Estimating "Onset Age of Production" for English Function Words Using Child Language Corpora and Growth Curve Models

Overview – Previous research has shown that content words are acquired earlier than function words (Bates et al. 1994; Braginsky et al. 2016; Frank et al. 2017). Yet, the exact trajectory and order of function word acquisition has remained relatively understudied at large scale. This is partly because age of acquisition for lexical items is often measured using Child Development Inventory (CDI) questionnaires (Fenson et al. 2007) which represent content words more than function words. On the other hand, function words have very high frequency of production which makes them easily detectable and measurable in corpus data. In this study, we use the largest available child language corpora as well as growth curve modeling to estimate the "onset of production" for English function words. Our estimates match previous reports in the literature in several cases (e.g. negation) and suggest that for a large number of function words, children start their productions as early as 12-30 months. Overall, corpus data and growth curve modeling prove promising in detecting early emergence of functional meaning in child language.

Methods – We used word tokens in the English-North America and English-UK collections of CHILDES corpora (MacWhinney 2000), accessed via the online database childes-db.stanford.edu and its associated R package {childsr} (Sanchez et al. 2019). After exclusions due to unintelligibility (N= 487,337 tokens) or missing information (N = 1,105,325 tokens), as well as age range of less than 12 and more than 72 months (N= 481,160 tokens), the collection contained 15,824,850 tokens from 1,146 children. We compiled and edited a list of about 150 English function words using the {stopword} package in R and the NLTK package in Python. We classified all tokens produced by parents and children in the corpora into content vs. function words. All tokens were annotated for the age of the child (in months) as well as whether the token was produced by parents or children. We computed the "cumulative relative frequency" of each function word by dividing its frequency up to that month by the frequency of all the words produced by children/parents up to that month.

Results – We first show that word cumulative relative frequencies have different developmental distributions for function vs. content words. For most function words, cumulative relative frequencies follow a nonlinear S curve across development with a period of accelerated growth, a period of slowdown, and a final stage of stable production. We fit separate Bayesian growth curve models (Figure 2, Right) to the cumulative relative frequency of each function word using the BRMS package (Bürkner 2017). We used growth curves with three parameters: 1. Lambda (lag): estimating the age at which production starts; 2. Miu: maximum rate of growth; and 3. A: maximum growth. Uniform priors were chosen for all three parameters. Figure 2 (Left) shows the onset of production estimates for a sample of 11 function words in our study.

Word Count: 497

References – Bates, Elizabeth, Virginia Marchman, Donna Thal, Larry Fenson, Philip Dale, J Steven Reznick, Judy Reilly, and Jeff Hartung. 1994. "Developmental and Stylistic Variation in the Composition of Early Vocabulary." J Child Lang 21 (01): 85–123.–Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. Journal of Statistical Software, 80(1). –Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. (2016). From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. In CogSci (Vol. 6). –Fenson, Larry, Elizabeth Bates, Philip S Dale, Virginia A Marchman, J Steven Reznick, and Donna J Thal. 2007. MacArthur-Bates Communicative Development Inventories. Brookes Publishing Company.– 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.–MacWhinney, Brian (2000) The CHILDES Project: The Database – Park, Tschang-Zin (1979) Some Facts on Negation: Wode's Four-Stage Developmental Theory of Negation Revisited. *JCL* – Sanchez, Alessandro, Stephan Meylan, Mika Braginsky, Kyle MacDonald, Daniel Yurovsky, and Michael C Frank (2019) Childes-Db: A Flexible and Reproducible Interface to CHILDES. *Behavior Research Methods*

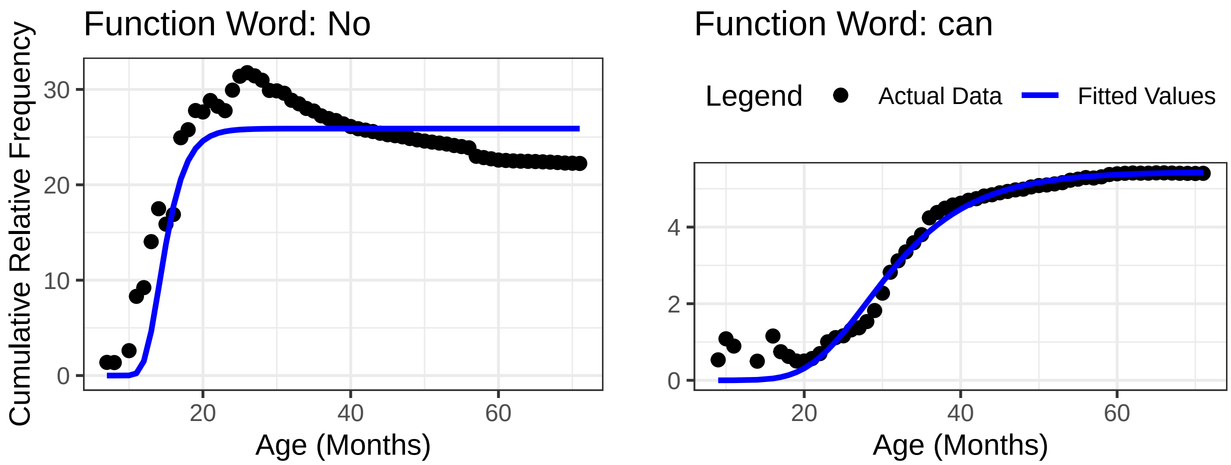

Figure 1: Cumulative relative frequency (words per thousand) for the sample words no' andcan’ in children’s productions.

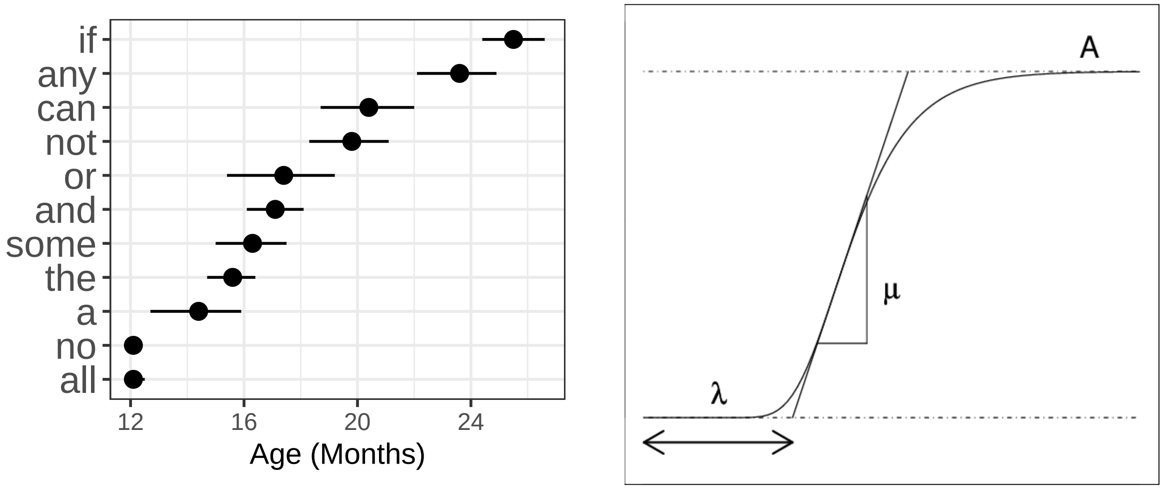

Figure 2: Estimated onset of production (lag value Lambda) for a sample of function words in the study.