PROJECT PROPOSAL: REASSESSING EARLY LANGUAGE ACQUISITION DATASETS USING SUPPORT ESTIMATION TECHNIQUES

SASHA CUI

Abstract. We propose to re-evaluate estimates of children's language knowledge by applying advanced support estimation techniques to CHILDES data. Traditional analyses typically underestimate a child's vocabulary, grammatical constructions, and speech acts due to undersampling and methodological shortcomings. By implementing refined estimators such as the Good–Turing, Chao, and bootstrap methods, our study aims to generate more accurate and robust estimates of children's linguistic repertoires. This interdisciplinary approach not only validates modern statistical methods but also challenges prevailing assumptions in child language acquisition, offering novel insights with significant implications for both linguistic theory and practical applications.

1. Research Plan

1.1. Research Objectives.

- **Primary Objective:** Demonstrate that existing estimates of children's language knowledge (vocabulary, unique grammatical constructions, and novel speech acts) are underestimates due to undersampling and methodological shortcomings.
- **Secondary Objectives:**
  - Validate the effectiveness of various support estimation methods when applied to linguistic datasets.
  - Identify potential pitfalls and errors in previous statistical analyses within child language research.
  - Provide a revised, more accurate account of children's linguistic repertoires that challenges conventional wisdom.

1.2. Methodology.

1.2.1. *Data Acquisition:* Utilize publicly available CHILDES corpora, selecting a diverse sample that represents different ages, language backgrounds, and conversational contexts. Preprocess the data (tokenization, lemmatization, etc.) and categorize it by linguistic features such as vocabulary, grammatical constructions, and speech acts.

1.2.2. *Statistical Analysis & Support Estimation:*

- Review current support estimation techniques (e.g., Good-Turing, Chao1, and bootstrapping methods) and justify their selection based on their robustness in undersampled ecological datasets.
- Implement these techniques using Python to generate revised estimates for:
  - Total vocabulary size.

Date: March 28, 2025.
I discussed this with Mike Tomasello and Linjun Zhang.

  - Diversity of grammatical constructions.
  - Frequency and variety of novel speech acts.
- Perform comparative analyses with published estimates and conduct sensitivity analyses to assess how sampling variations affect the support estimates.

1.2.3. *Validation & Robustness Checks:* Employ resampling techniques (e.g., bootstrap methods) to validate the stability of the support estimates. Identify and document sources of error in both the re-implemented methods and previous studies, focusing on the impact of dataset size, frequency of rare events, and method-specific biases.

1.3. **Expected Outcomes.** We anticipate this project will yield both **methodological advances** and **substantive findings**:

- **Novel Findings:** Revised estimates suggesting that children have significantly larger repertoires of words, grammatical constructions, and speech acts than previously reported. These results will also highlight systematic errors in earlier estimation efforts.
- **Validated Estimation Procedures:** A robust method (or software tool) for estimating a child's total vocabulary from transcript data, complete with uncertainty quantification. This tool will be directly useful to researchers conducting corpus studies in language acquisition.
- **Comparative Evaluation of Estimators:** A real-world testbed for evaluating advanced support estimation methods. By applying these techniques to child language data, we can identify their strengths and limitations, and document which methods perform best under varying data conditions (e.g., small vs. large sample sizes).
- **Implications for Linguistics:** A more complete and empirically grounded view of children's linguistic capabilities, which could challenge prevailing theories of language acquisition and stimulate further research into methodological rigor and developmental variation.
- **Exploration of Individual Differences:** Estimation may uncover differences in language use that are not apparent from raw counts alone. For instance, two children with similar observed vocabularies may differ substantially in estimated total vocabulary size, pointing to variation in language exposure, talkativeness, or conversational context.

1.4. **Feasibility & Timeline.**

1.4.1. *Feasibility:*

- CHILDES data is readily available and well-documented.
- Established support estimation methods can be efficiently implemented using existing programming libraries.
- The project is designed to be manageable within a 0.5-1 year research cycle.

1.4.2. *Timeline:*

Month 1: Literature review, dataset selection, and data preprocessing.
Months 2-3: Implementation of support estimation methods and preliminary analyses.
Months 4-5: Comparative analyses with existing studies; sensitivity and robustness checks.
Month 6: Synthesis of findings, manuscript preparation, and submission for publication.

1.5. **Significance & Impact.** This study promises to deliver novel insights that question long-held assumptions about children's language capacity. By providing more accurate estimates of vocabulary, grammatical constructions, and speech acts, the research could reshape debates on language acquisition and inform new theoretical models. Moreover, the project serves as a case study for the importance of proper support estimation in undersampled datasets, potentially influencing best practices in both statistical methodology and interdisciplinary applications.

1.6. **Expected Dissemination.** The primary product will be a research article. Given the interdisciplinary nature of the work (spanning NLP, statistics, and child language), a suitable target journal is the **Journal of Child Language (JCL)** or a similar venue that appreciates quantitative approaches to language development. JCL has previously published works on large datasets (e.g., the introduction of Wordbank itself by *Frank2017*) and on methodology for vocabulary assessment, making it a fitting choice. Alternatively, we could consider *Language Learning* or *Developmental Science* for a more cognitive science audience, but our current plan is to aim for JCL as it squarely targets language acquisition specialists while welcoming computational methods.

Finally, we will release any code developed as open-source (e.g., a GitHub repository) and possibly contribute data back to TalkBank or Wordbank if our analyses produce aggregated insights that could be of use (for example, a dataset of estimated vocabulary sizes for each child in certain corpora).

2. Collaborative Team and Required Expertise

We describe the key skill sets needed from collaborators:

- **Natural Language Processing / Computational Linguistics**: An NLP specialist with expertise in handling language corpora (text processing, parsing CHAT format from CHILDES), as well as familiarity with language modeling will help implement any language-model-based predictors and ensure that the data pipeline (from raw transcripts to frequency counts to analysis) is robust and reproducible.
- **Developmental Linguistics / Psycholinguistics**: A collaborator with background in child language acquisition is helpful for interpreting the results. They will provide insights on what estimated vocabulary sizes mean in developmental terms (e.g., are they typical for a given age?). They will also guide selection of appropriate datasets and ensure that the research questions remain grounded in theories of language development. This expertise helps in making sure we consider, for example, the difference between comprehension and production vocabulary, the types of words children learn (nouns vs verbs), and how that might affect sampling.
- **Statistical Modeling / Data Science**: A statistician or data scientist is needed to implement the estimators correctly, prove or verify their assumptions (perhaps deriving when each method should work best), and contribute to the design of simulations. This person will also help with the analytical evaluation, significance testing, and possibly the derivation of any new estimator variants. Knowledge of statistical programming (Python) and familiarity with concepts like maximum likelihood, bias-variance tradeoff, and nonparametric inference is expected.

- **Software Engineering (optional)**: If the project evolves into creating a usable toolkit for other researchers (for example, a Python package that takes child transcript data and outputs estimated vocabulary size with confidence intervals), having someone with solid coding practices will be helpful. This could be combined with the NLP role.

3. Background Information

3.1. **General Background.** Children's early vocabularies are a central focus of language acquisition research. Accurately determining how many words a child knows (and which words) is important for understanding linguistic development and for diagnosing or intervening in cases of atypical development. However, obtaining a complete inventory of a child's vocabulary is difficult in practice. Researchers and caregivers can only hear a fraction of the words a child actually understands or can say, since any given recording or observation session will capture only a small sample of the child's speech. As a result, the collected data on child language are often *undersampled* and incomplete, especially for low-frequency words.

This raises a critical question: *given a sample of utterances a child has produced, how can we infer the total vocabulary (support) from which those utterances are drawn?* This problem of inferring unseen elements from observed samples is analogous to a well-studied statistical problem known as **support estimation** or the **unseen species problem**. In statistics, support estimation refers to estimating the number of distinct elements (the support) in a distribution based on a finite sample. In our context, the distinct elements are the words in the child's lexicon, and the sample is the set of words observed in transcripts of the child's speech.

Classical approaches to this problem date back to the works of Fisher and colleagues in ecology and Good and Turing in cryptography. Indeed, the challenge of estimating how many new species (or new word types) one would observe with more sampling was posed by *Fisher43* and addressed by the famous Good–Turing estimator *Good53,GoodToulmin56*. Good–Turing estimation provides an unbiased estimate of the probability mass of unseen events but can suffer from high variance when extrapolating beyond the original sample size.

In vocabulary estimation, this means naive application of Good–Turing may underestimate the true lexicon size if the sample is small or the vocabulary is very diverse *EfronThisted76*.

In recent years, there have been significant new developments in the theory and methods of support estimation that promise improved accuracy for problems like vocabulary size inference. These include advanced nonparametric estimators and algorithms that achieve near-optimal sample efficiency.

At the same time, the availability of child language datasets has expanded and been modernized. Resources such as **CHILDES** (Child Language Data Exchange System) and related **TalkBank** databases now offer a large repository of transcript data across many languages, while **Wordbank** provides a broad compilation of vocabulary checklist data from infants and toddlers. These resources have become invaluable in linguistics and cognitive science research, enabling large-scale analyses of language acquisition.

However, they also come with inherent limitations: sampling bias, inconsistent data collection methods, and the fundamental issue of incomplete observation of each child's full language exposure and usage.

This proposal bridges the gap between the latest statistical estimation techniques and the needs of child language research. We will formulate the child's vocabulary inference as

a support estimation problem and apply cutting-edge estimators to data from CHILDES, Wordbank, and related corpora. We will evaluate how well these methods infer total vocabulary in realistic scenarios of data sparsity.

The research is expected to yield more reliable estimates of children's lexicon sizes from transcript data, which in turn can shed light on individual differences in language experience and development.

In the following sections, we first review the recent advances in support estimation (Section 2) and describe the current child language datasets and their characteristics (Section 3). We then detail a concrete research plan (Section 4) including proposed methodologies, experiments, and evaluation strategies. Section 5 outlines the required expertise from collaborators to successfully execute the project. Finally, Section 6 discusses the expected outcomes, significance of the work, and plans for dissemination (including a target journal for publication).

3.2. **Statistics Background: Recent Advances in Support Estimation.** Statistical support estimation has undergone a renaissance in the last decade, driven by new theoretical insights and algorithms that can handle the inference of unseen events more effectively than traditional techniques. In classical settings, as noted above, the Good–Toulmin estimator and its variants were early attempts to extrapolate the number of unseen species (or word types) from a given sample. *EfronThisted76* famously applied such methods to estimate how many new words Shakespeare would use if he had written more plays, introducing a smoothing approach to balance bias and variance. These early works laid the foundation but also highlighted the difficulties: naive extrapolation can severely underestimate or overestimate the support if the sample is small or the distribution of word frequencies has a long tail (as is often the case with language, where many words are rare).

Modern developments have substantially improved our ability to estimate the support size from limited samples. One milestone result was the formulation of estimators with proven near-optimal sample complexity. *ValiantValiant11, ValiantValiant13* developed an approach using linear programming and approximation theory to reconstruct the underlying distribution's tail, enabling accurate support estimates. Building on this, *Orlitsky16* (and concurrently *ValiantValiant16*) introduced a clever variant of the Good–Toulmin estimator that remains accurate even when extrapolating to sample sizes on the order of $n \log n$ (where n is the original sample size). This was a significant breakthrough: it showed that one can reliably predict the number of new words that would appear if we, say, doubled or tripled the amount of observed speech, up to a logarithmic factor in n. These new estimators match theoretical lower bounds for the support estimation problem, meaning they are essentially as good as any estimator can be in terms of worst-case sample requirements.

More recently, attention has turned toward practical algorithms and even incorporating machine learning to improve support estimates. *WuYang19*, for example, presented a linear estimator based on Chebyshev polynomials that achieves the optimal sample complexity order. In fact, their algorithm requires on the order of $n/\log n$ samples to reach a given accuracy and is proven optimal under certain assumptions.

In parallel, researchers have explored *learning-augmented* support estimation: using auxiliary information or predictors (such as a language model or neural network) to guide the estimation process. One such approach demonstrated that if one has a reasonably good model of the distribution (for instance, a predictor of word frequencies learned from external data), it can be used to further reduce the estimation error or sample size needed. This hybrid method was shown to yield up to 3x reduction in error on real datasets compared to

the best purely sample-based estimator. The intuition is that by partitioning the frequency range and focusing on the likely frequency of unseen elements, the algorithm can better estimate how many rare words are missing from the sample.

The key takeaway from these developments is that we now have tools that can infer the hidden portion of a distribution (such as a child's unobserved vocabulary) far more accurately than before. These advances include:

- **Unseen species estimators with guarantees**: New estimators (Orlitsky et al., Valiant & Valiant) that allow extrapolation beyond the sample with controlled error, leveraging assumptions about distribution tail behavior.
- **Optimal sample complexity algorithms**: Methods that reach the theoretical limits of performance (e.g., Wu & Yang's Chebyshev-polynomial-based estimator) and require minimal data to estimate support size with high confidence.
- **Learning-informed estimation**: Techniques that incorporate external models or predictors of word frequency to correct or enhance support estimates, reflecting a trend of combining NLP models with statistical estimation.

This progress in support estimation is highly relevant to language acquisition data. Children's vocabularies are dynamic and have heavy-tailed frequency distributions (a few words like *mommy*, *the*, *ball* appear very often, while many others appear rarely). Traditional vocabulary size measures (e.g., type-token ratios or naive extrapolations) often underestimate the lexicon because they miss those rare words. By applying these new estimators, we can aim to quantify a child's total vocabulary more rigorously and perhaps detect growth or differences that were previously obscured by sampling limitations.

Before detailing our research plan to do so, we review the datasets that will provide the empirical basis for this study and discuss why they both enable and necessitate such improved estimation methods.

3.3. **Linguitsics Background: Limitations in Child Language Acquisition Datasets.** A variety of datasets are available for studying child language, each with different methodologies for data collection. In this project, we focus on two main types of data sources: (1) spontaneous speech transcripts (as found in **CHILDES / TalkBank**) and (2) vocabulary checklist data (as in **Wordbank**). These resources are widely used in linguistics and developmental psychology and will supply the data for our analyses. It is important to understand their scope and limitations, as these factors will shape our approach to support estimation.

3.4. **Transcription Corpora: CHILDES and TalkBank.** The **Child Language Data Exchange System (CHILDES)** is a long-standing repository of transcripts of child–adult interactions *MacWhinney2000*. It has been incorporated into the broader **TalkBank** system, which also includes databases for aphasia, second language, multimodal communication, and more *MacWhinney2017*. CHILDES contains transcript collections from numerous studies, languages, and contexts, often including longitudinal recordings of children's speech from the one-word stage through multi-word stage. Researchers have contributed data from English and many other languages (over a dozen), and new data continue to be added.

The availability of CHILDES / TalkBank has led to thousands of published articles across diverse fields that analyze child speech patterns. These range from studies of grammatical development to investigations of parent-child interaction and beyond.

Despite its richness, CHILDES has well-known **limitations**. Firstly, the data are inherently *incomplete*: no matter how diligently a child is recorded, there will be words in the

child's vocabulary that simply never occur during the recorded sessions. Many transcripts come from structured play sessions or lab visits that last an hour or less, or from diary studies with sporadic entries. As a result, the observed vocabulary in a given corpus can significantly under-represent the child's full lexicon. In essence, we see the "tip of the iceberg" of vocabulary, which motivates the need for support estimation to infer the unseen portion.

Secondly, there are **sampling biases**. The families contributing to CHILDES corpora are often WEIRD (Western, Educated, Industrialized, Rich, Democratic) populations, skewing toward middle-class, highly educated parents, which may not generalize to all children. Even within a recording, the context (often playtime with a researcher or parent) might elicit certain vocabulary (toys, household objects) but not others (no child will spontaneously name every animal or color they know unless prompted). Thus, frequencies in transcripts are not a perfect reflection of true word knowledge. Furthermore, transcription practices vary, and some recordings might omit nonverbal context or have transcription errors, though the field has standard CHAT format to mitigate inconsistency *MacWhinney2008*.

Finally, the total size of CHILDES data for any given child is relatively small compared to data in adult language corpora. Preparing spoken data for analysis is labor-intensive, resulting in much less text than we would have from written sources. For example, a child might be recorded for a few hours yielding perhaps a few thousand words of transcribed speech. By contrast, an adult corpus like a section of a newspaper can easily have millions of words. This paucity of data means that naive estimators of vocabulary size (e.g., counting unique words in the sample) will severely undercount the true vocabulary. It also means that more sophisticated statistical estimation is required to make the most of the limited data we do have.

Nonetheless, CHILDES provides the essential *observational evidence* of language use that our study will leverage. We will select subsets of CHILDES data (detailed in the Research Plan) to apply support estimation algorithms, treating each subset as an "observed sample" from an underlying vocabulary distribution.

In addition to CHILDES, the **TalkBank** infrastructure includes related projects like **HomeBank**, a newer repository specifically for day-long audio recordings of children in home environments. HomeBank data, collected via wearable recorders (e.g., LENA devices), offer vastly larger raw audio data, though much of it is not transcribed due to scale. Automatic methods exist to tag segments (speaker diarization, etc.), but word-level transcripts are sparse or only for short clips. For our purposes, fully transcribed corpora from CHILDES are more immediately useful. However, we note that as automatic speech recognition for children improves, HomeBank could become another source of vocabulary observation (with its own set of estimation challenges due to noise). We will focus on CHILDES transcripts for concrete analysis, but the methods developed could potentially be applied to sample data drawn from HomeBank in the future to estimate vocabulary indirectly from acoustic data.

3.5. **Vocabulary Checklists: Wordbank (CDI Data).** Another invaluable resource is **Wordbank** *Frank2017*, an open database that compiles data from the MacArthur-Bates Communicative Development Inventories (CDIs) across many children and languages. The CDI is a parent-report checklist that asks which words (from a fixed list) a child understands or says at a given age (typically targeting infants and toddlers, e.g., 8 to 30 months old). Wordbank aggregates these checklist responses from numerous studies worldwide, providing a large sample (tens of thousands of children) with data on the words they know, as reported

by caregivers. It includes an interactive web interface and an R API (`wordbankr`) for querying the data.

Wordbank is extremely useful for establishing normative trends in early vocabulary growth. For instance, it can tell us the proportion of 18-month-olds who are reported to know the word *dog* or the median vocabulary size of children at 2 years old. It also enables cross-linguistic comparisons of early vocabularies since CDI instruments have been adapted to dozens of languages. Researchers have used Wordbank to explore theoretical questions like the relationship between comprehension and production vocabulary, the influence of language-specific frequency on acquisition, etc.

However, Wordbank (and CDI data in general) has its **own limitations**. The CDI forms have a limited list of words (around 400–680 words for infant and toddler versions in English, for example). This means the data are effectively censored: if a child knows a word not on the list, that knowledge is invisible to CDI data. Thus, while Wordbank might say a child knows 100 words, the child might actually know several more that just aren't asked about. Moreover, CDI is a recognition/production checklist, not an exhaustive inventory—parents might under-report words they forgot their child knows, or occasionally over-report if unsure. It is also limited to early ages; beyond 30 months, children often have vocabularies that exceed the checklist, and other methods must be used.

For our project, Wordbank provides a complementary perspective: it can act as a *benchmark* or partial ground truth for vocabulary size at certain ages. For example, if we take transcript data of a child around 18 months from CHILDES, we might compare our statistically inferred vocabulary size to the distribution of vocabulary sizes reported in Wordbank for 18-month-olds. This can validate whether our estimates are in a plausible range. Additionally, Wordbank data can inform priors or constraints for our estimators (e.g., typical vocabulary size for age, or known subsets of words the child likely knows even if they didn't say them in a recording).

In summary, the child language datasets at our disposal are rich but incomplete. CHILDES/TalkBank give us real usage evidence with linguistic context but only a fragment of total vocabulary. Wordbank gives broad coverage of vocabulary knowledge but only for a predefined list of words and primarily at younger ages. Both underscore the necessity of better inference: we need to stitch together what a child says (from transcripts) with what they are likely to know (from normative data) to estimate the full repertoire.

In the next section, we outline a research plan that uses modern support estimation methods to achieve this, leveraging the strengths of these datasets while accounting for their weaknesses.

4. Conclusion

This proposal addresses a fundamental challenge in child language research using modern statistical tools. By treating the estimation of a child's vocabulary size as an instance of the support estimation problem, we can apply and extend powerful techniques to infer what lies beyond our observations. The synergy between updated datasets (which offer broad and detailed glimpses of child speech) and advanced estimation methods (which compensate for missing data) is at the heart of our approach.

The project is firmly grounded in a concrete plan: we have identified data sources, a range of methods to test, and a strategy to validate them. The interdisciplinary nature of the

work is a strength, bringing together NLP, linguistics, and statistics. As such, the team with varied expertise is well-poised to execute the research.

The outcomes will contribute to multiple fields: providing NLP with a case study of applying distribution estimation in a new domain, giving developmental linguistics a tool to better measure an important construct (vocabulary knowledge), and offering statistics a demonstration of how theoretical advances can make a practical impact on data analysis.

In summary, this research will not only expand our knowledge of how to estimate "how many words children know" with greater accuracy, but it will also exemplify how computational and statistical methods can enrich traditional areas of linguistic inquiry. We look forward to carrying out this plan and reporting our findings in a leading interdisciplinary journal.

References

[Brown(1973)] Brown, R. (1973). *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.

[Efron & Thisted(1976)] Efron, B., & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika, 63*(3), 435–447.

[Frank et al.(2017)] 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.

[Fisher et al.(1943)] Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology, 12*(1), 42–58.

[Good(1953)] Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika, 40*(3–4), 237–264.

[Good & Toulmin(1956)] Good, I. J., & Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika, 43*(1–2), 45–63.

[MacWhinney(2000)] MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum.

[MacWhinney(2017)] MacWhinney, B. (2017). A shared platform for studying second language acquisition. *Language Learning, 67*(S1), 254–275.

[Orlitsky et al.(2016)] Orlitsky, A., Suresh, A. T., & Wu, Y. (2016). Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences, 113*(47), 13283–13288.

[Segbers & Schroeder(2017)] Segbers, J., & Schroeder, S. (2017). How many words do children know? A corpus-based estimation of children's total vocabulary size. *Language Testing, 34*(3), 297–320.

[Valiant & Valiant(2011)] Valiant, G., & Valiant, P. (2011). Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing* (pp. 691–700).

[Valiant & Valiant(2016)] Valiant, G., & Valiant, P. (2016). Instance optimal learning of discrete distributions. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing* (pp. 142–155).

[Wu & Yang(2019)] Wu, Y., & Yang, P. (2019). Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *Annals of Statistics, 47*(2), 857–883.

[Hsu et al.(2021)] Hsu, D., Ji, Z., & Zhou, X. (2021). Learning-based support estimation in sublinear time. In *9th International Conference on Learning Representations (ICLR)*.

Department of Statistics and Data Science, Yale University, New Haven, CT, USA
Email address: sasha.cui@yale.edu