From *doggy* to *dog*: Developmental shifts in children's use of register-specific words

Kennedy Casey
University of Chicago
kbcasey@uchicago.edu

Marisa Casillas
University of Chicago
mcasillas@uchicago.edu

Abstract

Child-directed language (CDL) features words such as *doggy*, *night-night*, and *tummy* that are rarely used in adult-directed language (ADL). Characteristics of CDL variants, 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 variants 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 variants in children's input. Notably, these differences emerged even in speech that was primarily child-directed. Learners may therefore be able to capitalize on these linguistic cues to support their discovery of register along with context-appropriate CDL/ADL pair use.

Keywords: child-directed language; word production; linguistic input; speech register; corpus analysis

Introduction

Across their first few years of life, children come to know 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). Here, we investigate one dimension of this dramatic developmental change: the appearance and use of words from distinct registers.

Vocabulary first gets off the ground, in part, with words that are specifically tailored to young learners (e.g., *doggy* and *tummy*: Ferguson, 1964). Hallmark features of child-directed language (CDL), 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 early word learning. These effects are in addition to the cross-cutting influence of a word's frequency, concreteness, length, and association with infancy (e.g., *bottle* and *bib*) on early learnability (Braginsky et al., 2019; Frank et al., 2017; Perry et al., 2018).

While CDL-specific words (e.g., *doggy*, *night-night*, *tummy*) are overrepresented in children's early vocabularies, they are eventually exchanged for ADL equivalents in most contexts (*dog*, *goodnight*, *stomach*). However, these words do not fully disappear. Instead, they become designated for use in a specific context—communication with infants and young children (e.g., Sachs & Devin, 1976; Shatz & Gelman, 1973).

The addition of a word like *stomach* to a child's vocabulary may mark their growing awareness that word choice should be tailored to the current interactional context (e.g., Clark, 1997, 2018). We do not, at present, know when children begin to make shifts from CDL to ADL word use or precisely how such a shift may be supported or initiated. Our investigation starts where this transition is most easily observed, with CDL/ADL word pairs (e.g., *doggy/dog*, *night-night/goodnight*, *tummy/stomach*), as opposed to words that become less relevant with time (e.g., *diaper* and *peekaboo*).

Some 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 overlap in phonological form (e.g., *bunny/rabbit* vs. *doggy/dog*). Indeed, learners often assume that new labels refer to new items (i.e., "mutual exclusivity": Markman & Wachtel, 1988; see Lewis et al., 2020, for a recent meta-analysis). Yet, children seem to learn multiple CDL/ADL variants without issue (see Clark, 1990).

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 variant vs. the other may indicate something meaningful about *who* is being addressed or producing the label. That is, differences in register could serve to 'explain away' the otherwise problematic redundancy of multiple labels in these pairs (Clark, 1990). Indirect evidence for this idea comes from findings that the mutual exclusivity effect is modulated by children's experience with multiple languages (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 et al., 2021). Further, children's modifications to their own speech when talking to infants and younger children (Sachs & Devin, 1976; Shatz & Gelman, 1973) and their awareness of socially meaningful linguistic variation (Ikeda, Kobayashi, & Itakura, 2018;

Liberman, Woodward, & Kinzler, 2017; Soley & Sebastian-Galles, 2020) suggest that they may be able to recognize the importance of social context for language use from relatively early in development.

We hypothesize that children may contend with CDL/ADL word pairs by associating the contrasting forms with different modes of use (i.e., by classifying each variant as belonging to a distinct 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 associated 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 variants become more dominant in children's own productions. That is, when do children switch from primarily using CDL variants to primarily using ADL variants? Our data suggest that the average age of CDL-to-ADL switch occurs around 2.5 years.

We then explore the features of children's input that could support this switch by examining the extent to which CDL and ADL words are used in distinct linguistic contexts. Further processing of nearly 70,000 non-target-child utterances (primarily from adult caregivers and addressed to the target child) 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 words *dog* and *stomach* do not entirely replace *doggy* and *tummy*—rather, the contrasting forms become reserved for use with different addressees.

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

We tracked children's use of 15 CDL/ADL word pairs (Table 1) from early infancy up to age seven. Since CDL variants rarely appear in ADL, we predicted that children would shift away from production of these CDL-specific words with increasing age. That is, we expected to see replacement of CDL variants with ADL variants 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). To further gain purchase on our question of interest with *longitudinal* data, we also analyzed children's productions in the Language Development Project (LDP) corpus (see Huttenlocher et al., 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 14 to 58 months.

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

	CDL tokens		ADL tokens	
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

Target words

Fifteen CDL/ADL word pairs (30 total target words) were selected based on two criteria: the appearance of at least one variant 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 variant). Pairs were also selected based on our own subjective judgment that the same object, animal, routine, or body part could be reasonably labeled with either variant by young children[1]. Across all transcripts, 64,852 child-produced utterances con-

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

tained at least one target word and were included in our analyses.

Results

We asked when CDL variants are replaced by ADL variants in children's own speech. Using the *lme4* package in R (Bates et al., 2015; R Core Team, 2021), we fit a mixed-effects binomial logistic regression model predicting children's production of CDL vs. ADL variants, with target child age (in months, scaled) as a single fixed effect. Random slopes and intercepts for word pairs were also included[2]. For each target word token, variant was coded as either 0 (CDL) or 1 (ADL). Thus, the model captures, for each age, the probability of using ADL variants over CDL variants.

Children significantly increased their production of ADL variants over age ($\beta = 0.54$, $SE = 0.11$, $t = 4.92$, $p < 0.001$ (Figure 1). The average CDL-to-ADL transition point (i.e., the point at which ADL variants were produced >50% of the time) was at approximately 28 months, or 2.5 years.

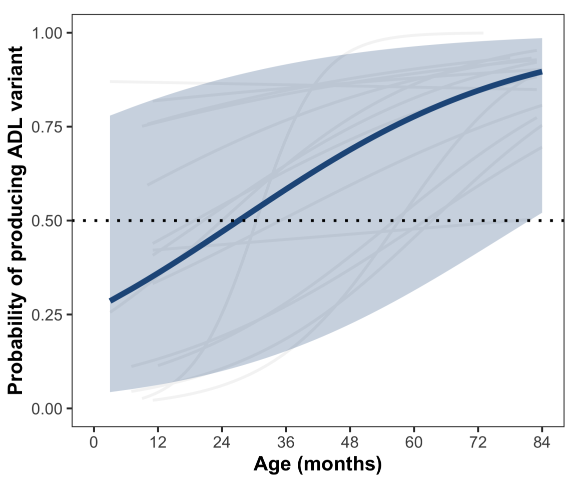

Figure 1: Model-predicted increase in production of ADL variants across age, with shaded standard error region. Gray lines depict individual word-pair trajectories.

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

To further examine the robustness of the overall effect of increasing ADL variant use over time, we ran subset analyses on all CHILDES transcripts (primarily cross-sectional, with hundreds of children), all LDP transcripts (longitudinal, $n = 59$, age range = 14–58 months), and all transcripts from the Providence corpus, a small, longitudinal subset of CHILDES ($n = 6$, age range = 11–48 months: Demuth, Culbertson, & Alter, 2006). The main finding was replicated in all three individual corpora. Children significantly increased their production of ADL variants over age, collectively across all CHILDES corpora ($\beta = 0.55$, $SE = 0.11$, $t = 4.95$, $p < 0.001$), as well as in the LDP corpus ($\beta = 0.38$, $SE = 0.04$, $t = 8.61$, $p < 0.001$) and Providence corpus ($\beta = 0.45$, $SE = 0.14$, $t = 3.23$, $p = 0.001$). Moreover, the average CDL-to-ADL transition point was estimated to be around 2.5 years in all corpora (CHILDES: 28 months, LDP: 30 months, Providence: 27 months).

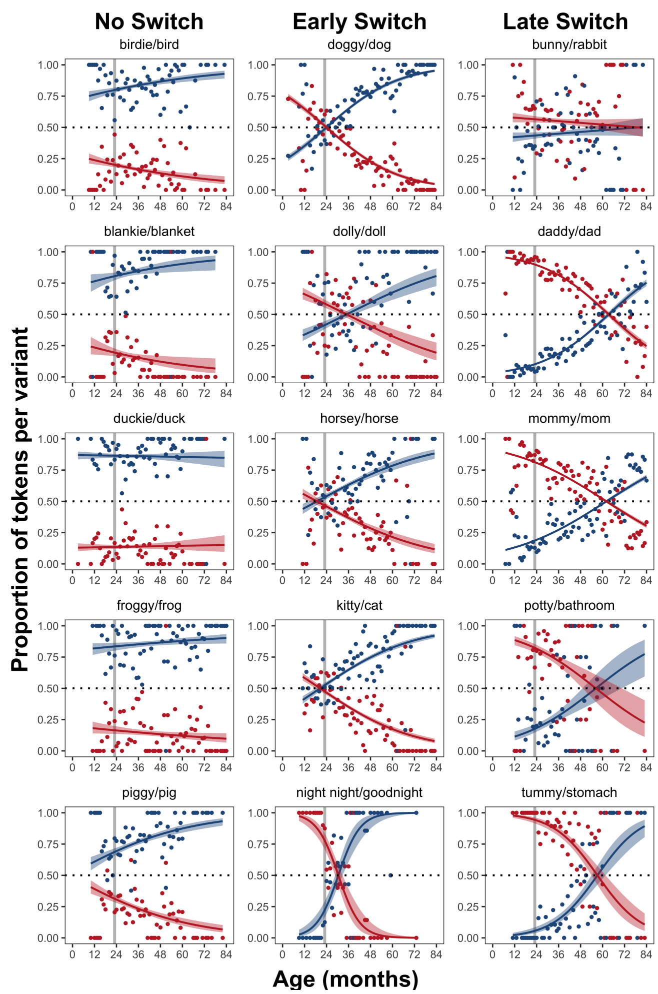

Figure 2: Individual word-pair trajectories for increasing production of ADL variants (blue) and decreasing production of CDL variants (red) with age. Points indicate proportions for each 1-month age bin. Vertical gray lines at 28 months indicate the overall model-predicted CDL-to-ADL transition point across all words.

[2] glmer(variant ~ age (months, scaled) + (1 + age | word pair), family = binomial)

Discussion

Analysis of children's own spontaneous speech revealed developmental shifts in their production of CDL vs. ADL variants, with the latter becoming increasingly more prominent over age. As an indicator of robustness, this effect emerged in three different corpora with vastly different sample sizes and distinct sampling strategies (i.e., cross-sectional vs. longitudinal). We additionally found substantial pair-level variation in the exact trajectories of CDL-to-ADL vocabulary shift, but overall, we take children's shifts away from CDL variants and toward ADL variants over time as indirect evidence of their early formation of CDL and ADL as distinct registers.

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

We next explored children's input (i.e., other-produced speech), asking what linguistic information could support their shift from CDL to ADL vocabulary. We conceptualize our second study as an investigation of the cues that could help learners associate CDL and ADL variants 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). 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 (e.g., Ko & Soderstrom, 2013; Vigliocco et al., 2020; but see Martin et al., 2016). CDL typically includes less lexical diversity (Hills, 2013) and more words that children already know (Foushee, Griffiths, & Srinivasan, 2016). Syntactically, CDL is characterized as less complex than ADL. CDL utterances are typically shorter (Brent & Siskind, 2001; Martin et al., 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 variants—even in speech that is primarily addressed to children from their adult caregivers (i.e., language from a single register). In other words, can the appearance of one variant vs. the other be predicted on the basis of individual utterance-level prosodic, lexical, or syntactic cues?

We hypothesized that utterances with CDL variants (vs. ADL variants) 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 ability to associate these lexical variants with their corresponding registers.

Method

Corpora

In addition to the child-produced utterances from Study 1, we analyzed 69,709 other-produced utterances (i.e., utterances not produced by the target child) in the same CHILDES transcripts. The majority of utterances were produced by children's primary caregivers (n = 58,071, or 83.3%). While, by and large, the utterances are not annotated for addressee, our manual scanning of the CHILDES corpora suggested that the vast majority of this speech is addressed to the target child. Study 2 analyses exclude the LDP corpus because it has not been comprehensively timestamped.

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, 41.4% of child-produced utterances). Utterances shorter than 58ms were excluded from analysis[3]. 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 et al., 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 considered more lexically complex. **Lexical rarity** was determined based on overall frequency in CHILDES. For all words with at least 10 tokens[4], 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 by the 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

[3]This lower bound was set by identifying the shortest possible duration of an utterance containing at least one word in four manually annotated North American English corpora (see Bergelson et al., 2019, for details).

[4]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.

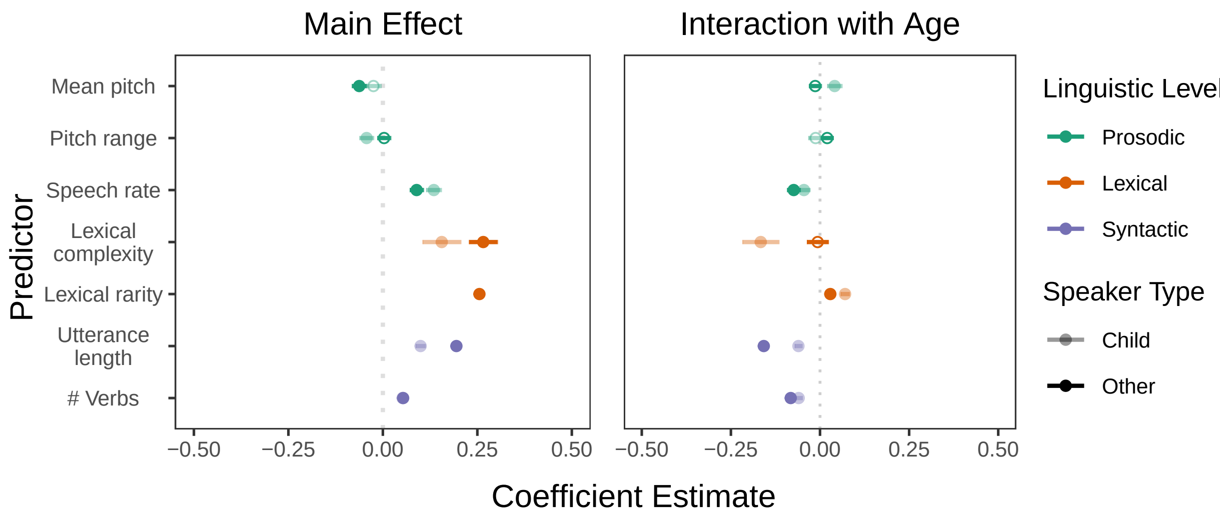

Figure 3: Coefficient estimates for linguistic predictors of variant. Positive main effects (left panel) indicate that utterances are more likely to contain ADL variants when they have higher values for that predictor (e.g., more verbs). Positive age interactions (right panel) indicate an increasing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

using the *childesr* package (Braginsky, Sanchez, & Yurovsky, 2021). The number of verb phrases per utterance was determined using *spaCy3*, an automatic syntactic parser (Honnibal et al., 2020).

Results

We ran individual mixed-effects binomial logistic regression models of CDL vs. ADL variant use 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[5]. For each target word token, the variant was coded as CDL (0) or ADL (1), so coefficient estimates provide a measure of the strength of association between a predictor and ADL variants.

First, we ran our models on all other-produced utterances (i.e., children's input). Main effects of linguistic predictors and interactions with age are shown in Figure 3. All models also revealed a positive main effect of target child age (all *p*s < 0.001), confirming that adults, like children in Study 1, increase ADL variant production as their child addressees get older.

At the prosodic level, we found significant effects for two of the three input predictors tested. Utterance-level **pitch range** was not predictive of variant ($\beta = 0.003$, $SE = 0.02$, $t = 0.18$, $p = 0.858$) and did not significantly interact with age ($\beta =$, $SE =$, $t =$, p). However, utterance-level **mean pitch** was a negative predictor of ADL variant ($\beta =$, $SE =$, $t =$, p). That is, utterances with higher overall mean pitch were more likely to contain CDL variants, with no significant interaction with age ($\beta =$, $SE =$, $t =$, p). **Speech rate** (i.e., words produced per second) was a positive predictor of ADL variant ($\beta =$, $SE =$, $t =$, p). Utterances spoken more quickly were more likely to contain ADL variants. This input predictor also negatively interacted with age ($\beta =$, $SE =$, $t =$, p), indicating a decreasing strength in predictive power across developmental time.

At the lexical level, we found significant effects for both input predictors. Utterances with higher levels of **lexical complexity** ($\beta =$, $SE =$, $t =$, p) and **lexical rarity** ($\beta =$, $SE =$, $t =$, p) were more likely to contain ADL variants. Lexical complexity did not interact with age ($\beta =$, $SE =$, $t =$, p); whereas, lexical rarity negatively interacted with age such that there was a decreasing effect of this predictor over time ($\beta =$, $SE =$, $t =$, p).

At the syntactic level, we found significant effects of **utterance length** and **number of verb phrases**. Utterances with more words ($\beta =$, $SE =$, $t =$, p) and more verb phrases ($\beta =$, $SE =$, $t =$, p) were more likely to contain ADL variants. Both linguistic predictors negatively interacted with age (Length: $\beta =$, $SE =$, $t =$, p ; Verbs: $\beta =$, $SE =$, $t =$, p), suggesting that the strength of these predictors decreases across developmental time.

Across all levels of linguistic representation, children's productions largely mirrored others' (all main effects and interactions shown in Figure 3). That is, children's own utterances showed reliably different patterns of prosodic, lexical, and syntactic information for utterances with CDL vs. ADL variants.

Discussion

Analyses of children's input revealed reliable differences in the patterns of linguistic information surrounding CDL

[5] glmer(variant ∼ linguistic predictor (numeric, scaled) * age (months, scaled) + (1 | word pair) + (1 | speaker), family = binomial)

vs. ADL variants. Many of the prosodic, lexical, and syntactic features that broadly differentiate CDL vs. ADL registers similarly partitioned utterances containing CDL vs. ADL variants. 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). This finding underscores the idea that register production reflects stylistic linguistic choices by the person producing the utterance and does not necessarily require the prototypical communicative context (e.g., a caregiver can use ADL-like utterances when talking to a young child).

While we do not yet know if these linguistic cues are indeed exploited by learners, this study identifies which patterns appear learnable in principle and which patterns are reflected in children's own productions. More broadly, this work provides support for the possibility that associations with CDL vs. ADL registers help learners grasp the differences in the contexts of CDL vs. ADL variant use and thereby support their gradual transition away from production of more contextually-constrained CDL-specific words. A next step is to experimentally test how well children across this age range perceive words as CDL- or ADL-relevant given the surrounding linguistic context, or how their expectations for hearing one variant vs. the other may be modulated by linguistic cues such as mean pitch, lexical complexity, and utterance length.

General Discussion

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

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

By analyzing spontaneous language production in the present study, we find variation in form that is often overlooked but may be crucial for understanding how children's vocabularies develop. Widely-used caregiver-reported (Fenson et al., 1994) and researcher-administered (Dunn & Dunn, 1965) vocabulary measures typically ask for a binary indication of whether a child 'knows' a word. For good reason, these surveys and tests often gloss over variations in form. This standardization helps with generalizing over many idiosyncrasies, which allows for large-scale, even cross-linguistic, comparisons (e.g., Frank et al., 2017, 2021). At the same time, glossing over this lexical variation may present a missed opportunity to investigate more nuanced but essential aspects of vocabulary development. The present findings on the transition between CDL and ADL variants help demonstrate that vocabulary development taps into other major aspects of children's language learning, including their recognition of multiple levels of linguistic information (e.g., prosodic, lexical, and syntactic) and their broader socialization as individuals who can effectively deploy language across variable contexts.

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 in children's input. Variation also appears across languages, dialects, accents, and other types of registers (e.g., pedagogical, narrative, etc.). 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). Children may therefore be able to leverage this social knowledge when learning language. We see examples of this in the context of word learning when children show flexibility in applying the mutual exclusivity heuristic in accordance with the social conditions under which new words are introduced (e.g., Weatherhead et al., 2021), and here, too, children's shifts from CDL to ADL vocabulary are likely supported by their emerging knowledge about the contexts in which different registers are used.

We focus here on the issue of encountering multiple labels for the same referent in early word learning, but children also face the inverse problem—one label for many different referents (e.g., Casey et al., 2021; Meylan et al., 2021). We see these puzzles of word learning as interrelated—and given children's early success in contending with both sources of variability, as evidence that learning happens at multiple levels.

Rather than conceptualizing vocabulary development as the simple tallying up of new 'known' words in a relatively low-dimensional semantic space, there is richness to be found in interactions with other types of information (linguistic, social, etc.) and in analyses of change over time. Exploring children's *use* words in varying forms and contexts can provide insight into the patterns of information that supported their learning in the first place.

Data Availability
All anonymized data and analysis scripts can be found at the following link: https://github.com/kennedycasey/RegisterShift.

Acknowledgements

We are grateful to the members of the University of Chicago Chatter Lab and Northwestern University Child Language Lab for valuable discussion and feedback on this work. Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P01HD040605; by a grant from the Successful Pathways from School to Work initiative of the University of Chicago, funded by the Hymen Milgrom Supporting Organization; by a grant from the Institute of Education Sciences, U.S. Department of Education, through grant R305A190467 to the University of Chicago; and by a grant from the Spencer Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the Institute of Education Sciences, or the U.S. Department of Education.

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.

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*.

Braginsky, M., Sanchez, A., & Yurovsky, D. (2021). *Childesr: Accessing the 'CHILDES' database*.

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.

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. E., 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. *PsyArXiv*.

Clark, E. V. (1990). On the pragmatics of contrast. *Journal of Child Language*, *17*(2), 417–431.

Clark, E. V. (1997). Conceptual perspective and lexical choice in acquisition. *Cognition*, *64*(1), 1–37.

Clark, E. V. (2018). Conversation and language acquisition: A pragmatic approach. *Language Learning and Development*, *14*(3), 170–185.

Demuth, K., Culbertson, J., & Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of english. *Language and Speech*, *49*(2), 137–173.

Dunn, L. M., & Dunn, L. M. (1965). *Peabody picture vocabulary test*.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., & 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*, 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. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, 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*.

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.

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., & Soderstrom, M. (2013). Additive effects of lengthening on the utterance-final word in child-directed speech. *Journal of Speech, Language, and Hearing Research*, *56*(1), 364–371.

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.

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.

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. *PsyArXiv*.

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

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

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.

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. *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.

Vigliocco, G., Shi, J., Gu, Y., & Grzyb, B. (2020). Child directed speech: Impact of variations in speaking-rate on word learning. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, *42*, 1043–1049.

Vosoughi, S., & Roy, D. K. (2012). A longitudinal study of prosodic exaggeration in child-directed speech. *Speech Prosody Special Interest Group (SProSIG)*.

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.