**Estimating "Onset Age of Production" for English Function Words Using Child Language Corpora and Bayesian 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), but the exact trajectory and order of function word acquisition has remained relatively understudied at large scale. In this study, we use the largest available child language corpora as well as Bayesian growth curve modeling to estimate the population level onset of production for more than 100 English function words. Our estimates suggest that for the large majority of function words, the earliest age of production lies between 12-24 months of age. The results of linear regression modeling show that longer function words as measured by the number of phonemes and function words that appear in longer utterances as measured by their Mean Length of Utterance (MLU) also have later estimated onset age of production. We did not find a significant effect of frequency in child directed-speech on our estimates of onset age of production for function words. Overall these results point to early emergence of abstract functional morphemes with production limitations as the main bottleneck.

**Methods** – We used word tokens in the English-NA 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). Our collection contained 16,294,147 tokens from 1,151 children. We compiled and edited a list of about 150 English function words. 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. For each function word, we also produced its relative frequency in child-directed speech (CDS), its length in phonemes, and its MLU.

**Results** – Word cumulative relative frequencies have different developmental distributions for function vs. content words overall. 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. Figure 1 (bottom right) shows an example curve for the modal "can". We fit separate Bayesian growth curve models 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 75 function words in our study. A linear regression showed a significant effect of function word length in phonemes ($\beta = 2.65$, $t = 3.182$, $p = 0.002$) and MLU ($\beta = 2.07$, $t = 3.137$, $p = 0.002$), but not relative frequency in CDS ($\beta = -40.09$, $t = -0.214$, $p = 0.83$).

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). –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*

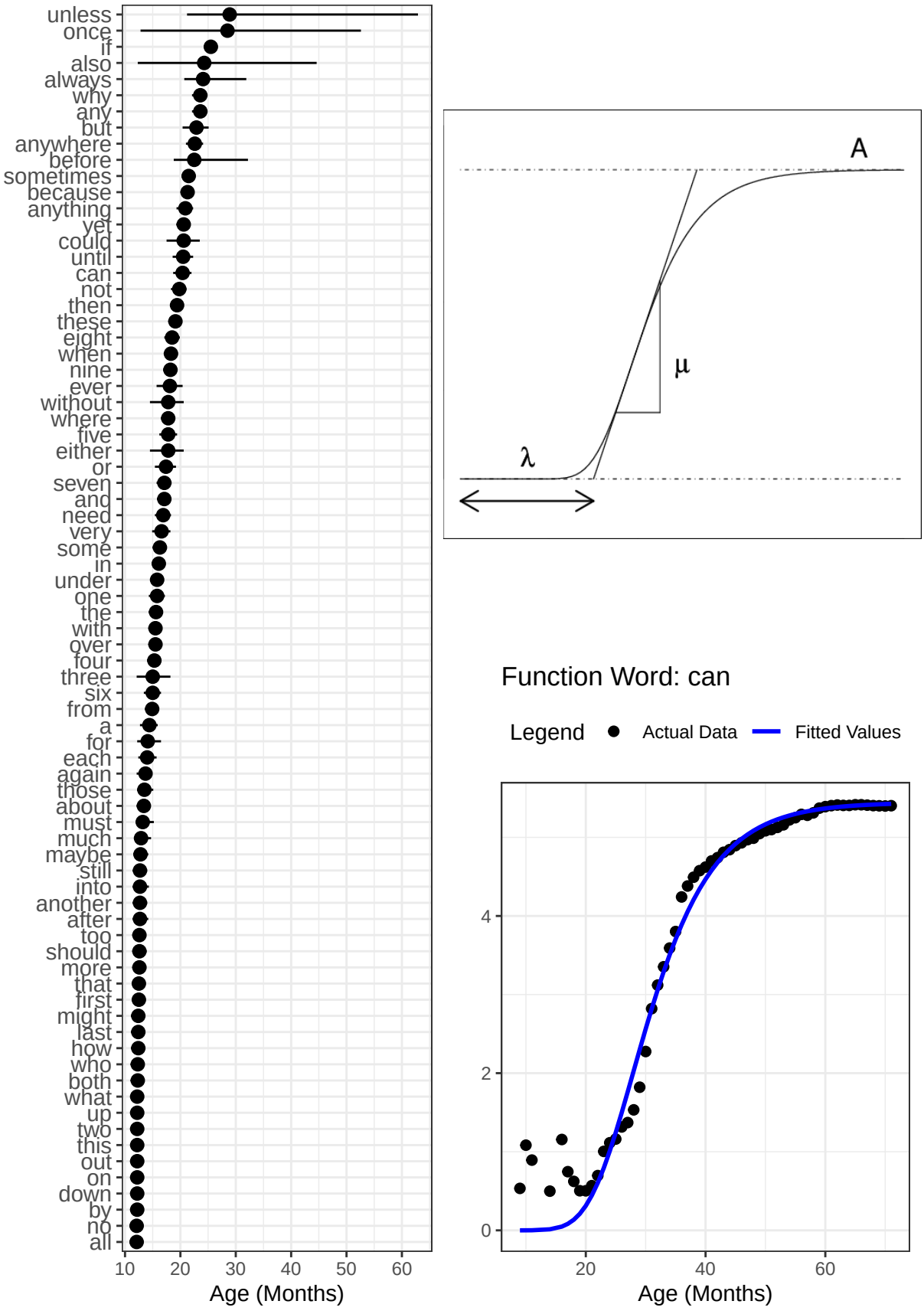


Figure 1: Estimated onset of production (lag value Lambda) for a sample of function words in the study.