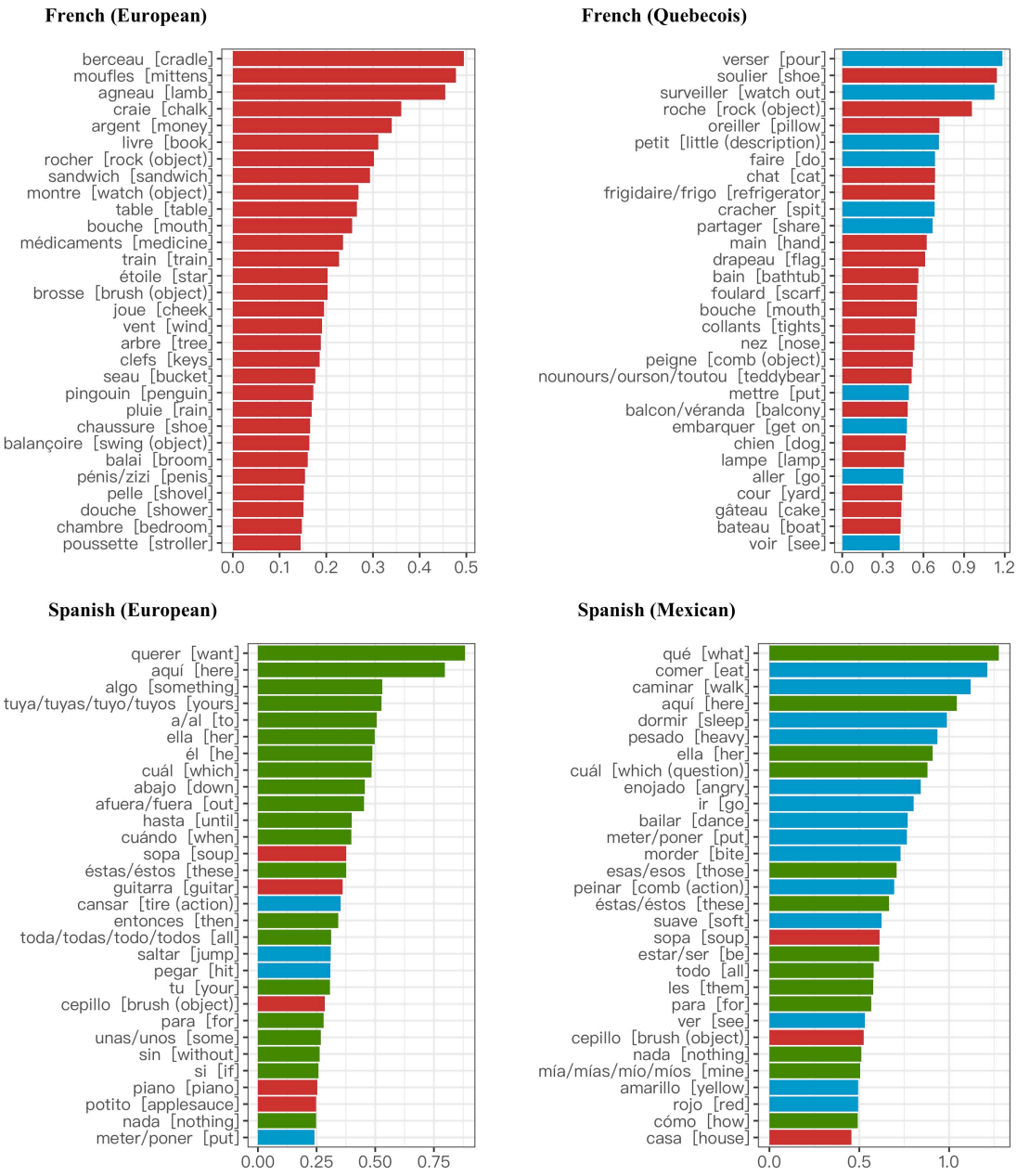
We can also look at the specific items where adding residualized LSTM surprisal beyond uni-gram surprisal makes the biggest positive difference and verify whether they also indeed tend to be function words and predicates. For each language, I list the top 30 items for which adding residualized LSTM surprisal decreased the absolute deviation (in months) the most between the predicted AoA and the actual estimated AoA based on Wordbank, see figure 2.6 over three pages. If there are multiple word alternatives listed for a given item they are separated by a forward slash. I also include the 'unilemma' of these words in brackets – the equivalent English concept listed in Wordbank – to help with crosslinguistic interpretation. Words are color coded by lexical category: nouns are red, predicates are blue, and function words are green.

In six of the ten languages the items whose absolute deviation decreases the most by adding residualized LSTM average surprisal beyond uni-gram average surprisal are almost exclusively predicates in function words. This observation is true of *English (Australian)*, *German*, *Spanish (European)*, *Spanish (Mexican)*, *Mandarin (Beijingese)*, and *Mandarin (Taiwanese)*. In *English (British)* and *French (Quebecois)* show no particular advantage for any lexical category in their top 30 words. The notable exceptions are *English (American)* and *French (European)* where most of the items are nouns. Unlike n-grams, LSTM average surprisal also takes into account semantic information about words because their probability in context is also conditional on the lexical embeddings these language models learn to represent these words. Notably, these are the languages where LSTM average surprisal was the least correlated with uni-gram surprisal (table 2.7), supporting the possibility that these LSTM surprisal values are determined using additional type of semantic information, which in turn may help predict AoA for specific nouns as well.

These results suggest that though adding more dynamic context sizes using residualized LSTM average surprisal does not help predict AoA for all words, it can better predictions of AoA for specific words, especially some function words and predicates across languages. Overall, however, most nouns may be best predicted by simply using uni-gram surprisals, or in other words log transformed frequencies.



(a) Part 1 : Difference for the top 30 words in each language between the absolute deviation of the uni-gram models and the augmented models. Nouns are red, predicates are blue, and function words are green.



(b) Part 2: Difference for the top 30 words in each language between the absolute deviation of the uni-gram models and the augmented models. Nouns are red, predicates are blue, and function words are green.

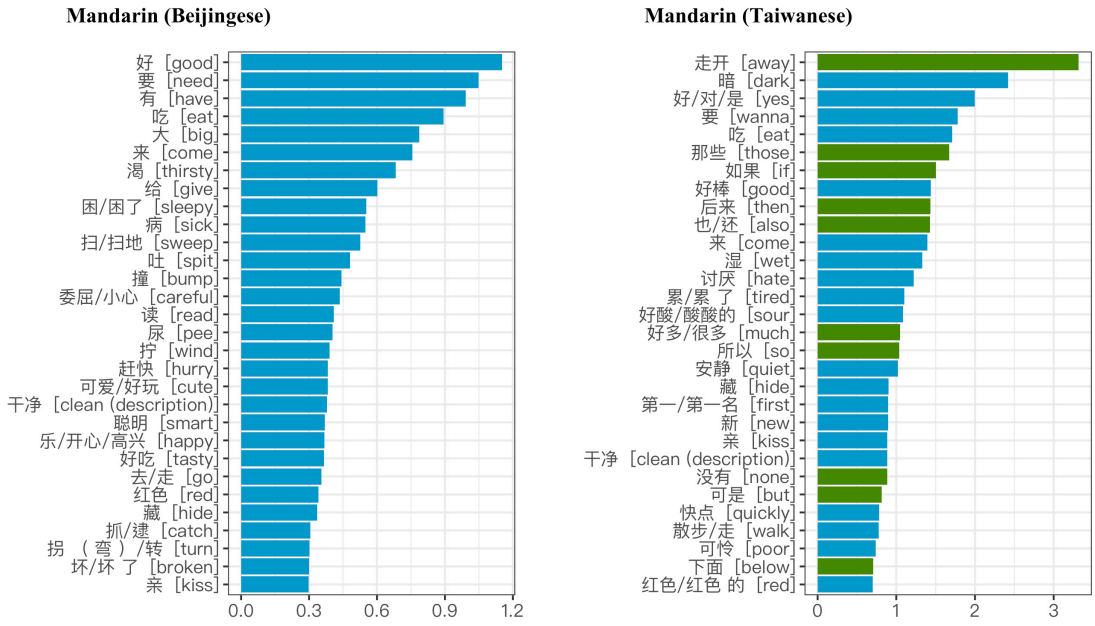


Figure 2.6: Difference for the top 30 words in each language between the absolute deviation of the uni-gram models and the augmented models. Nouns are red, predicates are blue, and function words are green.

2.6 Discussion

I started this chapter with two questions : (1) does the predictability of words in children's linguistic input help explain when they are acquired by children beyond previously studied variables like frequency? and (2) are words which are difficult for language models to learn also acquired later by children? To answer these questions I considered whether the surprisal of words for language models can predicted the AoA of these same words, thus considering whether the learning outcomes of language models as probability models showed similar ordering effects to the word learning outcomes of children. Based on my experiments, I can now formulate an answer to both of these questions.

First, word predictability is a very good predictor of the AoA of words across all the languages tested. That being said, I found that for most words, very little if no prior linguistic context was necessary to predict their AoA; for the items contained in each on the languages dataset, simply using uni-gram average surprisal is best. In other words, using log transformed frequencies is definitely better than non-log transformed ones. This is not a new predictor; Goodman et al., 2008, Braginsky et al., 2019, and Kuperman et al.,

2012 all used log transformed frequencies to reduce the long tail distribution that frequency counts often have. Considering word predictability given slightly more previous linguistic context from children's input however can help predict the AoA of specific words. In most languages, adding additional information about context using LSTM average surprisal beyond uni-gram average surprisal helps predict the AoA of some predicates and function words, much more than nouns. This may be because function words and predicates derive much of their meaning from linguistic context, while nouns are often referential, drawing meaning from real world objects and animate things. The nouns contained in the item lists from Wordbank used in these experiments are almost all rated as being highly concrete with a mean score of 4.81 (SD = 0.24) out of 5 on the scale used here from Brysbaert et al., 2014, while predicates have a mean concreteness score of 3.46 (SD = 0.85) and function words of 2.47 (SD = 0.88). Thus, it is safe to say that almost all the nouns considered do have concrete referents in the real world. Additionally, given that there is a clear correlation between an item's lexical category and its concreteness score in our data, it is difficult to differentiate whether function words and predicates benefit more from having LSTM average surprisal as a predictor of their AoA because of their lexical category or because of they are less concrete, since these concepts are in this case somewhat interchangeable. It is clear however that learning nouns may require relying on other types of contextual information beyond previous linguistic context, like the visual and social contexts in which children are learning language, while learning predicates and function words that are not as concrete may require relying on more linguistic context in addition to these other modalities.

Second, though there are some similarities between the word learning difficulties of LSTM language models and the order in which children acquire words, there are also some clear differences. Since the average surprisal of a word for an LSTM is in fact a measure of how difficult a word is to learn for the language model, we can consider how well LSTM average surprisal predicts the AoA of words in children as a study of the relation between word learnability for language models and word learning ordering effects in children. In experiment 1, I found that LSTM average surprisal was a good predictor of the AoA of function words and predicates in almost all languages. In almost all cases LSTM average surprisal as a predictor of AoA had a positive estimate for these items, meaning that function words and predicates which where difficult for the language model to learn were also learnt later in children. A notable counterexample was Mandarin predicates which showed a negative relation, such that Mandarin predicates with higher average surprisal

seemed to actually be learnt earlier in children. I discussed some possible explanations for this difference, including that Mandarin learners have previously shown other word learning order differences from other language learners. Importantly, I did not find a strong relationship between LSTM average surprisal and the AoA of nouns. These results suggest that though LSTM word learning shows similarities to children's word learning in many languages for words whose meaning often depends on linguistic context, like predicates and function words, there are definite differences especially when learning more concrete words like the nouns in our datasets. Additionally, the results from experiment 2 suggest that the LSTM language models I trained on CHILDES data most often relied on very little previous context to predict words, since for many words LSTM average surprisal was highly correlated with uni-gram surprisal (LSTM average surprisal is most highly correlated with uni-gram average surprisal, then bi-gram, and tri-gram and so on). I also found that uni-gram average surprisal was a much better predictor of the AoA of words than other n-gram average surprisal that considered more previous linguistic context, like bi-grams or tri-grams. Thus, both the LSTMs trained on child-directed utterances and children seem to favor little or no previous context to learn words. This result also supports the recent finding that LSTMs and other neural language model architecture seem to initially learn to rely on uni-gram frequencies, before learning a bit more about bi-grams and eventually a little more nuanced representations of previous context to predict words, as a function of how much data they have seen (Chang & Bergen, To appear).

Word predictability is clearly important for word learning, both for children and, by definition, language models as well. In most cases, early word learning seems to however only require keeping track of and conditioning predictability on very little previously linguistic context. Though considering more nuanced and dynamic previous linguistic context sizes may help both children and LSTM language models learn function words and predicates, the similarities between the word learning trajectories of these models and human learners do not carry over to more concrete words like nouns. Children likely dependent much less on linguistic context and more on visual context as well as social context to learn these words, two types of input contexts that language models like LSTMs do not a priori have access to during learning. In the next chapter, using VQA models, I address how the addition of visual context to models' input may affect word learning, specifically function word learning.

Appendix A

Additional experimentation and development from chapter 2

A.1 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. I used an Adam optimizer across all models during training (Kingma & Ba, 2014) and simple cross-entropy across the outputs of all states as my loss function.

A.1.1 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**]

References

Agrawal, A., Batra, D., & Parikh, D. (2016). Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1955–1960). Retrieved from `https://www.aclweb.org/anthology/D16-1203`

Andreas, J., & Klein, D. (2016). Reasoning about Pragmatics with Neural Listeners and Speakers. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1173–1182). Retrieved from `https://www.aclweb.org/anthology/D16-1125`

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual Question Answering. In *Proceedings of the ieee international conference on computer vision* (pp. 2425–2433). Retrieved from `https://arxiv.org/pdf/1505.00468.pdf`

Bates, E., Dale, P. S., Thal, D., et al. (1995). Individual differences and their implications for theories of language development. *The handbook of child language*, *30*, 96–151.

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.

Bengio, Y., Simard, P., Frasconi, P., et al. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166.

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.

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.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, *33*(2), 111–153.

Chang, T. A., & Bergen, B. K. (To appear). 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.

Cimpian, A., & Markman, E. M. (2005). The absence of a shape bias in children's word learning. *Developmental Psychology*, *41*, 1003–1019.

Clark, E. V. (1977). Strategies and the mapping problem in first language acquisition. *Language Learning and Thought*, 147–168.

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

Clark, E. V. (1993b). The mapping problem. In *The lexicon in acquisition* (p. 43–66). Cambridge University Press.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural Computation*, *1*(3), 372–381.

Dagan, G., Hupkes, D., & Bruni, E. (2020). Co-evolution of language and agents in referential games. *arXiv preprint arXiv:2001.03361*. Retrieved from `https://arxiv.org/abs/2001.03361`

Dale, P. S., Bates, E., Reznick, J. S., & Morisset, C. (1989). The validity of a parent report instrument of child language at twenty months. *Journal of child language*, *16*(2), 239–249.

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. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*(2-3), 195–225.

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*.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Foerster, J., Assael, I. A., de Freitas, N., & Whiteson, S. (2016). Learning to Communicate with Deep Multi-Agent Reinforcement Learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29, pp. 2137–2145). Retrieved from `https://proceedings.neurips.cc/paper/2016/file/c7635bfd99248a2cdef8249ef7bfbef4-Paper.pdf`

Frank, M. C., Braginsky, M., Marchman, V., & Yurovsky, D. (2019). *Variability and Consistency in Early Language Learning: The Wordbank Project.* (https://langcog.github.io/wordbank-book/index.html)

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.

Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., & Xu, W. (2015). Are you talking to a machine? Dataset and methods for multilingual image question answering. In *Proceedings of the conference on neural information processing systems.* Retrieved from `https://proceedings.neurips.cc/paper/2015/file/fb508ef074ee78a0e58c68be06d8a2eb-Paper.pdf`

Gers, F. A. (1999). Learning to forget: Continual prediction with lstm. In *9th international conference on artificial neural networks* (pp. 850–855).

Geurts, B., Katsos, N., Cummins, C., Moons, J., & Noordman, L. (2010). Scalar quantifiers: Logic, acquisition, and processing. *Language and cognitive processes*, *25*, 130–148.

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.

Grigoroglou, M., Johanson, M., & Papafragou, A. (2019). Pragmatics and spatial language: The acquisition of front and back. *Developmental psychology*, *55*, 729 – 744.

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

Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., ... others (2017). Grounded language learning in a simulated 3d world. *arXiv preprint arXiv:1706.06551*. Retrieved from `https://arxiv.org/abs/1706.06551`

Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., & Santoro, A. (2020). Environmental drivers of systematicity and generalization in a situated agent. In *International conference on learning representations.* Retrieved from `https://openreview.net/forum?id=SklGryBtwr`

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.

Hochmann, J.-R., Carey, S., & Mehler, J. (2018). Infants learn a rule predicated on the relation same but fail to simultaneously learn a rule predicated on the relation different. *Cognition*, *177*, 49–57.

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

Hosseini, H., Xiao, B., Jaiswal, M., & Poovendran, R. (2018). Assessing shape bias property of convolutional neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 1923–1931). Retrieved from `http://labs.ece.uw.edu/nsl/papers/Assessing_Shape_Bias.pdf`

Hudson, D. A., & Manning, C. D. (2019). Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 6700–6709). Retrieved from `https://arxiv.org/pdf/1902.09506.pdf`

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).

Jasbi, M., Jaggi, A., & Frank, M. C. (2018). Conceptual and prosodic cues in child-directed speech can help children learn the meaning of disjunction. In *Cogsci.* Retrieved from `https://cogsci.mindmodeling.org/2018/papers/0121/0121.pdf`

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2901–2910). Retrieved from `https://arxiv.org/abs/1612.06890`

Johnston, J. R. (1984). Acquisition of locative meanings: Behind and in front of. *Journal of Child Language*, *11*, 407–422.

Johnston, J. R., & Slobin, D. I. (1979). The development of locative expressions in english, italian, serbo-croatian and turkish. *Journal of child language*, *6*, 529–545.

Jones, S. S., Smith, L. B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child development*, *62*(3), 499–516.

Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*, 664–676. Retrieved from `https://ieeexplore.ieee.org/document/7534740`

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`

Kiros, R., Salakhutdinov, R., & Zemel, R. (2014a). Multimodal neural language models. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st international conference on machine learning* (Vol. 32, pp. 595–603). Retrieved from `http://proceedings.mlr.press/v32/kiros14.html`

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014b). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... Fei-Fei, L. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations.. Retrieved from `https://arxiv.org/abs/1602.07332`

Kuczaj, S. A., & Maratsos, M. P. (1975). On the acquisition of" front, back", and" side". *Child Development*, 202–210.

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

Landau, B., Smith, L. B., & Jones, S. (1992). Syntactic context and the shape bias in children's and adults' lexical learning. *Journal of Memory and Language*, *31*, 807–825.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*, 299–321.

Lazaridou, A., Hermann, K. M., Tuyls, K., & Clark, S. (2018). Emergence of Linguistic Communication from Referential Games with Symbolic and Pixel Input. In *Proceedings of the 6th international conference on learning representations, ICLR*. Retrieved from `https://openreview.net/forum?id=HJGv1Z-AW`

Lazaridou, A., Peysakhovich, A., & Baroni, M. (2017). Multi-Agent Cooperation and the

Emergence of (Natural) Language. In *5th international conference on learning representations, ICLR 2017, toulon, france, april 24-26, 2017, conference track proceedings.* Retrieved from `https://openreview.net/forum?id=Hk8N3Sclg`

Lazaridou, A., Pham, N. T., & Baroni, M. (2016). Towards multi-agent communication-based language learning. *arXiv preprint arXiv:1605.07133*. Retrieved from `https://arxiv.org/abs/1605.07133`

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.

Macnamara, J. (1982). *Names for things: A study of human learning.* Mit Press.

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

Malinowski, M., & Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proceedings of the conference on neural information processing systems.* Retrieved from `https://papers.nips.cc/paper/2014/file/d516b13671a4179d9b7b458a6ebdeb92-Paper.pdf`

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive science*, *14*(1), 57–77.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, *20*(2), 121–157.

Morris, B. J. (2008). Logically speaking: Evidence for item-based acquisition of the connectives AND & OR. *Journal of Cognition and Development*, *9*(1), 67–88.

Norris, D. (1990). A dynamic-net model of human speech recognition. In G. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and cognitive perspectives.* The MIT Press.

Norris, D. (1993). Bottom-up connectionist models of 'interaction'. In *Cognitive models of speech processing: The second sperlonga meeting* (pp. 211–234).

Paris, S. G. (1973). Comprehension of language connectives and propositional logical relationships. *Journal of experimental child psychology*, *16*(2), 278–291.

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

Quine, W. V. O. (1960). *Word and object mit press.* MIT Press.

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism*. MIT Press.

Ren, M., Kiros, R., & Zemel, R. (2015). Exploring models and data for image question answering. In *Proceedings of the conference on neural information processing systems.* Retrieved from `https://proceedings.neurips.cc/paper/2015/file/831c2f88a604a07ca94314b56a4921b8-Paper.pdf`

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. In *International conference on machine learning* (pp. 2940–2949). Retrieved from `https://arxiv.org/abs/1706.08606`

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

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.

Sukhbaatar, S., szlam, a., & Fergus, R. (2016). Learning multiagent communication with backpropagation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems.* Retrieved from `https://proceedings.neurips.cc/paper/2016/file/55b1927fdafef39c48e5b73b5d61ea60-Paper.pdf`

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.

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).

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *2015 ieee conference on computer vision and pattern recognition (cvpr)* (pp. 3156–3164). Retrieved from `https://ieeexplore.ieee.org/document/7298935`

Ware, E. A., & Booth, A. E. (2010). Form follows function: Learning about function helps children learn about shape. *Cognitive Development*, *25*(2), 124–137.

Washington, D. S., & Naremore, R. C. (1978). Children's use of spatial prepositions in two-and three-dimensional tasks. *Journal of Speech and Hearing Research*, *21*(1), 151–165.

Windmiller, M. (1973). *The relationship between a child's conception of space and his comprehension and production of spatial locatives* (Unpublished doctoral dissertation). University of California, Berkeley.

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).

Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., & Parikh, D. (2016). Yin and yang: Balancing and answering binary visual questions. In *2016 ieee conference on computer vision and pattern recognition (cvpr)* (p. 5014-5022). Retrieved from `https://ieeexplore.ieee.org/document/7780911`