UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Exploring the variable effects of frequency and semantic diversity as predictors for a word's ease of acquisition in different word classes

Permalink

https://escholarship.org/uc/item/83t6n1rq

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

ISSN

1069-7977

Authors

Siow, Serene
Plunkett, Kim

Publication Date

2021

Peer reviewed

eScholarship.org

Powered by the California Digital Library
University of California

Exploring the variable effects of frequency and semantic diversity as predictors for a word's ease of acquisition in different word classes

Serene Siow (serene.siow@psy.ox.ac.uk)
Department of Experimental Psychology, Anna Watts Building,
Oxford, OX2 6GG United Kingdom

Kim Plunkett (kim.plunkett@psy.ox.ac.uk)
Department of Experimental Psychology, Anna Watts Building,
Oxford, OX2 6GG United Kingdom

Abstract

Infant vocabulary development is inevitably dependent on the speech they hear in their environment. This paper reports an investigation of the vocabulary statistics that predict a word's age of acquisition, focusing on frequency and contextual diversity as derived from child-directed speech data along with associative norms generated by adults. Age of acquisition is operationalised using parental-report of infant word knowledge in a British English-speaking population. The work can be considered an extension of Hills, Maouene, Riordan, and Smith (2010) using a fully British English dataset. We found significant effects of both word frequency and word associations on age of acquisition. Interestingly, the strength of these predictors differed between word classes, with frequency being the strongest predictor for nouns and associations the strongest predictor for function words.

Keywords: vocabulary acquisition; child-directed speech; word frequency; semantic associations; contextual diversity

Introduction

Children learn language via experience and exposure. Therefore, unsurprisingly, analyses of child-generated productions reveal that the amount and variety of words used by children correlates highly with the words used by their mothers (Li & Fang, 2011). Both child-directed speech and child-produced speech are predominantly composed of concrete nouns and common pronouns, which have obvious referents. Hills (2013) found that child-directed speech possesses higher rates of word repetitions, less diversity and more semantic associations between adjacent words than adult-directed speech.

Quantifying word acquisition

A common methodology for measuring infant word knowledge is via parental-report vocabulary questionnaires. At the group level, a word's *Age of Acquisition* (AoA) is typically operationalised as the youngest age at which more than 50% of sampled children are reported to understand the word (in the case of AoA measured in comprehension) or produce it (AoA in production). This value of 50% is in truth arbitrary, however this simple threshold is sufficient for our purposes. For the analyses explored in this paper (and in fact much of the literature), we are less concerned about the exact value of AoA, and more about the *order* of acquisition. We use the quantified statistic of AoA to compare words against each other. This gives us a proxy of a word's ease of acquisition relative to other words. In all further discussions of AoA in this paper, we will be referring to this measure of relative acquisition ease.

Frequency and AoA

What factors make a word easier to learn? Children have to deal with complexities in both the linguistic and physical environment. One feature of this complexity is that children are not necessarily oriented towards the correct referent of the word at the time of naming. Naming events are just as likely to be highly unambiguous (with referent in dominant visual position) as highly ambiguous (referent not present or unclear) from the child's visual perspective (Yurovsky, Smith, & Yu, 2013). Repetitions and high frequency of words may make it easier to map the correct referent to the object, because they increase the frequency of the word co-occurring with its referent, and particularly occurrences where the child's attention is on the referent. Many studies have found correlations between a word's AoA and its frequency of occurrence in the language environment (Goodman, Dale, & Li, 2008; Hills et al., 2010; Hills, 2013; Braginsky, Yurovsky, Marchman, & Frank, 2019).

Quantifying frequency in CDS Transcripts of Child Directed Speech (CDS) can be used to obtain quantitative measures of children's early learning environments. A word's frequency is quantified by counting the total number of times that word appears in the corpus. Frequency may be counted for word forms (i.e., separate counts for "run", "ran", "running", etc.) or for lemmas (i.e., combining all the above instances under the concept "run"). It is important to note that frequency as derived from corpus data can only give us direct data on word form occurrences. However, word learning doesn't only involve learning a word form, but also the mapping of the word form to its referent. Studies of frequency using corpus data make the implicit assumption that words are typically uttered in the presence of their referent, allowing for word-referent mapping. Given the extensive literature on contingent responding and joint attention in caregiver-child interactions, this assumption seems a reasonable one. Parents preferentially name and refer to objects that their child is looking at. Joint attention on the referent by both parent and child during word utterances has been found to be predictive of successful learning (Tomasello & Todd, 1983).

Contextual diversity and AoA

The hypothesised effect of *contextual diversity* on word learning is rooted in the cross-situational learning literature. Infants are able to extract statistical information about a word-referent pairing by rapidly compiling information across multiple occurrences and evaluating the regularities within the input (Smith & Yu, 2008). Several explanations have been proposed regarding the mechanism underlying cross-situational learning. One possibility is that learners operate using associative learning mechanisms (McMurray, Horst, & Samuelson, 2012). Co-occurrences between word and referent may facilitate resolution of referential ambiguity. Alternatively, learners may engage in Bayesian hypothesis testing, where learners assign probabilities to each object in an ambiguous scene, representing the likelihood that a given object is the target word's referent (Xu & Tenenbaum, 2007). These probabilities are derived and adjusted using evidence over multiple scenes. Probabilities can be increased through systematic occurrences of a word-object pair, and decreased by the absence of a given possible referent during utterance of the word. The object with the highest probability can be inferred to be the intended referent, subject to collecting sufficient evidence.

Given the same number of occurrences, word-referent pairs are learnt more successfully when a target word-referent pair co-occurs in a wide variety of contexts than when it appears with only one or two unique word-referent pairs (Kachergis, Shiffrin, & Yu, 2009; Suanda, Mugwanya, & Namy, 2014). By occurring in a variety of contexts in the joint presence of different objects, word-object pairs can be assigned different weightings (whether by associative learning or hypothesis testing) depending on systematic co-occurrences or non-co-occurrences, thus facilitating inference of word meaning.

Quantifying contextual diversity in CDS Besides frequency, corpora of CDS transcripts can also be used to identify the words that a given word co-occurs with. The number of unique words a given word co-occurs with in CDS can be used to quantify the diversity of the linguistic environment (and by proxy the physical environment) that a given word occurs in. Hills et al. (2010) referred to this quantified variable as contextual diversity, forming a link with the cross-situational learning literature. As with corpus-derived frequency, this measure of contextual diversity also relies on several assumptions: firstly, that words occur in the presence of their referents, and secondly that the diversity of word co-occurrences corresponds reliably with variation in the physical environment, so as to give cues for inference of meaning. Hills et al. found that a window of 5 words was optimal when using child-directed speech to predict individual words' AoA. This measure of contextual diversity should include mostly syntagmatic word associations denoting sentence structure (e.g. "apple" – "eat"), though it can also include some paradigmatic associations of words belonging to the same category (e.g. "apple" – "pear"), as found by Wettler, Rapp, and Sedlmeier (2005) for adult speech. Chang and Deák (2020) showed that co-occurrences in child-directed speech include both syntactic and thematic relationships, and that both types of co-occurrence contribute unique variance to predicting AoA even after accounting for word frequency.

Semantic associations and AoA

Free associations have been used as a measure for semantic association strength between word pairs and also to quantify semantic richness surrounding a word. Adult-generated associative norms have been found to correspond with contiguity in adult speech (Wettler et al., 2005). Adult-generated associations were found to be predictive of words' AoA in infants (Hills et al., 2010; Hills, Maouene, Maouene, Sheya, & Smith, 2009). While these statistics have been widely supported as good predictors of performance in adult lexical decision tasks, much less it known how they might relate to AoA. Hills et al. (2010) proposed that the relationship between associations and AoA of early-learnt words may be partially accounted for by contextual diversity in everyday language.

When attempting to unpack what association norms may represent in an infant word acquisition context, we need to consider the sampling methodology that word association databases are built on. Unlike speech transcripts which record co-occurrences in everyday speech, word association data is collected experimentally by asking participants to list the first word(s) that come to mind. The South Florida Association Norms (Nelson, McEvoy, & Schreiber, 2004) used by Hills et al. (2010) is an example of data collected using a discrete word association task, where participants could only give one response to each cue. Discrete word association tasks have been linked to more reliable indices of association strength and set size, but under-representation of weaker associates as compared to tasks that allow participants to give more than one answer (Nelson, McEvoy, & Dennis, 2000). Importantly, word associations have directionality that reflects their associative relationship and ease of retrieval. The cue "turtle" may elicit the response "animal", but the cue "animal" may preferentially elicit more frequent words like "dog" from within the large set of competitors within the animal category. Hills et al. (2009) found that *associative indegree* (which only counts the instances when the target word is given as a response) was a better predictor of AoA than outdegree (when the target word is given as a cue) or overall degree (adding both together). Hills et al. proposed the *preferential acquisition model* for infant word acquisition, where words that are the most well-connected to other words in the learning environment via shared semantic relationships are most easily learnt. They suggested that this effect may stem from well-connected words being more salient within the learning environment and also that the richness of shared semantic context with related words can help inference of meaning.

The Present Study

The present study aims to investigate the word statistics of word frequency, contextual diversity in CDS and associative indegree as predictors for AoA, extending Hills et al. (2010)'s study with British English data.

High word frequency is predicted to facilitate word acquisition through repeated exposures, with high frequency words having earlier AoA. Contextual diversity and associative indegree are measures of the richness of the learning environment that a word occurs in, proposed to support resolution of referential ambiguity. The justification of including both predictors lies in composition differences as a result of their respective data sources. While contextual diversity as derived from natural speech largely represents syntagmatic relationships in CDS (including both syntactic and thematic features), associative indegree would include a higher proportion of paradigmatic relationships. Of course, these are not mutually exclusive, and some paradigmatic links may also occur in the contextual diversity measure, and vice versa.

From Hills et al. (2010)'s findings, we predict that there will be differences between word classes in the variance explained by these predictors. Frequency is expected to be a good predictor for all word classes, following the extensive literature linking word frequency and learning across several different word classes (Kachergis et al., 2009; Naigles & Hoff-Ginsberg, 1998; Hochmann, Endress, & Mehler, 2010). For contextual diversity, given its theoretical background in the domain of cross-situational learning, we expect that the effect of contextual diversity will be strongest in nouns (a word class with concrete word-referent mappings). We expect this predictor to have a smaller but still significant effect on the other more abstract word classes. And lastly, for associative diversity, we expect that it will explain unique variance for all word classes. However, in the case of nouns, most of the effect may be absorbed by contextual diversity given their close relationship. The effect of associative diversity was found by Hills et al. to be strongest for function words.

Methods

Age of acquisition

Full word-by-child comprehension and production scores were obtained from monolingual English data collected using the Oxford Communicative Development Inventory (CDI) (Hamilton, Plunkett, & Schafer, 2000). The Oxford CDI contains 418 words commonly known to young children. The inventory collects data via parental report for both comprehension and production of these words.

Three unpublished datasets were combined. The first dataset was collected between March 2020 and December 2020 using a 418-word Oxford CDI with random category presentation order (CDI2020), N = 180, age range 12–32 months old. The second dataset was made up of data collected between 2013 and 2020 using the 553-word extended version of the Oxford CDI (CDIExt), N = 330, age range 12–32 months old. In this version of the CDI, the 418 words of the standard-length CDI analysed in this paper were always presented as the first 418 words in static presentation order. Analysis showed no significant difference between vocabulary scores collected with randomised-category presentation (CDI2020) or static presentation order (CDIExt) after controlling for the child's age (t = 0.267, p = .789). Finally, this data was combined with open-source data available on Wordbank (Frank, Braginsky, Yurovsky, & Marchman, 2017), collected by the Plymouth BabyLab using the 418-word Oxford CDI, N = 1210, age range 12–25 months old. CDI data was divided into one-month chunks by child's age, representing completed months. Sample sizes in each age group ranged between 11 and 222. This data was used to calculate AoA in both comprehension and production for each word in the CDI. A word's AoA in comprehension was operationalised as the lowest age (in months) when the word reaches the threshold of being understood by at least 50% of toddlers at that age. Similarly, AoA in production was defined as the lowest age it is spoken by 50% toddlers. There was a correlation of .86 between comprehension and production AoA (Pearson's r).

Four word classes were included in the analysis – nouns (N = 211), verbs (N = 65), adjectives (N = 36) and function words (N = 36). As in Hills et al. (2010), we excluded words about time (a very small class of 8 words), sounds, games and routines (an ambiguous word class, many of which are not single words), 8 words duplicated in multiple categories (e.g. noun "drink" and verb "drink") and an additional 6 words that are not single words (e.g. belly button).

Adult-generated associations

Adult-generated word pair associations were obtained from the Small World of Words (SWOW-EN) dataset (De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019), a database of free association responses for 12,292 cues. The SWOW-EN dataset was collected between 2011 and 2018. Out of the sample, the majority were native speakers of American English (81%), while 13% of participants reported British English as their first language. For comparability with the University of South Florida Free Association Norms (Nelson et al., 2004) used in Hills et al. (2010), we included only data corresponding to the first response given. *Associative indegree* for a given word is defined as the total number of unique cues that elicit that word as a response. To avoid conflating this score with idiosyncratic responses, we only included words that were given by at least 2 participants for a given cue. This measure therefore represents the diversity of a word's strongest semantic associations. The associative indegree of CDI words calculated from the whole sample had a .96 correlation with the British subset.

Corpora of naturalistic CDS

Corpora of CDS transcripts were downloaded from the CHILDES online database (MacWhinney, 2000) via the *childesr* package (Braginsky, Sanchez, & Yurovsky, 2020). Data was extracted from the British English subset of the corpora. Only corpora with naturalistic parent-child interactions were included in the final sample. To achieve this, we excluded any corpora where the investigator played an active role (e.g., interviews, defined as >10% of total utterances classified as being produced by the investigator) and school recordings (defined as >10% total utterances by teacher). Three corpora were excluded using these criteria. An additional 2 corpora were excluded for only having single-word utterances from the target child with no caregiver corresponding utterances, suggesting that they may be transcripts from an experimental task. Additionally, one more corpus was excluded for having transcripts that contained high numbers of non-English utterances.

The final set consisted of data from 67 children from 9 corpora (Forrester, Howe, Korman, Lara, Manchester, Nuffield, Thomas, Tommerdahl, Wells). This included recordings of everyday interactions (N = 2) and free-play sessions (N = 7). Each child's transcript provided between 103 and 162518 utterances (including both child-directed and child-produced utterances), totalling 979251 utterances. In the compiled dataset, 50.9% of total utterances were by the target child's mother, 39.4% were by the target child, and other individuals (father, investigator, grandparents and siblings) made up the remaining 9.7%. Common English contractions (e.g "it's") were expanded to their full form (i.e. "it is"). All words were then lemmatised using the *textstem* package (Rinker, 2018). Additionally, if a given word had abbreviations (e.g. airplane / plane) or synonyms (e.g. bunny / rabbit) that are accepted as tokens of the same item in the CDI, the frequency and co-occurrence statistics of these word tokens were summed.

CHILDES word frequency *Word frequency* was obtained from the CHILDES data above by counting the number of occurrences of a given lemma, and log-transformed into zipf values (Zipf, 1949). To check if child-produced utterances differed greatly from child-directed utterances, we separated CHILDES transcripts into utterances produced by the target child and utterances produced by other speakers. Frequency scores extracted from these two sources had a correlation of .94. As the correlation was very high, we decided not to analyse these sources separately.

CHILDES co-occurrence degree (contextual diversity) Co-occurrence degree was also extracted from the CHILDES data. Hills et al. (2010) found that a relatively small sliding window of 5 words was the best predictor for AoA when building a measure of contextual diversity from word co-occurrence, reflecting the limitations of working memory in the developing brain. We therefore decided to use a 5-word window in this British extension. *Contextual diversity* for a given word was computed by summing the total number of unique words that the word co-occurred with in the dataset within a 5-word window.

Results

We investigated the contribution of the three factors (frequency, contextual diversity, associative indegree) in predicting AoA of individual words. Frequency was transformed to zipf values, and both contextual diversity and associative indegree were log-transformed. All mentions below of the predictors will refer to these transformed variables. To avoid the effects of influential outliers, words that were more than 3 SD from the mean of their word class in any of the three predictors were excluded from the analysis. This removed 4 nouns, 1 verb and 2 function words.

We ran linear regressions with AoA as the dependent variable using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) for all words together and also for each word class. Notably, frequency and diversity were highly correlated (.99, Pearson's r). In all cases, frequency was a better predictor than diversity, so diversity was dropped from the models to avoid collinearity. R^2 for the models with frequency and associative indegree are listed in Table 1 for comprehension and Table 2 for production. Model fit is visualised in Figure 1 for comprehension and Figure 2 for production using the *ggiraphExtra* R package (Moon, 2021).

Table 1: R^2 of models predicting AoA (comprehension) for each word class, for frequency alone (Freq), associative indegree alone (AI) and the increase in R^2 from adding AI as predictor after accounting for Freq.

	df	Freq	AI	ΔR^2 AI
All	341	.027**	.010*	.0008
Nouns	205	.293***	.018*	.027**
Verbs	62	−.014	−.006	
Adjectives	34	.022	.0008	
Function	32	−.025	.300***	.304***

p<.05*, p<.01**, p<.001***

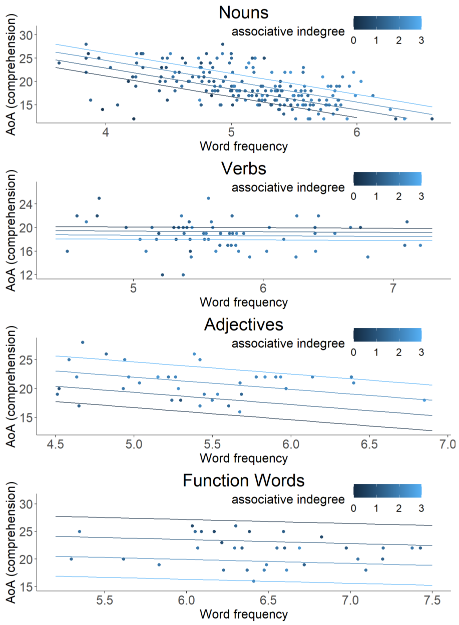

Figure 1: Scatterplot showing model fit, with AoA in comprehension (months) on x-axis, word frequency (zipf) on y-axis and associative indegree (log) as colour, split by word class.

Table 2: R^2 of models predicting AoA (production) for each word class, for frequency alone (Freq), associative indegree alone (AI) and the increase in R^2 from adding AI as predictor after accounting for Freq.

	df	Freq	AI	ΔR^2 AI
All	341	.043***	.025**	.008
Nouns	205	.404***	.027**	.039***
Verbs	62	−.013	−.008	
Adjectives	34	.152*	.039	
Function	32	.010	.410***	.415***

p<.05*, p<.01**, p<.001***

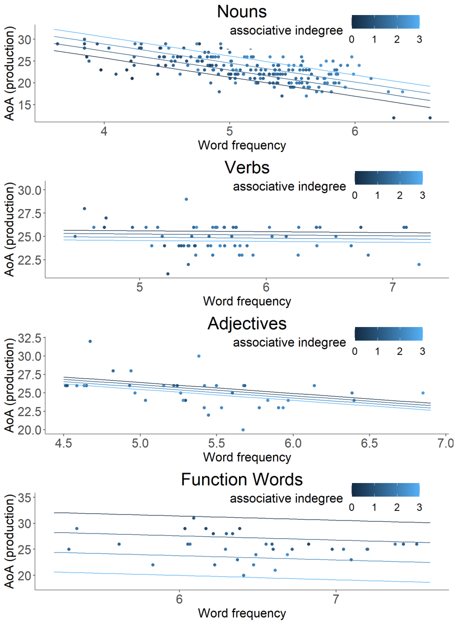

Figure 2: Scatterplot showing model fit, with AoA in production (months) on x-axis, word frequency (zipf) on y-axis and associative indegree (log) as colour, split by word class.

Discussion

This study contributes to the literature by replicating the procedures used by Hills et al. (2010) (Study 1) with an independent sample of British English toddlers and a different dataset of adult associations. We replicated their findings of significant but small R^2 for frequency and associative indegree in predicting AoA when looking across all words together. Stronger R^2 were found within word classes, together with notable variability depending on the word class and predictor, suggesting that different word classes rely on variable mechanisms for acquisition.

Additionally, unlike the CDI used by Hills et al. which only collects production scores, the Oxford CDI includes both comprehension and production scores. This allowed us to investigate whether frequency and associative indegree predict comprehension and production differently using data from the same children, controlling for potential confounds of developmental change. Aside from the relationship between frequency and AoA for adjectives that reached significance for production but not comprehension, the pattern of results was very similar between comprehension and production in our sample.

Frequency and AoA

When looking at each word class separately, nouns were the only word class where both predictors were significantly correlated to AoA. However, when both predictors were entered in a regression together, frequency accounted for a dominant amount of variance by itself. This supports the extensive literature that repeated exposure to words (and by proxy word-object mappings) predicts more successful word learning. The more times a word-object pair occurs, the easier it is to learn. The especially strong effect of frequency on the AoA of nouns relative to other word classes is a common finding in the literature (Goodman et al., 2008; Hills et al., 2010; Braginsky et al., 2019).

Unexpectedly, frequency was very bad at predicting acquisition order for verbs in our dataset. This poor predictive power of frequency is surprising considering findings in the literature of frequency predicting verb acquisition (Naigles & Hoff-Ginsberg, 1998). It also contrasts with Hills et al. (2010)'s findings of significant predictors. It is possible that the profile of the dataset may have masked any effect of frequency. The 64 verbs included in the analyses had a narrower spread of AoA in comprehension (mean = 18.78, IQR = 3) as compared to the 207 nouns (mean = 18.79, IQR = 5.5). Further exploration with datasets including a wider distribution of verbs is required to identify whether this poor effect of frequency is meaningful or simply a limitation of the data.

The literature is more mixed for function words. Infants have been shown to be sensitive to frequency as a cue to differentiate function words from other word classes (Hochmann et al., 2010). However, when investigating AoA within the word class, Braginsky et al. (2019) found that children's knowledge of function words was predicted better by word length and sentence complexity. This was in contrast to nouns and predicates which were strongly related to frequency. Braginsky et al. proposed that low sentence complexity could allow learners to decode a function word's meaning more easily. Hidaka (2013) proposed a computational model for the acquisition of function words that is insensitive to frequency, instead suggesting that function words are learnt through cognitive processes related to inference. This is again contrasted with the acquisition patterns of nouns, verbs and adjectives, which Hidaka found were best predicted by a model of cumulative learning. The divide between function words and other word classes presented in the abovementioned studies is consistent with our findings reported in this paper.

Semantic associations and AoA

Both nouns and function words showed a significant relationship between associative indegree and AoA, with the predictor explaining significant variance even after accounting for frequency. This effect was small for nouns, while associative indegree was by far the best predictor for function words out of the available options.

Like Hills et al. (2010), we found a small correlation between associative indegree and contextual diversity as derived from CHILDES when looking at all words together (.25, Pearson's *r*). As suggested by Hills et al., diverse semantic contexts could make a word more salient in the learning environment and support the disambiguation of meaning. There were stronger correlations within word classes (Nouns .56, Verbs .30, Adjectives .48) with the striking exception of function words ($-.003$). This suggests that while contextual diversity may explain part of the relationship between associative indegree and AoA of nouns, there is something different at play with function words.

We suggested in the introduction of this paper that associative indegree represented the richness of a word's semantic environment with a focus on paradigmatic relationships. An initial exploration of the composition of associations by word class supports this theory. On average, 37.1% of cues that elicited our sampled function words were other function words (28.9% cues were nouns). In comparison, for our sampled nouns, on average 69.5% of the cues were other nouns (only 3.4% were function words). Most frequent cue-response types include antonyms like "you"–"me", "this"–"that"; synonyms like "beneath"–"under", "additional"–"more"; and phrases like "thank"–"you", "upside"–"down". Unlike concrete nouns which have obvious referents, the learning of function words is dependent on the interpretation of its meaning from the context it appears in. Rich semantic associations may facilitate easier inference of a function word's meaning. Further work needs to be done to study the underlying mechanisms represented by the associative indegree measure, and whether paradigmatic and syntagmatic relationships between words contribute differently to word acquisition and inference of meaning.

Future directions

Disentangling collinearity of frequency and contextual diversity A strong correlation between frequency and contextual diversity as derived from CDS was found both across all words and within word classes. Shaoul and Westbury (2006) addressed the issue that very high frequency words tend to have higher scores in word co-occurrence measures simply through chance co-occurrences, in the context of the hyperspace analog to language (HAL) model. This issue is particularly relevant for corpus studies of infant language, where the earliest learnt words are often highly frequent. However, applying a simple standardisation procedure as in Hills (2013) where the contextual diversity score for a word was divided by the word's frequency did little to reduce this collinearity in our data. The resulting correlation between frequency and the standardised score (both log-transformed) was still very high at $-.985$. More exploration is needed to identify a suitable standardisation method to quantify contextual diversity without the influence of chance co-occurrences in higher frequency words. This will allow us to study whether contextual diversity plays an independent role in facilitating word learning outside of frequency effects.

Additionally, considering Chang and Deák (2020)'s findings that both adjacent and non-adjacent co-occurrences contribute unique variance to predicting AoA, a weighted co-occurrence system as applied by HAL and Shaoul and Westbury (2006)'s HiDEx models may allow us to explore the effect of contextual diversity in CDS at a more fine-grained level than the binary co-occurrence matrix used by Hills et al. (2010) and also the present study.

Evaluating the assumptions of corpus-derived statistics As mentioned in the introduction, the use of corpus-derived frequency and contextual diversity statistics makes strong assumptions about the linguistic environment systematically reflecting the physical learning environment. Studies that collect both video and audio data on children's day-to-day activities can provide us with crucial information on the physical environment that accompanies everyday speech. Systematic exploration of the rich data offered by such studies can help us reach past the abstractness of purely linguistic data to identify what these derived statistics actually represent in a real-world learning environment.

Conclusion

This study extends prior research on frequency, contextual diversity and associative diversity effects on monolingual AoA of early vocabulary. Frequency was the best predictor for nouns, with associative indegree explaining a small but significant amount of unique variance. Meanwhile, associative indegree was the strongest predictor for acquisition order of function words. The strong relationship between word frequency and nouns is consistent with previous findings in the literature, as is the difference in predictors between function words and other word classes. However there is still much work to be done to uncover how these predictors directly impact learning mechanisms in infant language learners, and why they are found to have variable strength for predicting acquisition of different word classes.

Acknowledgments

We are grateful to participating families and all researchers in the Oxford University BabyLab, past and present, who contributed to the collection of the Oxford CDI data used in this paper. We thank Dr Janette Chow for guidance provided on analysing corpus data and semantic networks.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4.

Journal of Statistical Software, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Braginsky, M., Sanchez, A., & Yurovsky, D. (2020). childesr: Accessing the 'childes' database [Computer software manual]. Retrieved from `https://github.com/langcog/childesr` (R package version 0.1.2)

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, *3*, 52–67.

Chang, L. M., & Deák, G. O. (2020). Adjacent and non-adjacent word contexts both predict age of acquisition of english words: A distributional corpus analysis of child-directed speech. *Cognitive Science*, *44*(11), e12899.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, *51*(3), 987–1006.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, *44*(3), 677.

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.

Hamilton, A., Plunkett, K., & Schafer, G. (2000). Infant vocabulary development assessed with a british communicative development inventory. *Journal of child language*, *27*(3), 689–705.

Hidaka, S. (2013). A computational model associating learning process, word attributes, and age of acquisition. *PLOS one*, *8*(11), e76242.

Hills, T. (2013). The company that words keep: comparing the statistical structure of child-versus adult-directed language. *Journal of child language*, *40*(3), 586–604.

Hills, 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.

Hills, T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological science*, *20*(6), 729–739.

Hochmann, J.-R., Endress, A. D., & Mehler, J. (2010). Word frequency as a cue for identifying function words in infancy. *Cognition*, *115*(3), 444–457.

Kachergis, G., Shiffrin, R., & Yu, C. (2009). Frequency and contextual diversity effects in cross-situational word learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 31).

Li, H., & Fang, A. C. (2011). Word frequency of the childes corpus: Another perspective of child language features. *ICAME Journal*, *35*, 95–116.

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk. transcription format and programs* (Vol. 1). Psychology Press.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological review*, *119*(4), 831.

Moon, K.-W. (2021). ggiraphextra: Make interactive 'ggplot2'. extension to 'ggplot2' and 'ggiraph' [Computer software manual]. Retrieved from `https://github.com/cardiomoon/ggiraphExtra` (R package version 0.3.0)

Naigles, L. R., & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs? effects of input frequency and structure on children's early verb use. *Journal of child language*, *25*(1), 95–120.

Nelson, D. L., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? *Memory & cognition*, *28*(6), 887–899.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

Rinker, T. W. (2018). textstem: Tools for stemming and lemmatizing text [Computer software manual]. Buffalo, New York. Retrieved from `http://github.com/trinker/textstem` (R package version 0.1.4)

Shaoul, C., & Westbury, C. (2006). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods*, *38*(2), 190–195.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *Journal of experimental child psychology*, *126*, 395–411.

Tomasello, M., & Todd, J. (1983). Joint attention and lexical acquisition style. *First language*, *4*(12), 197–211.

Wettler, M., Rapp, R., & Sedlmeier, P. (2005). Free word associations correspond to contiguities between words in texts. *Journal of Quantitative Linguistics*, *12*(2-3), 111–122.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental science*, *16*(6), 959–966.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.