Quantifying the amount and impact of label variation in early word learning

Kennedy Casey[1], Amalia Levitin[1], Jessica Kosie[2], & Casey Lew-Williams[1]
[1] Department of Psychology, Princeton University
[2] School of Social and Behavioral Sciences, Arizona State University

Background

Early word learning is traditionally framed as a **one-to-one mapping** problem[1,2], but this model oversimplifies learners' real-world experiences.[3]

When learning words, children often encounter many different labels for the exact same referent—for instance, dog can be called "doggy", "puppy", "woof-woof", etc.[4-6]

This kind of **label variation** is an attested feature of child-directed speech in English and other languages.[7]

We investigate (1) the naturalistic frequency of label variation in English child-directed speech, and (2) how this variation relates to children's word production.

Method

Naturalistic video corpus

10-minute, in-home free play interactions between children and one primary caregiver recorded on Zoom[8]

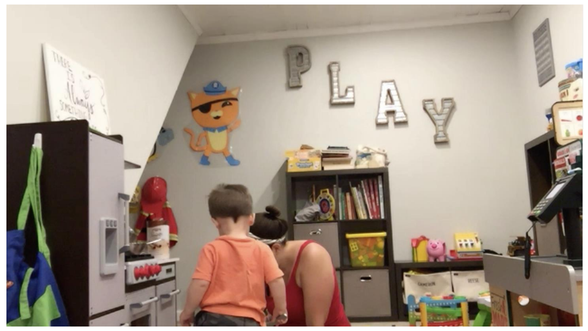

Figure 1. Example Zoom recording setup for Infant-Directed Communication corpus (Kosie & Lew-Williams, 2024). Videos are shared on Databrary.

Age range: 18-24 months (M_{age} = 20.5, SD_{age} = 1.9)
N: 44 caregiver-child dyads
Demographics: Monolingual English speakers
Children: 77% white, 18% multiracial, 5% Asian
Caregivers: 32% 4-year college degree, 61% graduate degree
Vocabulary measure: MacArthur-Bates CDI (Words & Sentences)[9] completed before the video recording session

Manual coding procedure

Videos were manually transcribed and annotated in ELAN.[10]

Two independent coders identified **referents** of 224 MB-CDI object/animal nouns and their corresponding **standard or variant labels**.

Note: Pronouns and production errors were excluded from variant counts. Inflected word forms were collapsed.

Reliability on 20% of videos: 93.7% agreement for referents and 97.1% agreement for labels

Results

Descriptive statistics for caregivers' use of label variation

Caregivers labeled **146 unique referents** in total across the sample:
- 65.6% standard labels
- 34.4% variant labels

M = 2.67 unique labels per referent (median = 2, range = 1-29).

Labels were categorized into 5 types:
standard: MB-CDI word forms
proper/relational: proper names, familial terms, etc.
wordplay: child-directed variants (e.g., diminutivization, reduplication, onomatopoeia, etc.)[4-6]
taxonomic: subordinate and superordinate labels
descriptive/functional: colors, shapes, functions, etc., replacing the standard noun

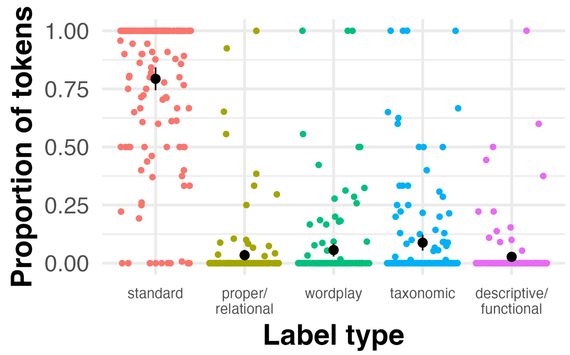

Figure 2. Distribution of variant label types. Each point corresponds to an individual MB-CDI referent. Point ranges show means and 95% CIs.

Finding #1: Early-learned words are associated with *more* wordplay variation

Overall, caregivers used significantly more wordplay variants for nouns with earlier AoA[11], controlling for word frequency.[12]

Negative binomial model: # of variants ~ AoA + freq
Effect of AoA: b = -0.27, SE = 0.09, p = 0.003
Effect of frequency: b = -0.13, SE = 0.27, p = 0.641
AoA x frequency interaction did not improve model fit

At the utterance level, caregivers were also more likely to use wordplay variants (but not other variant types) for earlier-learned nouns.

Binomial model: standard vs. variant ~ AoA + freq + (1 + AoA + freq | dyad)
Effect of AoA: b = -0.28, SE = 0.10, p = 0.006
Effect of frequency: b = -0.73, SE = 0.24, p = 0.003
AoA x frequency interaction did not improve model fit

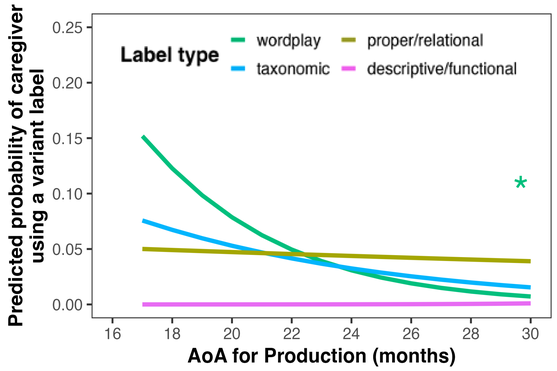

Figure 3. Utterance-level probability of using variant label by type.

Finding #2: Caregivers use more label variation for words their child does *not* yet produce

Figure 4. Probability of a caregivers using a variant label (vs. a standard MB-CDI label) predicted by parent-reported production of individual MB-CDI nouns. Points reflect raw probabilities for individual children and referents. Binomial model: standard vs. variant ~ produces + age + freq + (1 + produces + freq | dyad) + (1 | word)

Discussion

Label variants make up a sizable portion of English-hearing children's input for early word learning.

Early-learned words were associated with *more* wordplay variation than later-learned words, replicating recent studies using SEEDLingS[5,13] and CHILDES.[12,14]

Caregivers used more label variation for words that their child did *not* yet produce.

These findings build on prior work linking features of child-directed/wordplay variants to early learnability across languages[6,15,16] and new experimental evidence showing English-hearing toddlers' ability to learn novel words from input featuring wordplay variation[14].

This work suggests that **label variation does not hinder—and may instead support—early word learning.**

References

[1] Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3). 1558-1568.
[2] Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2017). The pursuit of word meanings. *Cognitive Science, 41*, 638-676.
[3] Wojcik, E. H., Zettersten, M., & Benitez, V. L. (2022). The map trap: Why and how word learning research should move beyond mapping. *WIREs Cognitive Science, 13*, e1596.
[4] Casey, K., & Casillas, M. (2022). From *doggy* to *dog*: Developmental shifts in children's use of register-specific words. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society* (pp. 49-56).
[5] Moore, C., & Bergelson, E. (2024). Wordform variability in infants' language environment and its effects on early word learning. *Cognition, 245,* 105694.
[6] 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.
[7] Ferguson, C. A. (1964). Baby talk in six languages. *American Anthropologist, 66*(6), 103-114.
[8] Kosie, J. E., & Lew-Williams, C. (2024). Infant-directed communication: Examining the many dimensions of everyday caregiver-infant interactions. *Developmental Science, 27*, e13515.
[9] deMayo, B. E., Kellier, D., Braginsky, M., Bergmann, C., Hendriks, C. A., Rowland, C. F., Frank, M. C., & Marchmann, V. (2021). Web-CDI: A system for online administration of the MacArthur-Bates Communicative Development Inventories. *Language Development Research, 1*(1), 55-98.
[10] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1556-1559).
[11] 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.
[12] MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs.* Lawrence Erlbaum Associates Publishers.
[13] Bergelson, E. (2017) Bergelson SEEDLingS HomeBank Corpus.
[14] Casey, K., & Lew-Williams, C. (2024, November 7-10). *Contending with label variation in early word learning* [Presentation]. Boston University Conference on Language Development. Boston, MA, United States.
[15] Kızıldere, E., Can Esmer, Ş., & Göksun, T. (2022). From *woof woof* to *dog*: Interactions between parents' use of sound symbolic words and infants' vocabulary development. *Infancy, 27*, 972-996.
[16] Han, M., & Gu, Y. (2023). Che-che ('car-car') and chi-chi ('eat-eat'): Reduplication in Mandarin Chinese child-directed speech. *Proceedings of the 45th Annual Meeting of the Cognitive Science Society* (pp. 920-927).

Contact: kcasey@princeton.edu

GitHub poster | corpus | Databrary

Acknowledgements: This project was supported by funding from the National Science Foundation (GRFP to KC, SPRF to JK), the National Institute of Child Health and Human Development (F32HD103439 to JK, R01HD095912 to CLW), the Princeton University Office of Undergraduate Research (AL). We thank Ruby Trujillo for critical annotation support.