Developmental Trajectories of Vocabulary Composition Across Languages

What sorts of words do children learn first, and why? Answers to these questions provide fundamental insights into the interplay of conceptual and linguistic factors in shaping children's lexical development. One long-standing claim is that children's first words are more likely to be concrete nouns than predicates or closed class forms, so early vocabularies show a "noun bias" (Gentner, 1982). An over-representation of nouns is seen in many languages, including English (Bates et al., 1994; Caselli et al., 1995). However, languages such as Korean (Gopnik & Choi, 1995) and Mandarin Chinese (Tardif, 1999) show either less or no noun bias, depending on methodology. The primacy of nouns over other lexical classes could reflect conceptual aspects of object categories, typological factors, morphological complexity, or input frequency (Gentner & Boroditsky, 2001). Previous research has been limited in its ability to offer a comprehensive look at cross-linguistic variability in the extent of noun bias, because studies have examined only one or few languages, relied on small samples of children, or explored a single age group, making inferences problematic (cf. Bornstein et al., 2004).

The current study aggregates cross-linguistic data from the MacArthur-Bates Communicative Development Inventory (MB-CDI), a family of parent-report questionnaires adapted to several languages. Using Wordbank (wordbank.stanford.edu), a searchable database of MB-CDI data, we analyze children's productive vocabularies in thirteen languages that vary along dimensions relevant to vocabulary composition. We include 32,568 administrations of Words & Sentences MB-CDI forms on 16–36-month-old children (median per language = 1,056; range = 253–12,969). In each language, we classify vocabulary checklist items by lexical category (nouns, predicates, function words, or other) and compute the proportion of items in each category that the child is reported to produce. We then examine the trajectory of these proportions as a function of vocabulary size. Unlike previous analyses, this method controls for both overall vocabulary size and the number of items in a category.

Figure 1 shows children's proportions for each category in English and Mandarin. The diagonal indicates equal distributions of words across lexical categories; deviations from the diagonal indicate over-representation or under-representation in each category. Composition biases across all languages are shown in Figure 2. As in earlier studies, most languages show a positive noun bias, producing more nouns than would be expected by chance. But the extent of this bias varies cross-linguistically, with Mandarin and Cantonese at the low end of the continuum. Degree of noun and predicate bias are strongly negatively correlated ($r = –0.74$). Across all languages, function words are substantially under-represented, although there is considerable variability in the degree of this negative bias.

By applying large-scale data aggregation techniques, our analysis provides the most comprehensive picture of cross-linguistic variation in early vocabulary composition to date. Results indicate that a noun bias is evident in many of the languages that we examined, yet there is considerable variation in its extent. Explaining this variability – whether in terms of cultural, morphosyntactic, semantic, or other factors – is an important challenge for detailed theories of vocabulary development.

(499 words)

Figure 1: Examples of our data and models for two languages. On the left, each MB-CDI administration is plotted as a point reflecting the proportion of nouns, predicates, and function words on the form that the child is reported to produce as a function of the proportion of all three combined that the child is reported to produce; on the right, constrained polynomial curves are fit to those data, with shaded areas indicating the area between the model predictions and the diagonal (our measure of bias).

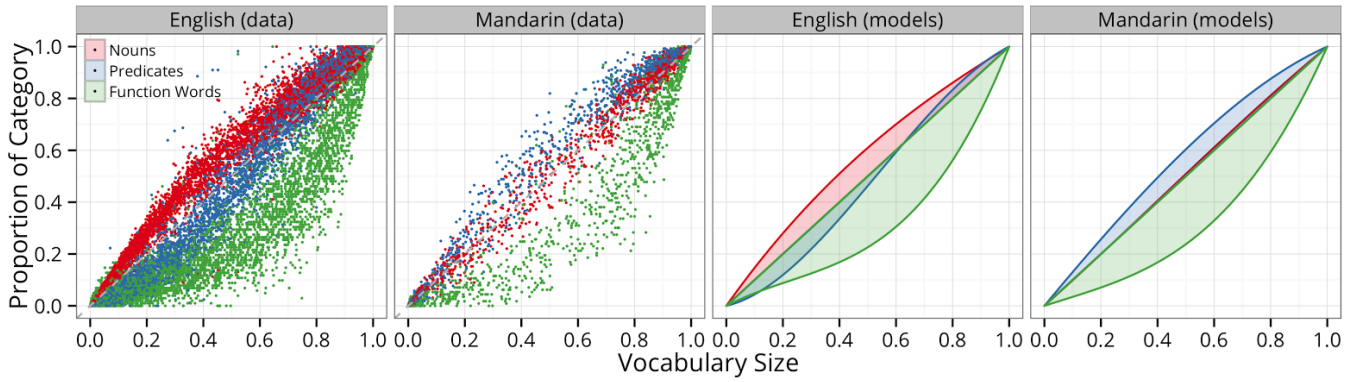
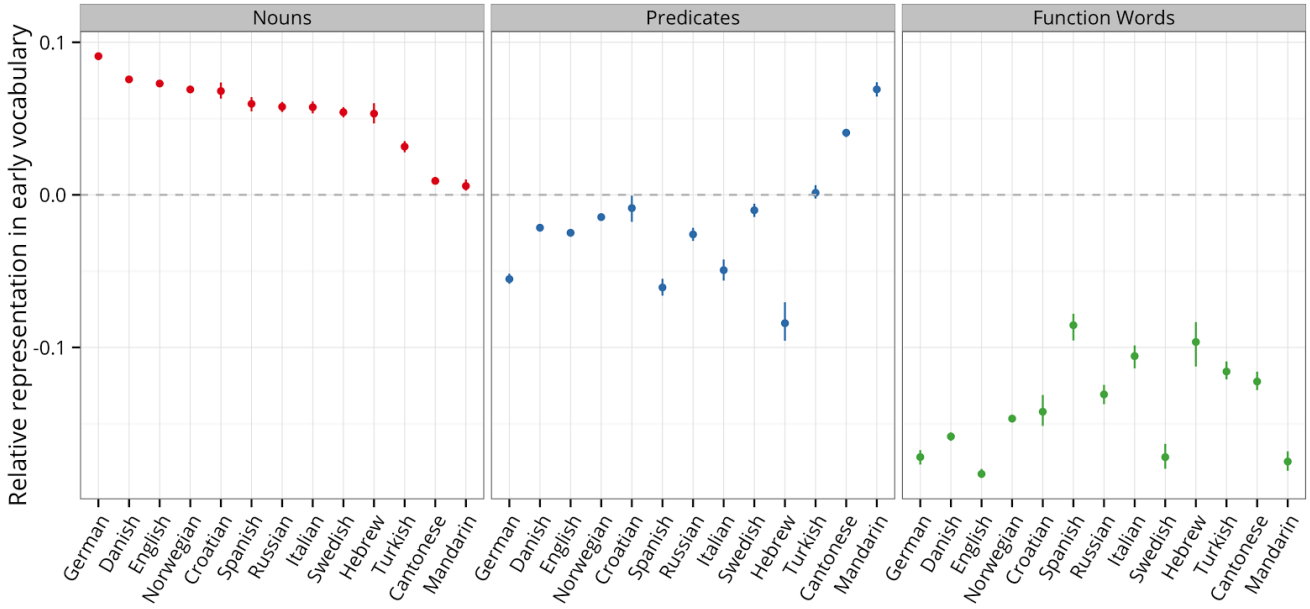


Figure 2: Bias estimates for each language and lexical category (line segments show bootstrapped 95% confidence intervals).



References

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., ... & Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of child language*, 21(01), 85-123.

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S. Y., Pascual, L., ... & Vyt, A. (2004). Cross-Linguistic Analysis of Vocabulary in Young Children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development*, 75(4), 1115-1139.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. Kuczaj (Ed.), *Language development: Vol. 2. Language, thought, and culture* (pp. 301–333). Hillsdale, NJ: Erlbaum.

Gentner, D., & Boroditsky, L. (2001). 8 Individuation, relativity, and early word learning. *Language acquisition and conceptual development*, 3, 215.

Gopnik, A., & Choi, S. (1995). Names, relational words, and cognitive development in English and Korean speakers: Nouns are not always learned before verbs.

Maital, S. L., Dromi, E., Sagi, A., & Bornstein, M. H. (2000). The Hebrew Communicative Development Inventory: Language specific properties and cross-linguistic generalizations. *Journal of Child Language*, 27(01), 43-67.

Tardif, T., Gelman, S. A., & Xu, F. (1999). Putting the "noun bias" in context: A comparison of English and Mandarin. *Child Development*, 70(3), 620-635.