Corpus-Based Population-Level Estimates of Onset Age-of-Acquisition for English Function Words

Masoud Jasbi, Aaron Pilapil, and Debbie Odufuwa

1. Introduction

Function words are a small group of highly frequent words that help build grammatical structure and convey abstract compositional meaning. This class includes pronouns (e.g. *she*, *you*, *us*, *their*), question words (e.g. *how*, *when*, *what*, *who*), prepositions (e.g. *in*, *on*, *at*, *from*), quantifiers (e.g. *all*, *some*, *many*, *none*), logical connectives (*and*, *or*, *not*, *if*), temporal adverbials (e.g. *always*, *usually*, *sometimes*, *never*), and modals (e.g. *might*, *can*, *should*, *may*), among other subclasses. A large body of research in language acquisition has examined children's comprehension and production of individual function words, testing specific hypotheses about their syntactic and semantic development. However, general developmental trends for function words as a whole, and in comparison with each other, remain largely unexplored. This paper uses corpus data and statistical modeling to provide a bird's-eye view of function word production in child language. We focus specifically on when children begin producing function words and the order in which different function words emerge in preschool children's speech, asking what factors determine this acquisition order.

Previous research has shown that function words are acquired later than content words (Braginsky, Yurovsky, Marchman, & Frank 2016; Bates, Marchman, Thal, Fenson, Dale, Reznick, Reilly, Hartung 1994; Frank, Braginsky, Yurovsky, & Marchman 2021). However, the precise age, trajectory, and developmental order of function word acquisition remain relatively understudied. This gap exists partly because standard measures of age of acquisition are better suited to content words. These measures typically fall into three categories:

1. Rating surveys with adults (Carroll & White 1973; Gilhooly & Logie 1980; Kuperman, Stadthagen-Gonzalez, & Brysbaert 2012; Scott, Keitel, Becirspahic, Yao, Sereno 2019)
2. Picture naming tasks with children (Morrison, Chappell, & Ellis 1997)
3. Parent reports using the MacArthur-Bates Communicative Development Inventories (MB-CDIs)(Fenson, Marchman, Thal, Dale, Reznick, & Bates 2006; Frank, Braginsky, Yurovsky, & Marchman 2021)

---

* Department of Linguistics, University of California Davis.

The first method asks adults (typically undergraduates) to estimate when they first encountered words from a presented list. However, these lists generally contain only content words such as nouns, verbs, and adjectives. The second method presents children at various ages with pictures of objects to elicit naming responses. While this approach provides objective age-of-acquisition data for concrete object names, it cannot readily assess abstract nouns, verbs, adjectives, or function words. The third method uses the MB-CDI[1] questionnaires in which parents check off words their child understands or produces from a standardized list. The lists contain many functional classes and function words including pronouns (e.g., *his, her, it, I, me, mine, my, that, this, you, your*), question words (*how, when, what, where, who, why*), prepositions (*away, back, down, in, inside, off, on, out, there, under, up*), and quantifiers (*all, another, more, none, not, other, same, some*). However, they also lack many function words such as disjunction (e.g. *or*), some quantifiers (e.g. *few*, *many*, *most*), modals (e.g. *might*, *maybe*, *must*), adverbs of frequency (e.g. *always*, *sometimes*), among other function words (e.g. *again*, *only*). Furthermore, it is not clear if parents' recollections of their children's comprehension and production of function words are as accurate and reliable as their recollections for content words. To date, there is no objective measure of age-of-acquisition, for children's production of function words.

This study provides a novel method for estimating age of acquisition for function words using child language corpora. Because function words occur with high frequency, they can be reliably detected even in small-to-medium sized corpora. We leverage the largest available collection of child language corpora combined with growth curve modeling to trace the developmental trajectory of function word production and estimate the earliest population-level age at which children begin producing different English function words. To understand what drives these acquisition patterns, we examine four potential predictors of our corpus-derived age-of-acquisition estimates: frequency in child-directed speech, concreteness ratings, number of phonemes, and average utterance length of the function word. Preliminary results from our linear model reveal that only phoneme count and average utterance length significantly predict the age at which function words are acquired.

2. Methods

2.1. Corpus Data Preprocessing

Word tokens from the English–North America and English–UK collections of the CHILDES corpora (MacWhinney 2000) were accessed via the CHILDES database (`https://childes-db.stanford.edu`) and using the `childesr` package in R (Sanchez et al. 2019). After excluding tokens with missing age information ($N = 1,114,702$) and tokens produced outside the target age range of 0 to 72 months ($N = 489,814$), the final dataset consisted of $16,294,147$ tokens produced by 1,151 children. All remaining tokens were annotated with the child's

[1] `https://mb-cdi.stanford.edu`

age in months and the speaker role, collapsed into *child* (target child) and *parent* (father and mother) categories.

A preliminary list of approximately 150 English function words was compiled using the `stopwords` package in R and the Natural Language Toolkit (NLTK) in Python. This list was manually reviewed and expanded (Table 1). All text was lowercased to standardize string comparisons. Tokens whose part-of-speech tags indicated babbling, unintelligible speech, or non-lexical vocalizations were retained in the corpus rather than excluded, so that they contributed to overall count of children's productions. These tokens were treated as valid communicative attempts and were included in denominator counts when computing relative frequencies. As far as we know removing these tokens do not alter the results and conclusions presented in this paper. Nine obvious transcription errors (e.g., lexical items such as *annotation* appearing before age two) were also removed. Contractions were handled by reducing tokens to their stems, with the exception of negative morphemes. For example, *he's* was treated as *he*, simply because *'s* is ambiguous as a token and we did not whish to track its development. However, *can't* was segmented into *can* and *n't*. The form *won't* was treated as a lexical exception and retained as a single token.

Table 1: A preliminary list of function words included in the study, organized by linguistic category.

Category	Function Words
Negation	no, not, n't
Conjunctions	and, or, but, if, although, because, therefore
Determiners	the, a, an, this, that, these, those
Pronouns	I, you, we, he, she, they, it, me, us, him, her, them, my, your, our, his, their, its, mine, yours, ours, hers, theirs
Auxiliaries	am, is, are, was, were, be, been, being, have, has, had, do, does, did
Modals	can, could, may, might, must, shall, will, would, should, ought, maybe
Frequency Adverbs	always, usually, often, sometimes, seldom, never
Quantifiers	all, none, some, any, most, few, many, several, both, each, every
Prepositions	in, on, at, to, of, from, with, without, about, by, for, under, over, between, into, through, against
wh-words	who, whom, whose, what, when, where, why, how
NPIs	any, ever, yet, anyone, anything, anywhere
Temporals	while, after, before, then, until, since, whenever, during
Other	again, too, also, another, other, others, still, only, just, even, indeed, either, neither, whether, as, else, almost, already, except, for, from, instead, such, with, without, about, by, very, unless, to, of, would, at

To characterize developmental change while mitigating month-to-month sampling variability, cumulative relative frequency trajectories were computed over age. Token counts were aggregated within monthly age bins for each speaker group and function word, and monthly relative frequencies were obtained by dividing the specific function-word token-counts by the total number of word tokens produced by that speaker in the same age bin. Cumulative relative frequency for a function word w up to age bin t was then computed by ordering observations by age and calculating the ratio of cumulative counts of the function word w up to t, divided by the cumulative total word token counts up to age t. Cumulative relative frequencies were rescaled to a per-thousand-token metric for interpretability.

$$y_{t,w} = \sum_{i \leq t} \mathrm{freq}_{w,i} / \sum_{i \leq t} \mathrm{total}_i \tag{1}$$

2.2. Statistical Modeling

We use two types of statistical models to study children's production of function words: 1. non-linear growth curves to model the trajectory of function word production in child language and estimate the onset age of acquisition for each function word; and 2. a linear regression to find significant predictors of the estimated onset age of acquisition for different function words. Below we provide more technical details on each model.

2.2.1. Growth-Curve Models

We use developmental growth curve analysis to model population-level trends in children's production of different function words (Kemper, Rice, & Chen 1995; Van Veen, Evers-Vermeul, Sanders, & Van den Bergh 2009). Specifically, we employ a re-parametrization of the Gompertz growth curve function (Gompertz 1825) originally used for modeling bacterial growth by Zwietering, Jongenburger, Rombouts, & van 't Riet (1990) shown in Equation (2) below. In this equation, $y_{t,w}$ represents the cumulative relative frequency of a function word w at age t. The letter e stands for Euler's number, exp stands for the exponential function (e^x), and the three main parameters of the model are: (1) the lag phase (λ), representing the period before growth begins; (2) the maximum growth rate (μ); and (3) the asymptote (A), representing the maximum level of growth. Figure (1) shows the shape of this modified Gompertz growth curve with the three parameters listed above. In the context of children's function word production, these parameters have specific developmental interpretations: the lag phase (λ) estimates the earliest age at which children begin producing the word, the maximum growth rate (μ) captures the steepest increase in production relative to other words in their lexicon, and the asymptote (A) reflects the stable production level often observed in adult usage. The parameter of primary interest in our study is the lag phase (λ), which estimates the earliest age at which there is corpus evidence for children producing a specific function word.

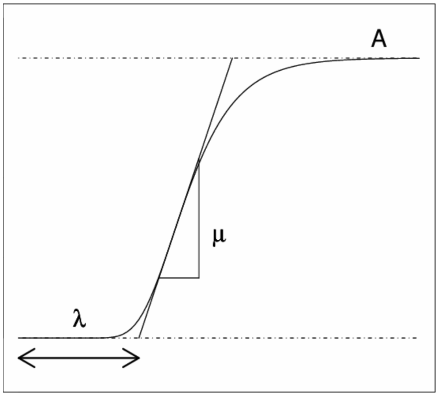

Figure 1: Modified Gompertz growth curve with three parameters: lag phase (λ), maximum growth rate (μ), and the asymptote (A).

$$y_{t,w} = A \times exp(-exp(\frac{\mu \times e}{A}(\lambda - t) + 1)) \quad (2)$$

We fit separate population-level Bayesian growth curves to the cumulative relative frequency of each function word using the BRMS package (Bürkner 2017). Uniform priors were chosen for all three parameters of the model. The lag parameter (λ) was bounded between (0, 72), the entire age range of children in our study. The maximum growth rate (μ) was bounded between (0, 10). We did not find any curves in which growth rate was close to our upper bound. Since the upper asymptote was highly variable per function word, the lower bound for this parameter was set to 0 and the upper bound was chosen as appropriately large to allow for the model to estimate the upper asymptote. All models ran with 4 chains, each for 10000 iterations (5000 warm-up) and converged with $\hat{R} = 1$.

2.2.2. Linear Model

We also collected data on four predictors hypothesized to influence the age of acquisition for function words. First, we used the available token-level data to calculate the frequency of all our function words in child-directed speech. Second, we transcribed all the function words using the International Phonetic Alphabet and computed the length of function words in phonemes. Third, we computed the average length of utterance for each function word. This measure tells us whether a function word typically appears in longer or shorter utterances. To avoid confusion with "Mean Length of Utterance", we call this measure "the average utterance length" for function words. The average utterance length is an important predictor because early in development, children's utterances are shorter and therefore function words that can appear in shorter utterances have a higher chance of being produced by children. Finally, we used the lexical concreteness measure provided in Byrsbaert et al. 2014, obtained using ratings from more than 4,000 study participants. The intuition is that some function words such as *up* or *down* refer to more concrete concepts than others like *the* and *each*, and therefore, may have a higher chance of being learned earlier in development. An additive linear regression model shown below examined the independent contributions of each of these factors to the corpus-derived age of acquisition estimates for function words.

$$\lambda_w = \beta_0 + \beta_1(\text{phoneme length}) + \beta_2(\text{frequency}) \\ + \beta_3(\text{average utterance length}) + \beta_4(\text{concreteness}) + \varepsilon \quad (3)$$

All code, data, and results for this study are available on the study repository at `https://github.com/jasbi/childes_function_words`.

3. Results

Figure 2 presents the percentage of function word tokens in parents' and children's productions at different monthly age bins. First, the results show that regardless of their children's age, roughly half of the tokens in parents' speech are function words. Overall, the production of function words in parents speech is stable across children's development. Children, on the other hand, start with producing no function words, and slowly increase their productions until they reach the stable parent level of 50% at about 40 months. A Bayesian Gompertz growth curve model estimated the onset age of production for all function words as 13.4 months (CI = [12.82, 13.95]), suggesting that children start producing some function words as early as a month and a half after their first birthday. Overall, the results in Figure 2 show that the age range of 1-3 years is an important period for the acquisition of function words.

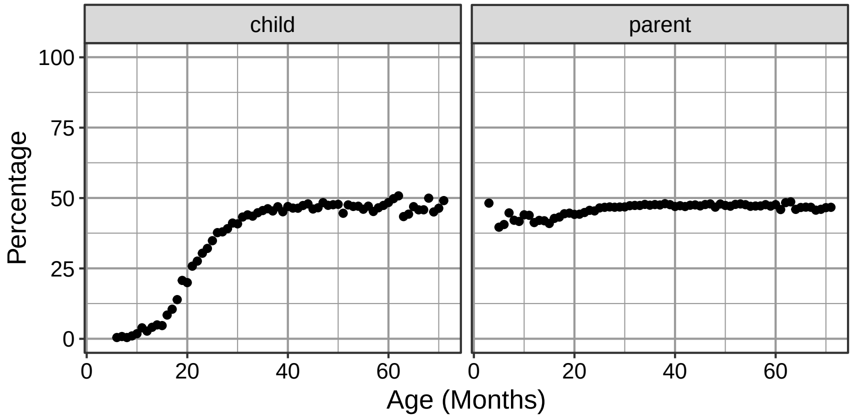

Figure 2: The percentage of function word tokens in parents' speech (right) and children's speech (left). The x-axis represents the age of children in months.

Figures 3 and 4 show the growth curves for a sample of 16 function words. The solid (gray) lines in the plots show the Gompertz growth curve model fit to the cumulative relative frequency data. Overall, Gompertz growth curves provided good model fit to the developmental trends of many English function words. However, as the data for "more" shows, some function words provided more complex non-linear trends that did not exactly match the growth pattern described by Gompertz curves. Due to space limitations we cannot present the growth curves for all the function words we studied in this paper. You can visit the study repository linked in the methods section to see the results for all other function words.

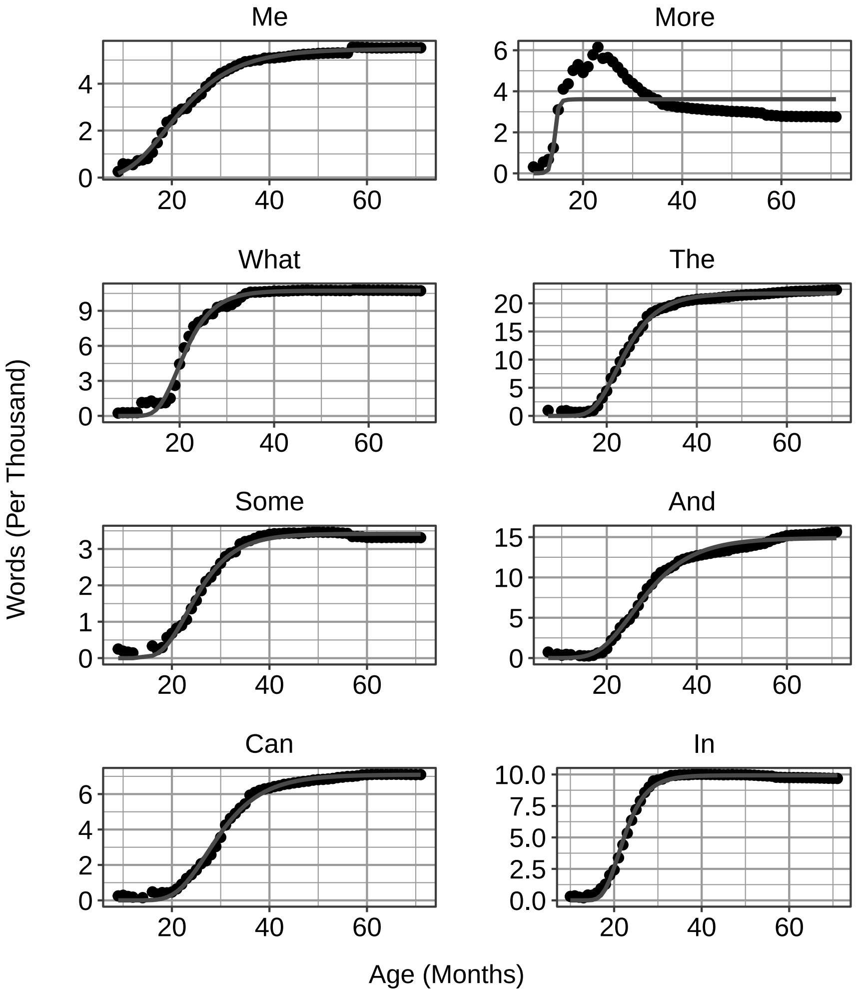

Figure 3: Growth curves for a sample of 8 function words. The x-axis shows the child age in months, the y-axis shows the cumulative relative frequency. The solid (gray) line is the Gompertz growth curve model fit to the data.

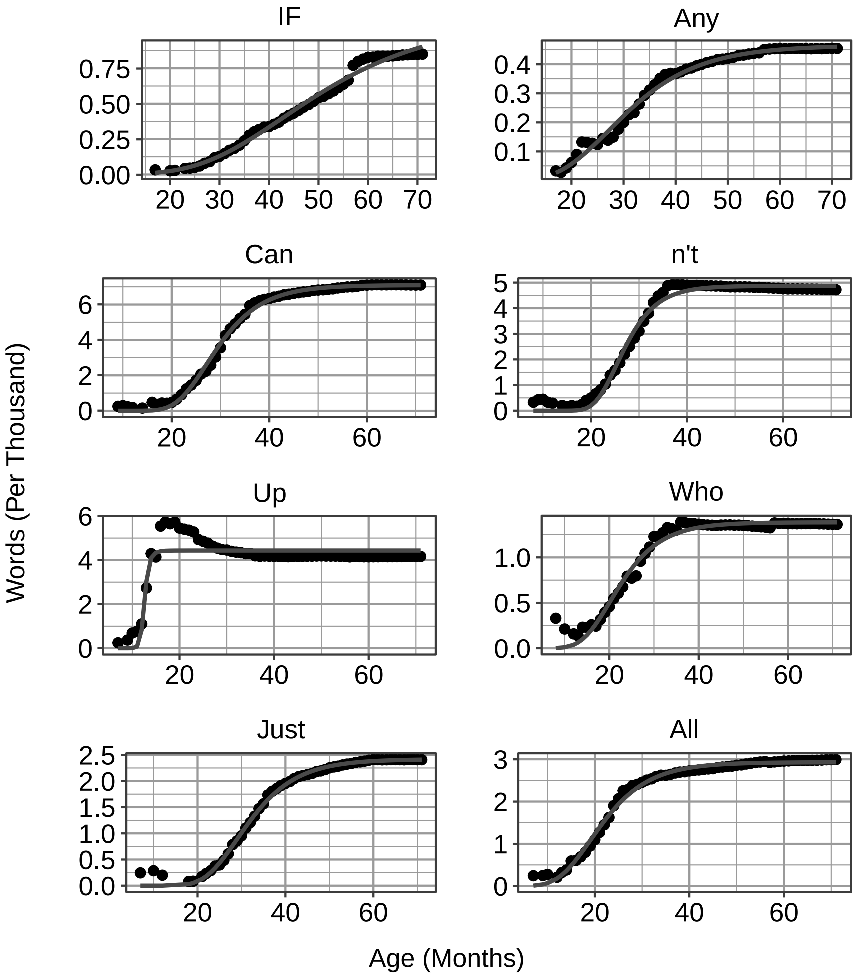

Figure 4: Growth curves for a sample of 8 function words. The x-axis shows the child age in months, the y-axis shows the cumulative relative frequency. The solid (gray) line is the Gompertz growth curve model fit to the data.

Figure 5 presents the lag phase estimates (λ) from our growth curve models for a sample of 60 English function words with their 95% credible intervals. As a reminder, the lag phase estimates represent the earliest age at which corpus evidence suggests at least some child or children can produce a specific function word. We use this estimate as our corpus-derived onset age of acquisition for function words. As the figure shows, almost all function words in this sample, with the exception of "except", are produced by some children in the 12-27 months age range. In other words, most function words in our sample could be produced quite early in children's linguistic development.

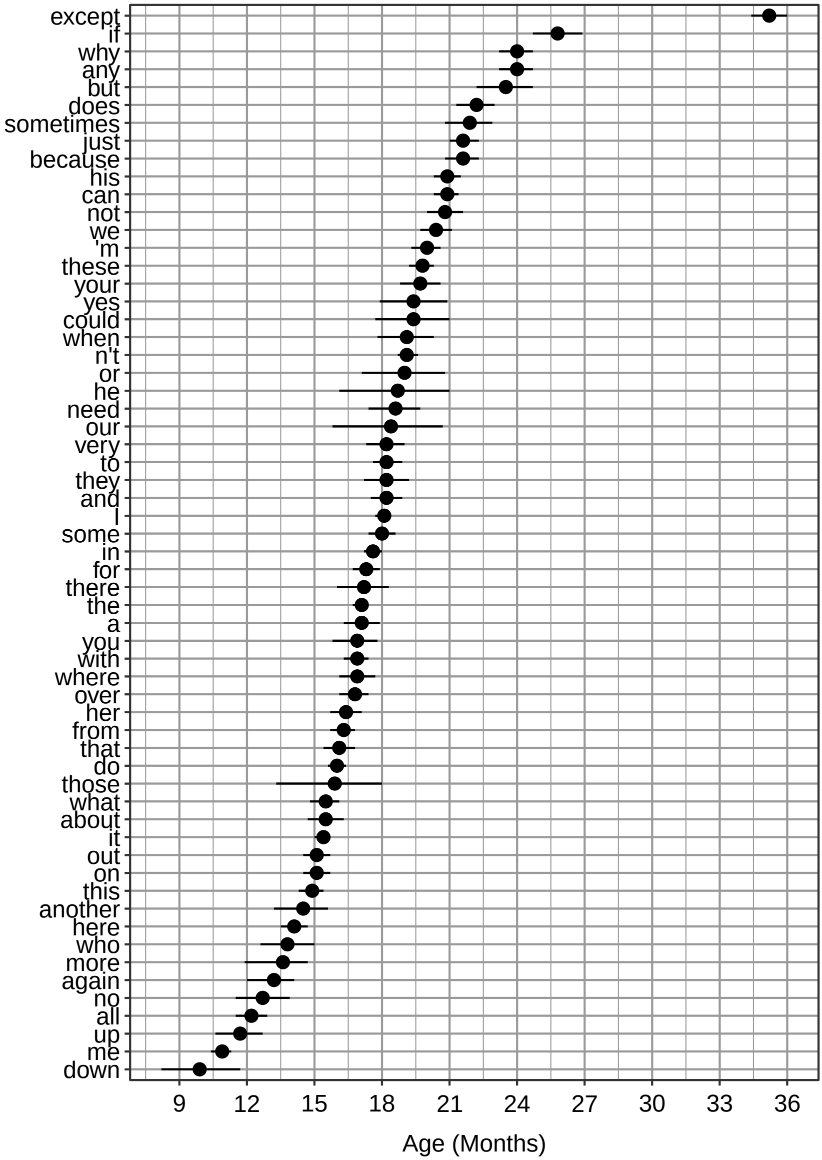

Figure 5: Lag phase (λ) estimates for 60 English function words. The line range represents 95% credible intervals.

Many function words that emerge early in corpus data such as “me”, “up”, “no”, and “more”, are function words that can be produced as single word utterances in response to questions by parents or as commands/requests to parents. This suggests that an important factor in children’s production of function words may be the communicative utility of a function word given children’s productive limitations. For example, a word such as “no” or “more” can communicate important information for the child, for example a rejection or a request for more food, while requiring the child to produce little. On the other hand, a function word such as “if” requires the child to be able to combine several other words before it can be communicatively useful. This trade-off of communicative utility and children’s production bottleneck may be an important predictor of the order in which function words emerge in childrens speech.

We also examined individual-child level trends to better understand the diversity and variability of developmental pathways for function words. A sample of 10 function words out of our list of about 150 English function words were selected. We then selected ten children who produced the most tokens between the ages of 12-48 months out of the English - North America and English - United Kingdom collections of CHILDES. These children were selected on the basis of production of tokens in order to ensure there was enough data to produce developmental curves. Figure 6 shows the cumulative relative frequency of the word “and” in the speech of the 10 children selected for individual-level analysis. The solid black line in the plot shows the aggregate development across all 10 children. The comparison of individual level cumulative relative frequencies for each child and the aggregate trend (black line) suggests considerable variability in the production of function words in children. Similar type of variability in the developmental trends in production can be observed in the other sample of function words provided in Figure 7. These results suggest that the aggregate developmental curves and the onset age of acquisition estimates should be treated as population level trends and estimates that do not easily translate to individual level trends and estimates. Individual children likely show much more variable production onsets and developmental curves for each function word.

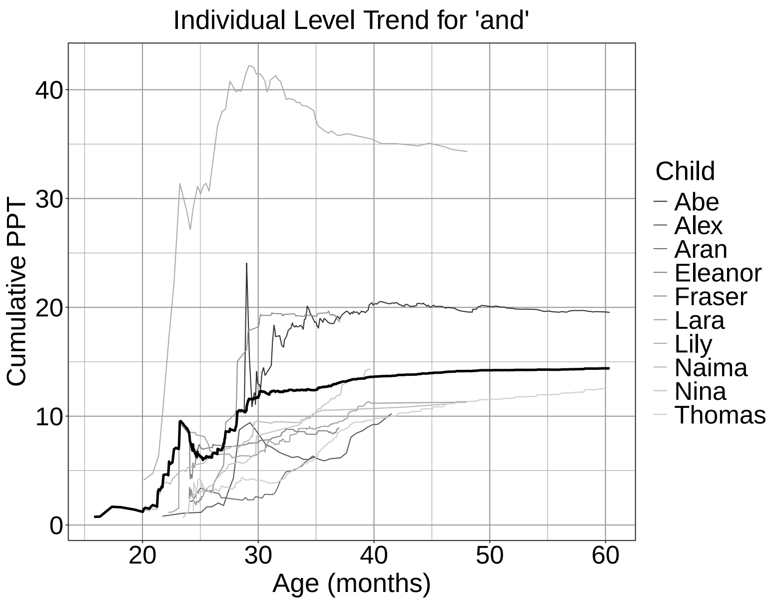

Figure 6: Cumulative relative frequency for the word "and" in the speech of 10 sample children with the aggregate trend denoted in black.

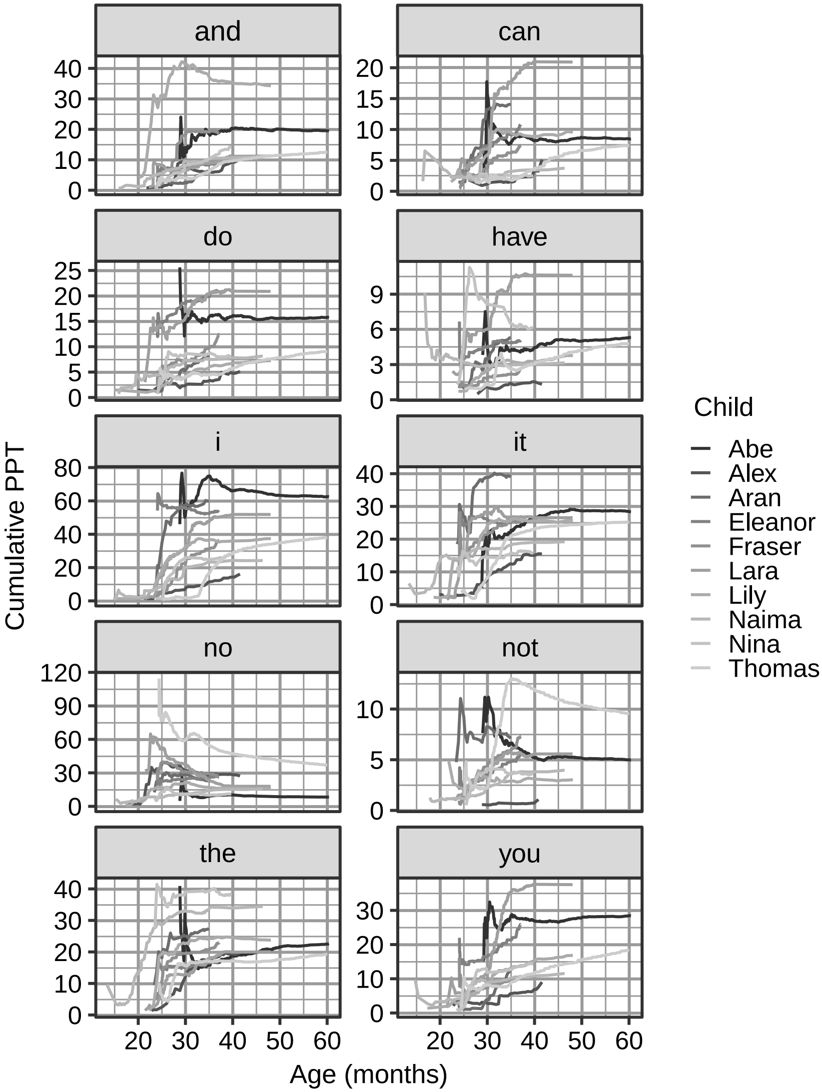

Figure 7: Cumulative PPT trajectories of ten function words for ten children who produced the most tokens between 12-48 months of age.

Finally, our preliminary investigation of factors predicting function word age of acquisition suggested that only the length of a function word in phonemes and the average utterance length for a function word may influence its onset age of acquisition. An additive linear regression found the length of a function word in phonemes ($\beta = 1.06$, $p < 0.05$) and the average utterance length for a function word ($\beta = 0.97$, $p < 0.001$) to be significant predictors, while neither frequency in child directed speech ($\beta = -5.8$, $p = 0.91$), nor concreteness ratings were found significant ($\beta = -0.88$, $p = 0.13$). While these results are interesting, it is important to emphasize their preliminary nature and that a wide variety of other factors need to be considered before we can make any definitive conclusions. For example, it is important for future research to consider the role of syntactic category of a function word in predicting its age of acquisition.

4. Conclusion

Age of acquisition for lexical items is typically measured via adult retrospective surveys or parent reports. While such measures are often suitable for content words, function words present a challenge: age of acquisition data are either missing or the reliability of second-hand reports for function words remains unestablished. To address this gap, we provide the first corpus-derived measure of age of acquisition for English function words. Using Bayesian Gompertz growth curve modeling, we estimated the earliest age at which corpus data document the production of 59 English function words. Our analysis revealed that nearly all function words examined were first produced by some children between 12 and 27 months of age. We further investigated whether four factors could predict these corpus-derived age of acquisition estimates. Linear regression analysis showed that shorter function words and function words produced in shorter utterances appeared earlier in children's speech. However, neither function word frequency in child-directed speech nor concreteness ratings emerged as significant predictors of function word age of acquisition.

We emphasize that the results presented here are preliminary. Several limitations point to important directions for future research. First, our analysis is restricted to a subset of English function words; extending this approach to additional function words is necessary for a more comprehensive picture. Some function words exhibit sparse production data in early development, which poses challenges for the growth curve modeling approach we employed—addressing these sparse data patterns will require methodological refinements. Second, the corpus-derived age of acquisition estimates reported here require validation against other available measures, particularly parent reports. Finally, our linear regression analysis examined only a limited set of predictive factors; a fuller understanding of function word acquisition will require consideration of a broader variety of linguistic and cognitive variables.

References

Bates Elizabeth, Marchman Virgina, Thal Donna, Fenson Larry, Dale Philip, Reznick J. Steven, Reilly J, Hartung Jeff. (1994) Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*. 21(1):85-123. doi: 10.1017/s0305000900008680.

Braginsky Mika, Yurovsky Daniel, Marchman Virginia A., & Frank Michael C. (2016). From uh-oh to tomorrow Predicting age of acquisition for early words across languages. Proceedings of the Annual Meeting of the Cognitive Science Society, 38. Retrieved from https://escholarship.org/uc/item/9sn4p5f9

Brysbaert Marc, Warriner Amy Beth. & Kuperman Victor. Concreteness ratings for 40 thousand generally known English word lemmas. Behav Res 46, 904–911 (2014). https://doi.org/10.3758/s13428-013-0403-5

Bürkner Paul-Christian. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Carroll John B.. & White Margaret White N. (1973) Age-of-acquisition norms for 220 picturable nouns. Journal of Verbal Learning and Verbal Behavior. 12, 563-576.

Elsherif M. M., Preece, E., & Catling, J. C. (2023). Age-of-acquisition effects: A literature review.Journal of Experimental Psychology: Learning, Memory, and Cognition, 49(5), 812–847. https://doi.org/10.1037/xlm0001215

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2006). MacArthur-Bates Communicative Development Inventories, Second Edition (CDIs) [Database record]. APA PsycTests.

Frank, Michael C., Mika Braginsky, Daniel Yurovsky, & Virginia A. Marchman (2021). Variability and Consistency in Early Language Learning: The Wordbank Project. The MIT Press. https://doi.org/10.7551/mitpress/11577.001.0001

Gilhooly, K.J., Logie, R.H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. Behavior Research Methods & Instrumentation 12, 395–427. https://doi.org/10.3758/BF03201693

Gompertz, Benjamin (1825). ”On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies”. Philosophical Transactions of the Royal Society of London. 115: 513–585.

Kemper Susan, Rice Karla, & Chen Ying-Ju. (1995). Complexity metrics and growth curves for measuring grammatical development from five to ten. First Language, 15(44), 151–166. https://doi.org/10.1177/014272379501504402

Kuperman Victor, Stadthagen-Gonzalez Hans & Brysbaert Marc (2012). Age-of-acquisition ratings for 30,000 English words. Behav Res 44, 978–990. https://doi.org/10.3758/s13428-012-0210-4

Morrison Catriona M., Chappell Tameron D., & Ellis Aandrew W. (1997). Age of Acquisition Norms for a Large Set of Object Names and Their Relation to Adult Estimates and Other Variables. The Quarterly Journal of Experimental Psychology Section A, 50(3), 528-559. https://doi.org/10.1080/027249897392017

Michael J. Panik. (2014). Growth curve modeling : theory and applications, 1st ed. edition. John Wiley & Sons, Inc., Hoboken, New Jersey.

Scott Graham G, Keitel Anne, Becirspahic Marc, Yao Bo, Sereno Sara C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. Behav Res Methods. 51(3):1258-1270. doi: 10.3758/s13428-018-1099-3. PMID: 30206797; PMCID: PMC6538586.

Van Veen Robert, Evers-Vermeul Jacqueline, Sanders Ted, & Van den Bergh Huub. (2009). Parental input and connective acquisition: A growth curve analysis. First Language, 29(3), 266–288. https://doi.org/10.1177/0142723708101679

Zwietering MH, Jongenburger I, Rombouts FM, van ’t Riet K. (1990) Modeling of the bacterial growth curve. Applied Environmental Microbiology. 56(6):1875-81.

Proceedings of the 50th annual Boston University Conference on Language Development

edited by Romi Hill, Xinhan Jiang,
Danutham Worapipat, and Aditya Yedetore

Cascadilla Press Somerville, MA 2026

Copyright information

Proceedings of the 50th annual Boston University Conference on Language Development

ISSN 1080-692X
ISBN 978-1-57473-047-0 (2 volume set, paperback)

Ordering information

To order a copy of the proceedings or to place a standing order, contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, sales@cascadilla.com, www.cascadilla.com