Developmental shifts in children's use of register-specific words

Anonymous CogSci submission

Abstract

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

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

Introduction

Across their first few years of life, children's expressive vocabulary grows to include hundreds if not thousands of words (Fenson et al., 1994; Mayor & Plunkett, 2011). Word production typically begins around age one, followed by a vocabulary 'explosion' or 'spurt' during toddlerhood (Ganger & Brent, 2004; see also McMurray, 2007) and continued, measurable increases in vocabulary size thereafter (Rice & Hoffman, 2015).

Predictors of early word comprehension and production

Which words do children understand and say first? Cross-linguistic predictors of early age of acquisition include overall frequency, frequency of occurrence in short or isolated utterances, and concreteness (Braginsky, Yurovsky, Marchman, & Frank, 2019). Words that adults rate as more associated with infancy (e.g., *bottle* and *bib*) are typically understood earlier (Perry, Perlman, & Lupyan, 2015), and shorter words with fewer phonemes are also produced earlier (Braginsky, Yurovsky, Marchman, & Frank, 2019).

Additional features of word forms, such as iconicity (Laing, Vihman, & Keren-Portnoy, 2017), diminutivization (Kempe, Brooks, & Gillis, 2005), and reduplication (Ota, Davies-Jenkins, & Skarabela, 2018), have been shown to support children's early learning. Words with these features (e.g., *doggy* and *choo-choo*) are typically overrepresented in children's early productive vocabularies (Frank, Braginsky, Yurovsky, & Marchman, 2017; Perry, Perlman, Winter, Massaro, & Lupyan, 2018) and serve as a hallmark of child-directed language (CDL) across many communities (Ferguson, 1964). Thus, children's productive vocabulary gets off the ground with many words that are particularly relevant for them. These include words for the key people, routines, and objects in their everyday lives (e.g., Tardif et al., 2008), along with words that appear specifically tailored to their level of linguisitc maturity (i.e., CDL-specific words).

However, at some point (likely gradually) children shift away from these baby-centric early words. This transition is perhaps most marked for common child language word forms that have direct substitutes in adult language (e.g., *doggy/dog*, *night-night/goodnight*, *tummy/stomach*; hereafter, 'CDL/ADL word pairs') as opposed to words that become less relevant over the years (e.g., *diaper*, *pacifier*, *peekaboo*). Note, though, that CDL-specific words do not fully disappear—children themselves come to use these lexical items (along with other CDL features) when talking to infants and younger children (e.g., Sachs & Devin, 1976; Shatz & Gelman, 1973).

Comprehension and production of language varieties

Classically, we might expect the appearance of both CDL and ADL labels for the same referent to be a problem for early word learning—particularly when the variants have little to no overlap in phonological form (e.g., *tummy/stomach*, *bunny/rabbit*). Indeed, children often apply a mutual exclusivity (ME) heuristic when learning new words; they assume that new labels refer to new items rather than interpreting them as synonyms for words that they already know (Markman & Wachtel, 1988; see Lewis, Cristiano, Lake, Kwan, & Frank, 2020, for a recent meta-analysis). Yet, children seem to contend with multiple CDL/ADL variants without issue.

One potential way to explain children's learning of both labels is to consider the social context of CDL vs. ADL use. While labeling an animal as *doggy* vs. *dog* may not communicate anything distinct about the referent itself, the production of one form vs. the other may indicate something meaningful about *who* is hearing or producing the label (i.e., CDL and children's own productions are more likely to feature

talk about *doggies*, while we would expect much more frequent talk about *dogs* in ADL). That is, differences in register could serve to 'explain away' the otherwise problematic redundancy of multiple labels (Clark & MacWhinney, 1987).

Over time, children become increasingly aware of the fact that language style is modulated by a variety of social factors, including the identities of speakers (e.g., from different social groups: Liberman, Woodward, & Kinzler, 2017) along with their addressees (e.g., young children vs. adults: Ikeda, Kobayashi, & Itakura, 2018; Soley & Sebastian-Galles, 2020). Moreover, children are exposed to many forms of socially meaningful variation in their linguistic input (see Johnson & White, 2020, for a recent review) – not just CDL vs. ADL words and registers (Bunce et al., 2020; Loukatou, Scaff, Demuth, Cristia, & Havron, 2021) but many different speakers (Bulgarelli & Bergelson, 2021; Rost & McMurray, 2009), accents (Buckler, Oczak-Arsic, Siddiqui, & Johnson, 2017; Potter & Saffran, 2017), dialects (Durrant, Delle Luche, Cattani, & Floccia, 2015; Edwards et al., 2014), and languages (Kremin, Alves, Orena, Polka, & Byers-Heinlein, 2020). Rather than viewing this variation as noise or a complication for learning, accumulating evidence suggests that experience with variable input affords children greater ability to detect, learn from, and apply social meaning to such variation. For instance, children's tendency to show the ME effect is modulated by their experience with multiple languages (and thus multiple labels: Byers-Heinlein & Werker, 2009; Houston-Price, Caloghiris, & Raviglione, 2010) as well as the social conditions under which multiple labels are introduced (e.g., by speakers of a familiar or unfamiliar race: Weatherhead, Kandhadai, Hall, & Werker, 2021).

Returning to the 'issue' of multiple CDL/ADL variants, we hypothesize that children may be able to make sense of this lexical variation by beginning to associate different forms with different modes of use (i.e., incipient representations of register). To test this idea, we first need to establish (a) when children begin to shift away from producing CDL-specific words, and (b) how children may be able to use bottom-up linguistic input cues to associate lexical variants with their appropriate registers (i.e., CDL vs. ADL).

Current investigation

We examine a small but core subset of 15 CDL-specific words in English (e.g., *doggy*, *night-night*, *tummy*) that are prevalent in children's early vocabularies but are eventually 'replaced' by ADL words—*dog*, *goodnight*, *stomach*. In Study 1, we analyze over 60,000 utterances of spontaneous speech from children up to seven years of age to establish when ADL forms become more dominant in children's own productions. That is, when do children 'switch' from using CDL forms to using ADL forms?

After establishing that the average age of 'switchover' occurs around 2.5 years, we next explored the features of children's input that could support this developmental shift. We investigated the extent to which CDL and ADL words are used in distinct linguistic contexts. Further processing of nearly 70,000 adult utterances revealed that CDL and ADL variants co-occur with reliably different patterns of prosodic, lexical, and syntactic information—cues that likely help learners associate them with different modes of use, or emerging representations of register.

Together, these studies push us to consider children's vocabulary development not as a simple accumulation of words or numeric increase in vocabulary size but rather a deepening and restructuring of the lexicon with growing linguistic and social maturity. The word *dog* does not entirely replace *doggy*—rather, the two forms become reserved for use with different addressees.

Study 1: When do children shift from CDL to ADL forms?

We tracked children's use of 15 CDL/ADL word pairs (Table 1) from early infancy up to age seven. Since CDL forms rarely appear in ADL, we predicted that children would shift away from production of these CDL-specific forms with increasing age. That is, we expected to see 'replacement' of CDL forms with ADL forms in children's own speech across time.

Method

Corpora

We analyzed 8,251 transcripts in the North American English collection of the Child Language Data Exchange System (CHILDES) database (MacWhinney, 2000) for children up to 7 years of age. The included transcripts were drawn from 52 individual corpora and featured 980 children (age range = 1–84 months, M = 33.5 months, with the earliest target word production detected at 3 months). Child production data from the Language Development Project (LDP) longitudinal corpus were also analyzed (see Huttenlocher, Waterfall, Vasilyeva, Vevea, & Hedges, 2010; Rowe, 2008, for further details regarding participating families, recording procedures, and transcription). LDP data included 622 transcripts from 59 English-learning children recorded every 4 months for approximately 90 minutes from age 14 to 58 months.

Target words

Fifteen CDL/ADL word pairs (30 total target words) were selected based on two criteria: the appearance of at least one form on the MacArthur-Bates Communicative Development Inventory (CDI, Fenson et al., 1994), and sufficient frequency of occurrence in CHILDES (at least 100 child-produced tokens and 100 other-produced tokens per form). Pairs were also selected based on researcher intuition to ensure that the same object, animal, routine, or body part could be reasonably labeled with either form by young children[1] Across

[1]While onomatopoeic words can be used in a similar manner to the CDL-specific words in our test set (e.g., *choo-choo* serving as a CDL-specific label for *train*, or *quack-quack* for *duck*), these iconic items were not included because they are primarily used as sound effects rather than labels for objects or animals (Skarabela, Pool, &

all transcripts, 64,852 child-produced utterances contained at least one target word and were included in our analysis.

Table 1: CHILDES frequency for 15 CDL/ADL word pairs. Child-produced counts include tokens produced only by the target child.

	CDL tokens by speaker		ADL tokens by speaker	
Pair	**Child**	**Other**	**Child**	**Other**
doggy/dog	2,249	2,644	3,519	5,113
kitty/cat	1,552	3,309	2,779	4,443
tummy/stomach	435	623	112	360
daddy/dad	9,603	10,048	2,313	1,031
mommy/mom	20,294	17,070	7,616	2,552
bunny/rabbit	1,237	2,597	1,060	1,397
duckie/duck	307	647	1,933	3,003
blankie/blanket	174	224	825	874
froggy/frog	154	434	970	1,846
potty/bathroom	511	786	161	270
night night/goodnight	149	153	102	446
dolly/doll	745	1,054	674	2,697
horsey/horse	1,149	1,034	1,749	2,575
piggy/pig	405	1,212	1,276	2,139
birdie/bird	399	588	1,879	3,358

Results

We asked when CDL forms are 'replaced' by ADL forms in children's own speech. Using the *lme4* package (version 1.1.27.1: Bates, Mächler, Bolker, & Walker, 2015) in R (version 4.1.0: R Core Team, 2021), we fit a mixed-effects binomial logistic regression model predicting children's production of CDL vs. ADL forms, with target child age (in months, scaled) as a single fixed effect. Random slopes and intercepts for word pairs were also included. For each target word token, the form was coded as either 0 (CDL) or 1 (ADL). Thus, the model captures, for each age, the relative proportion of CDL vs. ADL forms in children's own speech.

Children significantly increased their production of ADL forms over age ($\beta = 0.54$, $SE = 0.11$, $t = 4.92$, $p < 0.001$). The average CDL-to-ADL transition point (i.e., the point at which ADL forms were produced >50% of the time) was between 24 and 36 months of age (i.e., around 2.5 years; Figure 1).

The trend of increasing ADL form production was significant for 13 of 15 word pairs, but the exact trajectory of shift varied greatly across items (Figure 2). In some cases, CDL forms were 'replaced' by ADL forms early on (e.g., *doggy/dog* and *kitty/cat* around 2 years). For other pairs, the age 'switchover' was much later (e.g., *tummy/stomach* and *potty/bathroom* around 5 years). Finally, a clear point of 'switchover' was not observed for some pairs because ADL forms were produced >50% of the time from the earliest ages sampled (e.g., *duckie/duck* and *blankie/blanket*).

Ota, 2018). The polysemous nature of iconic word usage does not provide as clear of a test of 'replacement' of CDL forms with ADL forms over time.

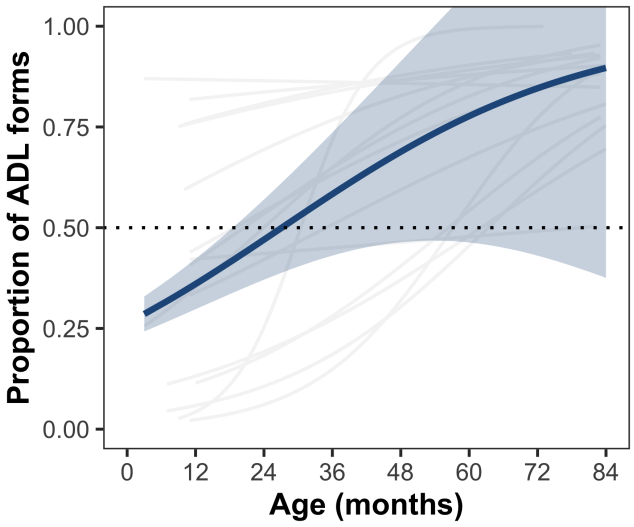

Figure 1: Model-predicted increase in production of ADL forms across age, with shaded standard error region. Gray lines depict individual word-pair trajectories.

Discussion

Analysis of children's own spontaneous speech revealed development shifts in their production of CDL vs. ADL forms, with the latter becoming increasingly more prominent over age. While we found substantial variation in the exact trajectories of CDL-to-ADL vocabulary shift for the 15 word pairs, the overall trend towards ADL production was clear. We take children's shifts away from CDL forms and towards ADL forms as indirect evidence of their early formation of CDL and ADL as distinct registers.

Study 2: What linguisitc information in children's input supports their shift from CDL to ADL forms?

We next explored children's input (i.e., other-produced speech), asking what linguistic information could support their shift from CDL to ADL forms. We conceptualize our second study as an investigation of the cues that could help learners associate CDL and ADL words with their appropriate registers.

CDL, as a register, is differentiated from ADL at multiple linguistic levels, including prosodic, lexical, and syntactic (e.g., Soderstrom, 2007). At least in English, CDL is associated with higher overall pitch as well as greater variability in pitch contours (Fernald, 1989; Vosoughi & Roy, 2012). CDL utterances are often produced more slowly (Ko & Soderstrom, 2013; e.g., Vigliocco, Shi, Gu, & Grzyb, 2020; but see also Martin, Igarashi, Jincho, & Mazuka, 2016). CDL also typically includes less lexical diversity (Hills, 2013) and more words that children already know (Foushee, Griffiths, & Srinivasan, 2016). Syntactically, CDL is characterized as

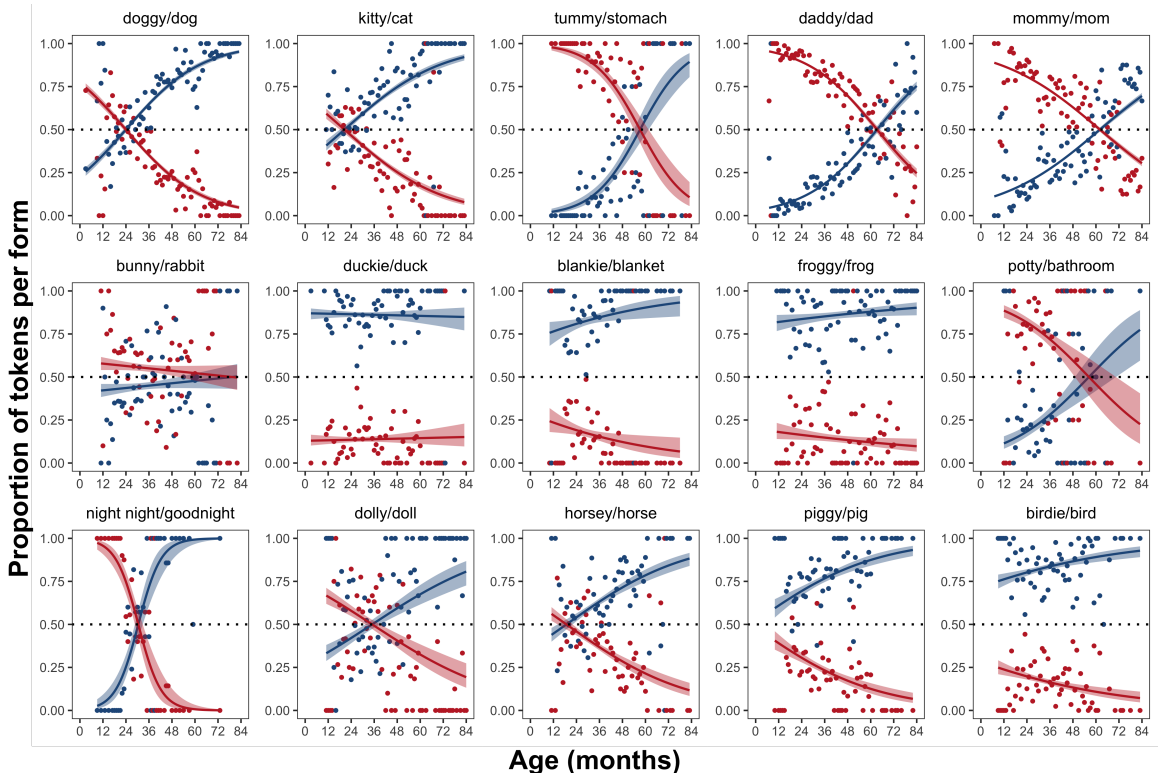

Figure 2: Inidividual word-pair trajectories for increasing production of ADL forms (blue) and descreasing production of CDL forms (red) with age. Points indicate proportions for each 1-month age bin.

less complex than ADL. CDL utterances are typically shorter (Brent & Siskind, 2001; Martin, Igarashi, Jincho, & Mazuka, 2016) and feature simpler constructions (Cameron-Faulkner, Lieven, & Tomasello, 2003).

Here, we tested whether the linguistic features that differentiate CDL vs. ADL at the register level also differentiate the local speech contexts surrounding CDL vs. ADL forms—even in speech that is primarily addressed to children from their primary caregivers (i.e., language from a single register). In other words, can form be predicted on the basis of individual utterance-level prosodic, lexical, or syntactic cues?

We hypothesized that utterances with CDL forms, relative to utterances with ADL forms, would be associated with (1) higher mean pitch, (2) greater pitch variability, (3) slower speaking rates, and (4) less lexical complexity. We also predicted that CDL utterances would contain (5) fewer rare words, (6) fewer words overall, and (7) fewer verb phrases. If these linguistic cues reliably differentiate CDL vs. ADL word usage contexts, then they could provide a viable source of information to support children's association between these words and their corresponding registers.

Method

Corpora

We analyzed 69,709 other-produced utterances (i.e., utterances not produced by the target child) in the same CHILDES transcripts from Study 1. The majority of utterances were produced by children's primary caregivers (n = 58,071, or 83.3%).

Linguistic input predictors

All input analyses were conducted over individual utterances containing at least one of the 30 target words from Study 1. We quantified prosodic, lexical, and syntactic information to describe each utterance.

Prosodic level We measured three types of prosodic information: **mean pitch** (Hz), **pitch range** (Hz), and **speech rate** (words per second). These measures were calculated over all timestamped utterances in CHILDES (42.3% of other-produced utterances). Utterances shorter than 58 ms were excluded from analysis.[2] 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

[2]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).

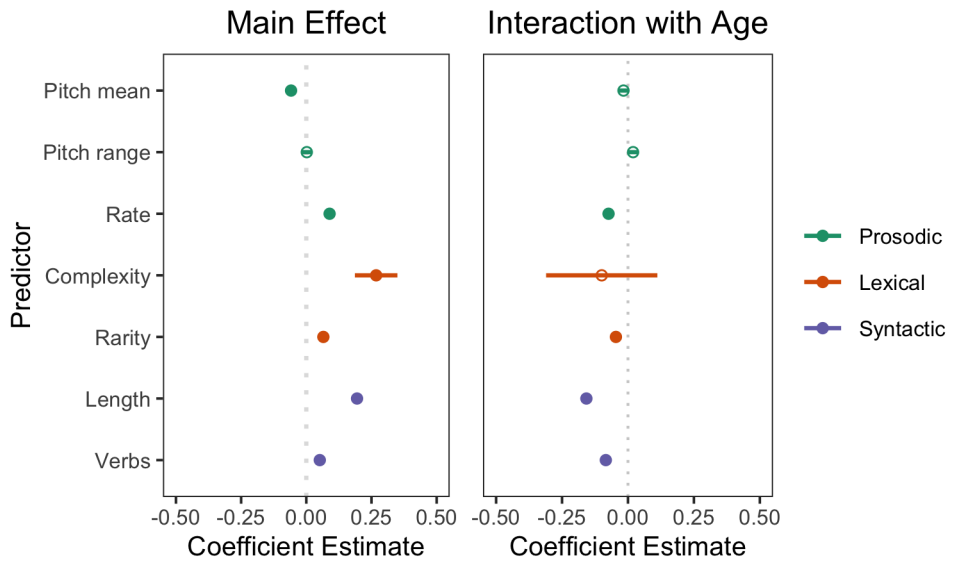

Figure 3: Coefficient estimates for linguistic predictors of form. Positive main effects indicate that utterances are more likely to contain ADL forms when they have higher values for the predictor (e.g., faster speech rates). Positive age interactions indicate an increasing effect of the predictor with age. Error bars depict standard errors of the coefficient estimates, and filled circles represent significant effects ($p < 0.05$).

al., 1994; Frank, Braginsky, Yurovsky, & Marchman, 2017) was less than or equal to the age of the target child when they heard the utterance. Utterances with proportionally fewer known words are considered more lexically complex. **Lexical rarity** was determined based on overall frequency in CHILDES. For all words with at least 10 tokens[3], we calculated a rarity score as the negative log proportion of other-produced tokens in CHILDES (i.e., number of tokens for a given word divided sum of all tokens of all words in the full corpus). We then averaged the rarity scores for all individual words in a given target utterance. Utterances with more low-frequency words are considered more lexically rare.

Syntactic level Syntactic measures included both the utterance **length** (in words) and **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

We ran individual mixed-effects binomial logistic regression models for each of seven linguistic input predictors. Models included fixed effects of linguistic predictor (scaled), target child age (in months, scaled), and their interaction as well as random intercepts for individual word pairs and speakers. For each target word token, the form was coded as CDL (0) or ADL (1), so coefficient estimates provide a measure of the strength of association between a predictor and ADL form. All main effects of linguistic predictors and interactions with age are shown in Figure 3.

At the prosodic level, we found significant effects for two of three input predictors tested. Utterance-level **pitch range** was not predictive of form ($\beta = 0.002$, $SE = 0.02$, $t = 0.10$, $p = 0.919$) and did not significantly interact with age ($\beta = 0.02$, $SE = 0.02$, $t = 1.11$, $p = 0.268$). However, utterance-level **mean pitch** was a negative predictor of ADL form ($\beta = -0.058$, $SE = 0.02$, $t = -3.02$, $p = 0.003$). That is, utterances with higher overall mean pitch were more likely to contain CDL forms, with no significant interaction with age ($\beta = -0.02$, $SE = 0.02$, $t = -0.96$, $p = 0.337$). **Speech rate** (i.e., words produced per second) was a positive predictor of ADL form ($\beta = 0.09$, $SE = 0.02$, $t = 4.86$, $p < 0.001$). Utterances spoken more quickly were more likely to contain ADL forms. This input predictor also negatively interacted with age ($\beta = -0.07$, $SE = 0.02$, $t = -3.96$, $p < 0.001$), indicating a decreasing strength in predictive power across developmental time.

At the lexical level, we found significant effects for both input predictors tested. Utterances with higher levels of **lexical complexity** ($\beta = 0.27$, $SE = 0.08$, $t = 3.28$, $p = 0.001$) and **lexical rarity** ($\beta = 0.07$, $SE = 0.01$, $t = 5.73$, $p < 0.001$) were more likely to contain ADL forms. Lexical complexity did not interact with age ($\beta = -0.10$, $SE = 0.21$, $t = -0.47$, $p = 0.637$); whereas, lexical rarity negatively interacted with age such that there was a decreasing effect of this predictor over time ($\beta = -0.05$, $SE = 0.01$, $t = -3.95$, $p < 0.001$).

At the syntactic level, we found significant effects of **utterance length** and **number of verb phrases**. Utterances with

[3]Manual checks revealed that many of the lowest-frequency words in CHILDES included idiosyncratic or erroneous transcriptions, so we excluded words with fewer than 10 tokens from our estimates of lexical rarity to reduce noise in this measure.

more words (β = 0.19, *SE* = 0.01, t = 16.30, $p < 0.001$) and more verb phrases (β = 0.05, *SE* = 0.01, t = 4.44, $p < 0.001$) were more likely to contain ADL forms. Moreover, both linguistic predictors negatively interacted with age (Length: β = -0.16, *SE* = 0.01, t = -14.61, $p < 0.001$; Verbs: β = -0.08, *SE* = 0.01, t = -7.70, $p < 0.001$), suggesting that the strength of these predictors decreases across developmental time.

Discussion

Analyses of children's input revealed reliable differences in the patterns of linguistic information surrounding CDL vs. ADL forms. Many of the prosodic, lexical, and syntactic features that broadly differentiate CDL vs. ADL registers similarly partitioned utterances containing CDL vs. ADL forms. Notably, these differences in local speech context emerged even in language that was primarily addressed to children from their primary caregivers (i.e., language likely from a single register—CDL).

While we do not yet know if these linguistic cues are actually exploited by learners, this study identifies which patterns appear learnable in principle. More broadly, this work provides support for the possibility that associations with CDL vs. ADL registers are helping learners grasp the differences in the context of CDL vs. ADL form use and gradually transition away from use of more contextually-constrained CDL-specific words.

End with a possible note about how this work can motivate new experimental investigations.

General Discussion

In the current work, we establish that children shift away from production of CDL-specific words over age. As predicted, these child-centric words are 'replaced' by ADL equivalents—at least until they again become relevant when talking to younger children. Further, we identify patterns in children's linguistic input that could support their discovery of associations between CDL/ADL words and their typical modes of use (i.e., incipient representations of register).

Developing linguistic and social knowledge in tandem

Children's linguistic knowledge builds around and together with their social knowledge. The lexical variants of CDL vs. ADL registers are just one example of socially meaningful linguistic variation. Variation also appears across languages, dialects, accents, and other types of registers (e.g., pedagogical, narrative, etc.). Early word learning is not just a process of one-to-one mapping. Here, we focus on the multiple labels for the same referent, but learners are also faced with the opposite problem: one label for many different referents (e.g., Casey, Potter, Lew-Williams, & Wojcik, 2021; Meylan, Mankewitz, Floyd, Rabagliati, & Srinivasan, 2021). Critically, learning is happening at multiple levels—not just words but associations between words and surrounding context (linguistic, social, etc.). Considering the interaction between these different factors can help us reason about learning mechanisms and children's representations—again not just of words but related social information too.

More than vocabulary size: Understanding words and using them in context

A merit of the approach of analyzing spontaneous language production is that we actually get to see variation in form that might otherwise be overlooked. For instance, typical caregiver-report measures, such as the MacArthur-Bates Communicative Development Inventories (CDIs: Fenson et al., 1994) , ask for a binary indication of whether a child 'knows' a word. For good reason, these surveys often gloss over variations in form. This standardization helps with generalizing over idiosyncratic words and focusing more on children's knowledge rather than adult-like articulation (allowing for large-scale, even cross-linguistic comparisons, e.g., Frank, Braginsky, Yurovsky, & Marchman, 2017, 2021). At the same time, glossing over variations in form presents a missed opportunity to investigate more nuanced aspects of vocabulary development.

In the end, returning to the point that was raised in the intro, which is that vocabulary development isn't just adding words to a bucket and the transition for these CDL/ADL forms is just one (pretty clear) way to see that.

References

10 Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. http://doi.org/10.18637/jss.v067.i01

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

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.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33–B44.

Buckler, H., Oczak-Arsic, S., Siddiqui, N., & Johnson, E. K. (2017). Input matters: Speed of word recognition in 2-year-olds exposed to multiple accents. *Journal of Experimental Child Psychology*, *164*, 87–100.

Bulgarelli, F., & Bergelson, E. (2021). Talker variability shapes early word representations in english-learning 8-month-olds. *Infancy*.

Bunce, J., Soderstrom, M., Bergelson, E., Rosemberg, C., Stein, A., Migdalek, M., ... others. (2020). A cross-cultural examination of young children's everyday language experiences.

Byers-Heinlein, K., & Werker, J. F. (2009). Monolingual, bilingual, trilingual: Infants' language experience influences the development of a word-learning heuristic. *Developmental Science*, *12*(5), 815–823.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27*(6), 843–873.

Casey, K., Potter, C., Lew-Williams, C., & Wojcik, E. H. (2021). Moving beyond" nouns in the lab": Using naturalistic data to understand why infants' first words include uh-oh and hi.

Clark, E. V., & MacWhinney, B. (1987). The principle of contrast: A constraint on language acquisition. *Mechanisms of Language Acquisition*, 1–33.

Durrant, S., Delle Luche, C., Cattani, A., & Floccia, C. (2015). Monodialectal and multidialectal infants' representation of familiar words. *Journal of Child Language*, *42*(2), 447–465.

Edwards, J., Gross, M., Chen, J., MacDonald, M. C., Kaplan, D., Brown, M., & Seidenberg, M. S. (2014). Dialect awareness and lexical comprehension of mainstream american english in african american english–speaking children. *Journal of Speech, Language, and Hearing Research*, *57*(5), 1883–1895.

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.

Ferguson, C. A. (1964). Baby talk in six languages. *American Anthropologist*, *66*(6_PART2), 103–114.

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Development*, 1497–1510.

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.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The wordbank project*. MIT Press.

Ganger, J., & Brent, M. R. (2004). Reexamining the vocabulary spurt. *Developmental Psychology*, *40*(4), 621.

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.

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

Houston-Price, C., Caloghiris, Z., & Raviglione, E. (2010). Language experience shapes the development of the mutual exclusivity bias. *Infancy*, *15*(2), 125–150.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology*, *61*(4), 343–365.

Ikeda, A., Kobayashi, T., & Itakura, S. (2018). Sensitivity to linguistic register in 20-month-olds: Understanding the register-listener relationship and its abstract rules. *Plos One*, *13*(4), e0195214.

Johnson, E. K., & White, K. S. (2020). Developmental sociolinguistics: Children's acquisition of language variation. *Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(1), e1515.

Kempe, V., Brooks, P. J., & Gillis, S. (2005). Diminutives in child-directed speech supplement metric with distributional word segmentation cues. *Psychonomic Bulletin & Review*, *12*(1), 145–151.

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.

Ko, E.-S., & Soderstrom, M. (2013). Additive effects of lengthening on the utterance-final word in child-directed speech.

Kremin, L. V., Alves, J., Orena, A. J., Polka, L., & Byers-Heinlein, K. (2020). Code-switching in parents' everyday speech to bilingual infants. *Journal of Child Language*, 1–27.

Laing, C. E., Vihman, M., & Keren-Portnoy, T. (2017). How salient are onomatopoeia in the early input? A prosodic analysis of infant-directed speech. *Journal of Child Language*, *44*(5), 1117–1139.

Lewis, M., Cristiano, V., Lake, B. M., Kwan, T., & Frank, M. C. (2020). The role of developmental change and linguistic experience in the mutual exclusivity effect. *Cognition*, *198*, 104191.

Liberman, Z., Woodward, A. L., & Kinzler, K. D. (2017). Preverbal infants infer third-party social relationships based on language. *Cognitive Science*, *41*, 622–634.

Loukatou, G., Scaff, C., Demuth, K., Cristia, A., & Havron, N. (2021). Child-directed and overheard input from different speakers in two distinct cultures. *Journal of Child Language*, 1–20.

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

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.

Martin, A., Igarashi, Y., Jincho, N., & Mazuka, R. (2016). Utterances in infant-directed speech are shorter, not slower. *Cognition*, *156*, 52–59.

Mayor, J., & Plunkett, K. (2011). A statistical estimate of infant and toddler vocabulary size from CDI analysis. *Developmental Science*, *14*(4), 769–785.

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

McMurray, B. (2007). Defusing the childhood vocabulary

explosion. *Science*, *317*(5838), 631–631.

Meylan, S., Mankewitz, J., Floyd, S., Rabagliati, H., & Srinivasan, M. (2021). Quantifying lexical ambiguity in speech to and from english-learning children.

Ota, M., Davies-Jenkins, N., & Skarabela, B. (2018). Why choo-choo is better than train: The role of register-specific words in early vocabulary growth. *Cognitive Science*, *42*(6), 1974–1999.

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in english and spanish and its relation to lexical category and age of acquisition. *PloS One*, *10*(9), e0137147.

Perry, L. K., Perlman, M., Winter, B., Massaro, D. W., & Lupyan, G. (2018). Iconicity in the speech of children and adults. *Developmental Science*, *21*(3), e12572.

Potter, C. E., & Saffran, J. R. (2017). Exposure to multiple accents supports infants' understanding of novel accents. *Cognition*, *166*, 67–72.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rice, M. L., & Hoffman, L. (2015). Predicting vocabulary growth in children with and without specific language impairment: A longitudinal study from 2; 6 to 21 years of age. *Journal of Speech, Language, and Hearing Research*, *58*(2), 345–359.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349.

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language*, *35*(1), 185–205.

Sachs, J., & Devin, J. (1976). Young children's use of age-appropriate speech styles in social interaction and role-playing. *Journal of Child Language*, *3*(1), 81–98.

Shatz, M., & Gelman, R. (1973). The development of communication skills: Modifications in the speech of young children as a function of listener. *Monographs of the Society for Research in Child Development*, 1–38.

Skarabela, B., Pool, E., & Ota, M. (2018). The train goes' choo choo': A corpus analysis of onomatopoeic words in child-directed speech and early production. In *BUCLD 43*.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532.

Soley, G., & Sebastian-Galles, N. (2020). Infants' expectations about the recipients of infant-directed and adult-directed speech. *Cognition*, *198*, 104214.

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

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.

Vigliocco, G., Shi, J., Gu, Y., & Grzyb, B. (2020). Child directed speech: Impact of variations in speaking-rate on word learning. In *Proceedings of the 42nd annual meeting of the cognitive science society* (Vol. 42, pp. 1043–1049). Cognitive Science Society.

Vosoughi, S., & Roy, D. K. (2012). A longitudinal study of prosodic exaggeration in child-directed speech.

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

Weatherhead, D., Kandhadai, P., Hall, D. G., & Werker, J. F. (2021). Putting mutual exclusivity in context: Speaker race influences monolingual and bilingual infants' word-learning assumptions. *Child Development*, *92*(5), 1735–1751.