Children's shift from CDS to ADS vocabulary across early childhood

Anonymous CogSci submission

Abstract

abstract text here. abstract text here.

Keywords: child-directed speech; word production; linguistic input; social register; corpus analysis; developmental change

Introduction

Methods

Corpus and item selection

We analyzed 8251 transcripts in the North American English collection of the Child Language Data Exchange System (CHILDES) database (MacWhinney, 2000). The included transcripts were drawn from 52 individual corpora and featured 980 children up to 7 years of age (range = 1–84 months, M = 33.5 months).

15 CDS/ADS word pairs were selected as test items based on their appearance on the MacArthur-Bates Communicative Development Inventory (Fenson et al., 1994) and their frequency of occurrence in CHILDES (at least 100 child-produced tokens and 100 other-produced tokens; see Table 1).

Cases where the same object, animal, or routine could be reasonably labeled with either form in typical communicative interactions with young children.

Linguistic predictors

All analyses were conducted over individual utterances. We quantified prosodic, lexical and syntactic information to describe each utterance containing one of the 30 target words.

Prosodic level For all timestamped utterances (41.4% of child-produced and 42.3% of other-produced utterances),

		CDS tokens by speaker		ADS tokens by speaker	
	Pair	**Child**	**Other**	**Child**	**Other**
1	*doggy/dog*	2249	2644	3519	5113
2	*kitty/cat*	1552	3309	2779	4443
3	*tummy/stomach*	435	623	112	360
4	*daddy/dad*	9603	10048	2313	1031
5	*mommy/mom*	20294	17070	7616	2552
6	*bunny/rabbit*	1237	2597	1060	1397
7	*duckie/duck*	307	647	1933	3003
8	*blankie/blanket*	174	224	825	874
9	*froggy/frog*	154	434	970	1846
10	*potty/bathroom*	511	786	161	270
11	*night night/goodnight*	149	153	102	446
12	*dolly/doll*	745	1054	674	2697
13	*horsey/horse*	1149	1034	1749	2575
14	*piggy/pig*	405	1212	1276	2139
15	*birdie/bird*	399	588	1879	3358

Table 1: CHILDES frequency for 15 CDS/ADS word pairs. Child-produced counts include tokens produced only by the target child. All other speakers' productions are included in the other-produced counts.

mean pitch and pitch range were extracted using Praat software (Boersma & Weenink, 2016). Speech rate was calculated as the number of words produced per second. Utterances shorter than 58 ms were excluded from analysis. This lower bound was set by identifying the the shortest possible duration of an utterance containing at least one word in four manually annotated North American English corpora in HomeBank (Bergelson, 2016; McDivitt & Soderstrom, 2016; VanDam et al., 2016; VanDam, 2016; Warlaumont & Pretzer, 2016; see also Bergelson et al., 2019).

Lexical level Typically, measures of lexical complexity are calculated over at least minutes or hours of transcribed speech rather than individual utterances. Here, we aimed to estimate local lexical complexity in two ways.

First, we calculated the negative log proportion of known words in each utterance (consistent with Foushee, Griffiths, & Srinivasan, 2016; Kidd, Piantadosi, & Aslin, 2012). A word was considered 'known' if the age of acquisition estimate (AoA) Frank, Braginsky, Yurovsky, & Marchman (2017) was less than or equal to the age of the target child when they heard or produced the utterance. Utterances with proportionally fewer known words are *more* lexically complex.

To account for more individual variation in word production, we included a measure

Syntactic level Syntactic measures included both the number of verb phrases and length (in words) of each utterance. Words were chosen over morphemes in the length analysis because we identified systematic errors in automatic morpheme counts. Manual checking of 5% of target utterances (i.e., XXXX utterances) revealed an error rate of XX%.

Results

Measuring production: When do children produce CDS vs. ADS forms?

Characterizing the input: In what linguistic contexts do children hear CDS vs. ADS forms?

We used mixed-effects binomial logistic regression models to predict the appearance of CDS vs. ADS forms in given utterance on the basis of target child's age, several linguistic properties of the utterance, and interactions between each property and age. Models included random intercepts for individual word pairs and speakers and were fitted to all utterance data from speakers other than the target child.

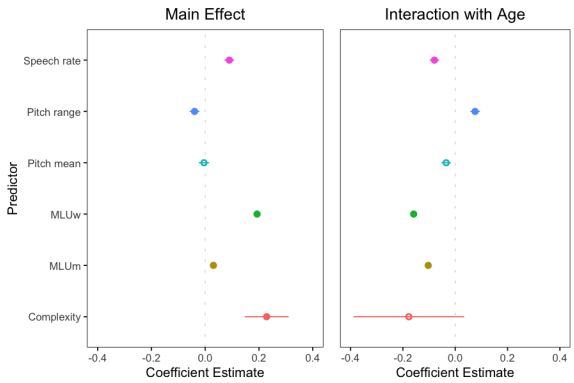

Figure 1: Coefficient estimates for linguistic predictors of form. Positive main effects indicate that utterances with higher levels of the predictor are more likely to contain ADS forms, relative to CDS forms. Positive age interactions indicate an increasing effect of the predictor with age. Error bars depict standard errors of the coefficient estimates, and filled circles represent significant effects ($p < 0.05$).

Modeling learning: What linguistic information can be used to distinguish CDS vs. ADS contexts?

Discussion

Acknowledgements

We are grateful to Claire Bergey, Ruthe Foushee, Ben Morris, and the members of the University of Chicago Chatter Lab and Northwestern University Child Language Lab for valuable discussion and feedback on this work.

References

10 Bergelson, E. (2016). Bergelson HomeBank corpus. https://doi.org/10.21415/T5PK6D.

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What do north american babies hear? A large-scale cross-corpus analysis. *Developmental Science*, *22*(1), e12724.

Boersma, P., & Weenink, D. (2016). Praat software. *Amsterdam: University of Amsterdam.*

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, i–185.

Foushee, R., Griffiths, T., & Srinivasan, M. (2016). Lexical complexity of child-directed and overheard speech: Implications for learning. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 1697–1702).

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), 677–694.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS One*, *7*(5), e36399.

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.

McDivitt, K., & Soderstrom, M. (2016). McDivitt HomeBank corpus. https://doi.org/10.21415/T5KK6G.

VanDam, M. (2016). VanDam2 HomeBank corpus.

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., De Palma, P., & MacWhinney, B. (2016). Homebank: An online repository of daylong child-centered audio recordings. https://homebank.talkbank.org.

Warlaumont, A. S., & Pretzer, G. M. (2016). Warlaumont HomeBank corpus. https://doi.org/10.21415/t54s3c.