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

Anonymous CogSci submission

Abstract

Child-directed speech (CDS) features words such as *doggy*, *night-night*, and *tummy* that are rarely used in adult-directed speech (ADS). Characterisitcs of CDS word forms, such as reduplication and diminutivization, explain why they may be learned and produced earlier by children. However, it is not yet clear how or when children switch to using ADS equivalents—*dog*, *goodnight*, *stomach*. Through analysis of transcripts from CHILDES and the Language Development Project corpus, we show that children significantly increase their production of ADS word forms across age, with the average CDS-to-ADS transition point at 2.5 years. Many of the linguistic features that distinguish CDS vs. ADS registers (e.g., speech rate, lexical complexity, etc.) similarly differentiate the local speech contexts surrounding CDS vs. ADS word forms. To test whether these patterns in children's input…classifier details…Learners may therefore be able to capitalize on these cues to support their discovery of context-appropriate CDS/ADS pair use.

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

Introduction

(1) CDS overview - beneficial for learning and x, y, z features

(2) Then register-specific words One characteristic feature of CDS is the presence of register-specific words, such as *doggy*, *night-night*, and *tummy*.

(3) Then relationship to social/contextual information Infants as young as 12 months show an emerging understanding of the relationship between registers and addressees (**soley2020infants?**). That is, infants hear CDS and adults hear ADS.

Word learners are not just determining what any given word means but also when it's

After establishing that a shift in children's productions occurs in early childhood, we next explored what information in their linguistic input could support this shift.

Methods

Corpora

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).

Additional transcripts from the Language Development Project (LDP) longitudinal corpus were also analyzed. These included XX transcripts from XX children recorded every X months for approximately one hour from age X to XX months.

Target words

Fifteen CDS/ADS word pairs (30 total target words) were selected 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). Test items were chosen such that the same object, animal, or routine could be reasonably labeled with either word in a given pair in typical communicative interactions with young children (e.g., *doggy* or *dog*).

		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.

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 We measured three types of prosodic information: mean pitch (Hz), pitch range (Hz), and speaking rate (words per second). These measures were calculated over all timestamped utterances in CHILDES (41.4% of child-produced and 42.3% of other-produced utterances). 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). Pitch information was extracted using Praat software (Boersma & Weenink, 2016).

Lexical level We measured two types of lexical information: complexity and rarity. Lexical complexity was defined as 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 (AoA) estimate (Fenson et al., 1994; 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. Rare words were determined based on overall CHILDES frequency. The rarest words have the lowest frequency. We calculated the average frequency for the words in each utterance that appeared at least one other time in the full corpus.

Syntactic level Syntactic measures included both the length of the utterance (in words) and the number of verb phrases. The number of words per utterance was automatically extracted using the `childesr` package (Braginsky, Sanchez, & Yurovsky, 2021). The number of verb phrases per utterance was determined using `spaCy3`, an automatic syntactic parser (Honnibal, Montani, Van Landeghem, & Boyd, 2020).

Results

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

We first asked when CDS forms are replaced by ADS forms in children's own speech. We fit mixed effects binomial logistic regression model predicting children's production of CDS vs. ADS forms with target child age (in months) as a single fixed effect. Random slopes and intercepts for individual word pairs and children were included.

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.

Main effects and interactions with age are shown in Figure

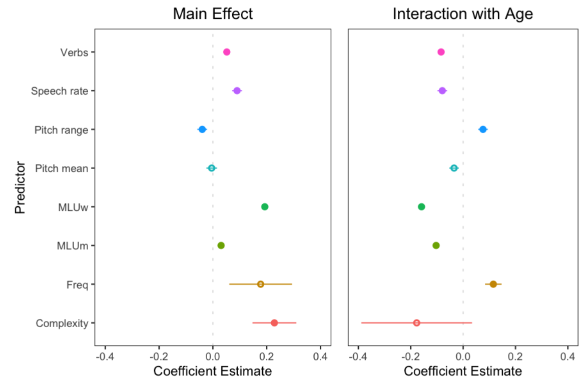

Figure 1: Coefficient estimates for linguistic predictors of form. Positive main effects indicate that utterances are more likely to contain ADS forms when they have higher values for the predictor (e.g., faster speech rates). Positive age interactions indicate an increasing effect of the predictor with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

Modeling learning: What linguistic information is most useful for distinguishing CDS vs. ADS contexts?

Discussion

References

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

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

Braginsky, M., Sanchez, A., & Yurovsky, D. (2021). *Childesr: Accessing the 'CHILDES' database*. Retrieved from `https://CRAN.R-project.org/package=childesr`

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.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. http://doi.org/10.5281/zenodo.1212303

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.