Children do not overuse "the" in natural production

Yuanfan Ying, Valentine Hacquard, Alexander Williams, Jeffrey Lidz
University of Maryland, College Park

English-learning children (age 3-5) have been reported to overuse "the": in elicited production studies, they sometimes use "the" when the referent is either not in the common ground [1][2][3] or not unique [4][5][6]. However, comprehension studies show no clear support for that observation [7][8]. The conflicting data suggests two possibilities: either comprehension data overestimates children's competence, or elicited production results reflect performance issues. We conducted two studies—one quantitative, one qualitative—examining children's natural production and found no evidence for systematic overuse of "the" in children's productions even with 2-year-olds, suggesting that English-learning children have a grasp of the definite/indefinite distinction early.

Through a corpus study, we first examined the distribution of definite phrases produced by 27 children and mothers in several CHILDES corpora (912,530 words) [9][10][11][12] and found no support for *the*-overuse. Children's percentage of definites (number of 'the N_{SG}' divided by number of 'a+the N_{SG}') does not exceed the baseline of mothers'. Crucially, children's determiner use follows their mothers', even when we break down instances into different pragmatic categories (assertions vs. questions; Figure 1) and syntactic environments (subject vs. object vs. fragment; Figure 2). They do not overuse "the" in any of the contexts examined, and they use more definite phrases in subjects than in objects. This is in line with their mothers' use, which reflects the generalization that in English, subjects tend to refer to things mentioned in prior discourse.[13] Moreover, if children frequently misuse "the", we'd expect to see more signs of misunderstanding from mothers when children ask questions with "the". Yet we found only 3.4% (7 out of 205) cases of seeming referent-related miscommunication with the 2-year-old corpora, which is small compared to the high overuse rates in elicited production studies.

To determine whether children's natural use of definites is adult-like, we conducted a determiner-guessing experiment, where adults (N = 240) were presented with short excerpts of mother-child conversations from 23 mother-child pairs, where a determiner produced by either child or mother was missing, and had to guess whether it was a definite or an indefinite. Those excerpts consisted of 10-line dialogs randomly drawn from the Gleason corpus in CHILDES (80,347 words).[14] If children systematically use definites where adults would use indefinites, we would expect adults to make more wrong guesses for children's definites than for mothers'. We found no such evidence for *the*-overuse: our results show that 1) adults were above chance at guessing definites used by either mothers or children for 3-, 4-, and 5-year-olds' corpora, and 2) they were never significantly better at guessing mothers' definites than children's (Table 1). Regarding individual test items, adults' average error rate was not significantly different for mothers' and children's definites with 3- and 4-year-olds; with 5-year-olds, adults' error rates were higher for *mothers'* definites (Table 2).

Taken together, *the*-overuse is not reflected in natural production data, suggesting that children's elicited production errors could be due to experimental artifacts, and that an alternative performance account is needed to explain the alleged *the*-overuse with children.

(Word count: 498 words)

Table 1. Adults' percentage of definite determiner match

Age	Speaker	N	Mean	SD	Wilcoxon test	Mann-Whitney U test
3-year-olds	child	36	0.844	0.144	$V = 595$, $p < .001$	$W = 516$, $p = .124$
	mother	36	0.900	0.096	$V = 666$, $p < .001$	
4-year-olds	child	35	0.857	0.127	$V = 595$, $p < .001$	$W = 571.5$, $p = .212$
	mother	39	0.895	0.102	$V = 780$, $p < .001$	
5-year-olds	child	34	0.894	0.110	$V = 595$, $p < .001$	$W = 726.5$, $p = .007$**
	mother	31	0.765	0.236	$V = 389$, $p < .001$	

Table 2. Adults' error rates with mothers and children

Age	Speaker	N	Mean	SD	Kruskal-Wallis test
3-year-olds	child	20	16.667	17.763	$p = .067$
	mother	20	9.356	16.232	
4-year-olds	child	20	15.262	21.082	$p = .320$
	mother	20	11.173	19.225	
5-year-olds	child	20	12.854	15.206	$p = .014$*
	mother	20	25.375	21.567	

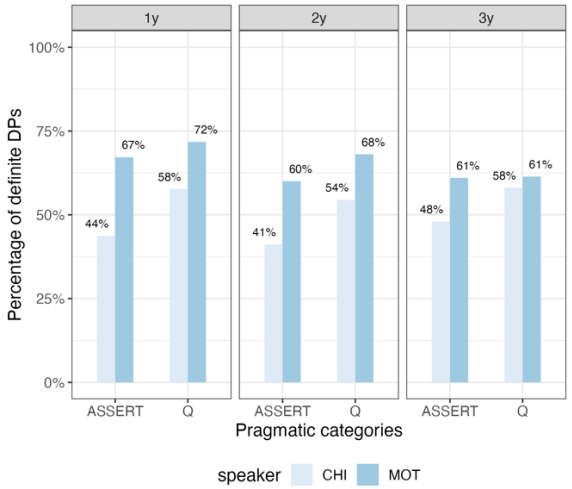

Figure 1. Percentage of definites by pragmatic categories

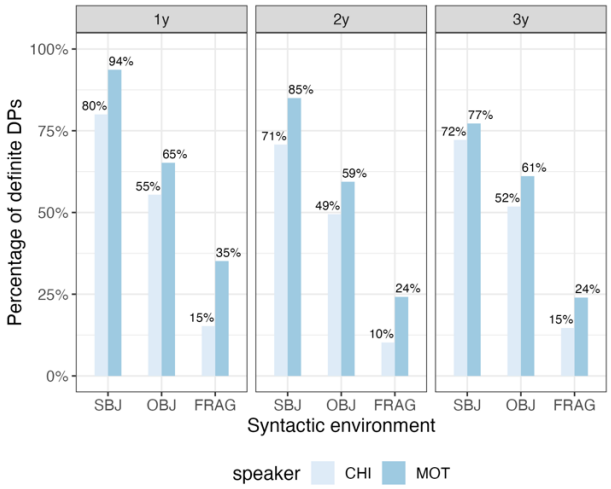

Figure 2. Percentage of definites by syntactic contexts in assertions

References

[1] Maratsos, M. (1974). Preschool children's use of definite and indefinite articles. *Child Development*, *45*(2), 446-455. [2] Schaeffer, J., & Matthewson, L. (2005). Grammar and pragmatics in the acquisition of article systems. *Natural Language and Linguistic Theory*, *23*(1), 53-101. [3] Schafer, R., & de Villiers, J. (2000). Imagining articles: What a and the can tell us about the emergence of DP. In S. C. Howell, S. A. Fish, & T. Keith-Lucas (Eds.), *Proceedings of the 24th Boston University Language Development Conference* (pp. 609-620). Cascadilla Press. [4] Karmiloff-Smith, A. (1979). *A functional approach to child language: A study of determiners and reference*. Cambridge University Press. [5] Van Hout, A., Harrigan, K., & de Villiers, J. (2010). Asymmetries in the acquisition of definite and indefinite NPs. *Lingua*, *120*, 1973-1990. [6] Wexler, K. (2011). Cues don't explain learning: Maximal trouble in the determiner system. In E. Gibson, & N. J. Pearlmutter. (eds.), *The processing and acquisition of reference* (pp. 15-42). MIT Press. [7] Choi, Y. J., Song, H. J., & Luo, Y. (2018). Infants' understanding of the definite/indefinite article in a third-party communicative situation. *Cognition*, *175*, 69-76. [8] Syrett, K., Kennedy, C., & Lidz, J. (2010). Meaning and context in children's understanding of gradable adjectives. *Journal of Semantics*, *27*(1), 1-35. [9] Brown, R. (1973). *A first language*. Harvard University Press. [10] Soderstrom, M., Blossom, M., Foygel, R., & Morgan, J. L. (2008). Acoustical cues and grammatical units in speech to two preverbal infants. *Journal of Child Language*, *35*, 869-902. [11] Suppes, P. (1974). The semantics of children's language. *American Psychologist*, *29*, 103-114. [12] Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. *Cognition*, *40*, 21-81. [13] Prince, E. F. The ZPG letter: Subjects, definiteness, and information-status. In W. C. Mann and S. A. Thompson (Eds.), *Discourse descriptions: Diverse linguistic analyses of a fund-raising text* (pp. 295-325). Philadelphia: John Benjamins. [14] Menn, L., & Gleason, J. B. (1986). Babytalk as a stereotype and register: Adult reports of children's speech patterns. In J. A. Fishman (Ed.), *The Fergusonian Impact, Volume I*. Mouton de Gruyter.