MIT Libraries | DSpace@MIT

MIT Open Access Articles

childes-db: A flexible and reproducible interface to the child language data exchange system

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

As Published: https://doi.org/10.3758/s13428-018-1176-7

Publisher: Springer US

Persistent URL: https://hdl.handle.net/1721.1/131922

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of Use: Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.

Massachusetts Institute of Technology

childes-db: a flexible and reproducible interface to the Child Language Data Exchange System

Alessandro Sanchez*[1], Stephan C. Meylan*[2], Mika Braginsky[3], Kyle E. MacDonald[1], Daniel Yurovsky[4], & Michael C. Frank[1]

[1] Stanford University

[2] University of California, Berkeley

[3] MIT

[4] University of Chicago

Author Note

Co-first authorship indicated with ∗. Thanks to Brian MacWhinney for advice and guidance, and to Melissa Kline for her work on ClanToR, which formed a starting point for our work. This work is supported by a Jacobs Advanced Research Fellowship to MCF.

Correspondence concerning this article should be addressed to Alessandro Sanchez*, Department of Psychology, 450 Serra Mall, Stanford, CA 94305. E-mail: sanchez7@stanford.edu

Abstract

The Child Language Data Exchange System (CHILDES) has played a critical role in research on child language development, particularly in characterizing the early language learning environment. Access to these data can be both complex for novices and difficult to automate for advanced users, however. To address these issues, we introduce `childes-db`, a database-formatted mirror of CHILDES that improves data accessibility and usability by offering novel interfaces, including browsable web applications and an `R` application programming interface (API). Along with versioned infrastructure that facilitates reproducibility of past analyses, these interfaces lower barriers to analyzing naturalistic parent-child language, allowing for a wider range of researchers in language and cognitive development to easily leverage CHILDES in their work.

Keywords: child language; corpus linguistics; reproducibility; R packages; research software

Word count: 2925

childes-db: a flexible and reproducible interface to the Child Language Data Exchange System

Introduction

What are the representations that children learn about language, and how do they emerge from the interaction of learning mechanisms and environmental input? Developing facility with language requires learning a great many interlocking components – meaningful distinctions between sounds (phonology), names of particular objects and actions (word learning), meaningful sub-word structure (morphology), rules for how to organize words together (syntax), and context-dependent and context-independent aspects of meaning (semantics and pragmatics). Key to learning all of these systems is the contribution of the child's input – exposure to linguistic and non-linguistic data – in the early environment. While in-lab experiments can shed light on linguistic knowledge and some of the implicated learning mechanisms, characterizing this early environment requires additional research methods and resources.

One of the key methods that has emerged to address this gap is the collection and annotation of speech to and by children, often in the context of the home. Starting with Roger Brown's (1973) work on Adam, Eve, and Sarah, audio recordings – and more recently video recordings – have been augmented with rich, searchable annotations to allow researchers to address a number of questions regarding the language learning environment. Focusing on language learning in naturalistic contexts also reveals that children have, in many cases, productive and receptive abilities exceeding those demonstrated in experimental contexts. Often, children's most revealing and sophisticated uses of language emerge in the course of naturalistic play.

While corpora of early language acquisition are extremely useful, creating them requires significant resources. Collecting and transcribing audio and video is costly and extremely time consuming – even orthographic transcription (i.e., transcriptions with minimal phonetic detail) can take ten times the duration of the original recording

(MacWhinney, 2000). Automated, machine learning-based methods like automatic speech recognition (ASR) have provided only modest gains in efficiency. Such systems are limited both by the less-than-ideal acoustic properties of home recordings, and also by the poor fit of language models built on adult-directed, adult-produced language samples to child-directed and child-produced speech. Thus, researchers' desires for data in analyses of child language corpora can very quickly outstrip their resources.

Established in 1984 to address this issue, the Child Language Data Exchange System (CHILDES) aims to make transcripts and recordings relevant to the study of child language acquisition available to researchers as free, public datasets (MacWhinney, 2000, 2014; MacWhinney & Snow, 1985). CHILDES now archives tens of thousands of transcripts and associated media across 20+ languages, making it a critical resource for characterizing both children's early productive language use and their language environment. As the first major effort to consolidate and share transcripts of child language, CHILDES has been a pioneer in the move to curate and disseminate large-scale behavioral datasets publicly.

Since its inception, a tremendous body of research has made use of CHILDES data. Individual studies are too numerous to list, but classics include studies of morphological over-regularization (Marcus et al., 1992), distributional learning (Redington, Chater, & Finch, 1998), word segmentation (Goldwater, Griffiths, & Johnson, 2009), the role of frequency in word learning (Goodman, Dale, & Li, 2008), and many others. Some studies analyze individual examples in depth (e.g., Snyder, 2007), some track multiple child-caregiver dyads (e.g., Meylan, Frank, Roy, & Levy, 2017), and still others use the aggregate properties of all child or caregiver speech pooled across corpora (Montag, Jones, & Smith, 2015; e.g., Redington et al., 1998).

Nonetheless, there are some outstanding challenges working with CHILDES, both for students and for advanced users. The CHILDES ecosystem uses a specialized file format (CHAT), which is stored as plain text but includes structured annotations grouped into parallel information "tiers" on separate lines. These tiers allow for a searchable plaintext transcript of an utterance to be stored along with structured annotations of its phonological, morphological, or syntactic content. These files are usually analyzed using a command-line program (CLAN) that allows users to count word frequencies, compute statistics (e.g., mean length of utterance, or MLU), and execute complex searches against the data. While this system is flexible and powerful, mastering the CHAT codes and especially the CLAN tool with its many functions and flags can be daunting. These technical barriers decrease the ease of exploration by a novice researcher or in a classroom exercise.

On the opposite end of the spectrum, for data-oriented researchers who are interested in doing large-scale analyses of CHILDES, the current tools are also not ideal. CLAN software is an excellent tool for interactive exploration, but – as a free-standing application – it can be tricky to build into a processing pipeline written in Python or R. Thus, researchers who would like to ingest the entire corpus (or some large subset) into a computational analysis typically write their own parsers of the CHAT format to extract the subset of the data they would like to use (e.g., Kline, 2012; Meylan et al., 2017; Redington et al., 1998; Yang, 2013).

The practice of writing custom parsers is problematic for a number of reasons. First, effort is wasted in implementing the same features again and again. Second, this process can introduce errors and inconsistencies in data handling due to difficulties dealing with the many special cases in the CHAT standard. Third, these parsing scripts are rarely shared – and when when they are, they typically break with subsequent revisions to the dataset – leading to much greater difficulty in reproducing the exact numerical results from previous published research that used CHILDES (see e.g., Meylan et al., 2017 for an example). Fourth, the CHILDES corpus itself is a moving target: computational work using the entire corpus at one time point may include a different set of data than subsequent work as corpora are added and revised. Currently, there is no simple way for researchers to document exactly which version of the corpus has been used, short of creating a full mirror of the data. These factors together lead to a lack of *computational reproducibility*, a major problem that keeps researchers from verifying or building on published research (Donoho, 2010; Stodden et al., 2016).

In the current manuscript, we describe a system for extending the functionality of CHILDES to address these issues. Our system, `childes-db`, is a database-formatted mirror of CHILDES that allows access through an application programming interface (API). This infrastructure allows the creation of web applications for browsing and easily visualizing the data, facilitating classroom use of the dataset. Further, the database can be accessed programmatically by advanced researchers, obviating the need to write one-off parsers of the CHAT format. The database is versioned for access to previous releases, allowing computational reproducibility of particular analyses.

We begin by describing the architecture of `childes-db` and the web applications that we provide. Next, we describe the `childesr` API, which provides a set of `R` functions for programmatic access to the data while abstracting away many of the technical details. We conclude by presenting several worked examples of specific uses of the system – both web apps and the `R` API – for research and teaching.

Design and technical approach

As described above, CHILDES is most often approached as a set of distinct CHAT files, which are then parsed by users, often using CLAN. In contrast to this parsing approach, which entails the sequential processing of strings, `childes-db` treats CHILDES as a set of linked tables, with records corresponding to intuitive abstractions such as words, utterances, and transcripts (see Kline, 2012 for an earlier example of deriving a singular tabular representation of a CHILDES transcript). Users of data analysis languages like `R` or Julia, libraries like Pandas, or those familiar with Structured Query Language (SQL) will be familiar with operations on tabular representations of data such as filtering (subsetting), sorting, aggregation (grouping), and joins (merges). These operations obviate the need for users to consider the specifics of the CHAT representation – instead they simply request the entities they need for their research and allow the API to take care of the formatting details. We begin by orienting readers to the design of the system via a top-level description and motivation for the design of the database schema, then provide details on the database's current technical implementation and the versioning scheme. Users primarily interested in accessing the database can skip these details and focus on access through the `childesr` API and the web apps.

Database format

At its core, `childes-db` is a database consisting of a set of linked tabular data stores where records correspond to linguistic entities like words, utterances, and sampling units like transcriptions and corpora. The smallest unit of abstraction tracked by the database is a *token*, treated here as the standard (or citation) orthographic form of a word. Using the standardized written form of the word facilitates the computation of lexical frequency statistics for comparison or aggregation across children or time periods. Deviations from the citation form – which are particularly common in the course of language development and often of special interest to researchers – are kept as a separate (possibly null) field associated with each token.

Many of the other tables in the database describe hierarchical collections built out of tokens – *utterance*, *transcript*, *corpus*, and *collection* – and store attributes appropriate for each level of description. Every entity includes attributes that link it to all higher-order collections, e.g., an utterance lists the transcript, corpus, and collection to which it belongs. An *utterance* contains one or more word tokens and includes fields such as the utterance type (e.g., *declarative*, *interrogative*, etc.), total number of tokens, and the total number of morphemes if the morphological structure is available in the original CHAT file. A *transcript* consists of one or more utterances and includes the date collected, the name of the target child, the age in days if defined, and the filename from CHILDES. A *corpus* consists of one or more transcripts, corresponding to well-known collections like the Brown (Brown, 1973) or Providence (Demuth, Culbertson, & Alter, 2006) corpus. Finally, a *collection* is a superordinate collection of corpora generally corresponding to a geographic region, following the convention in CHILDES. Because every record can be linked to a top-level collection (generally corresponding to a language), each table includes data from all languages represented in CHILDES.

Participants – generally children and caregivers – are represented separately from the token hierarchy because it is common for the same children to appear in multiple transcripts. A participant identifier is associated with every word and utterance, including a name, role, 3-letter CHILDES identifier (CHI = child, MOT = mother, FAT = father, etc.), and the range of ages for which they are observed (or age of corresponding child, in the case of caregivers). For non-child participants (caregivers and others), the record additionally contains an identifier for the corresponding target child, such that data corresponding to children and their caregivers can be easily associated.

Technical implementation

`childes-db` is stored as a MySQL database, an industry-standard, open-source relational database that can be accessed directly from a wide range of programming languages. The `childes-db` project provides access to hosted, read-only databases on a publicly-accessible server for direct access and `childesr` (described below). The project also hosts compressed .sql exports for local installation. While the former is appropriate for most users, local installation can provide performance gains by allowing a user to access the database on their machine or on their local network, as well as allowing users to store derived information in the same database.

In order to import the CHILDES corpora into the MySQL schema described above, it must first be accurately parsed and subsequently vetted to ensure its integrity. We parse the XML (eXtensible Markup Language) release of CHILDES hosted by childes.talkbank.org using the NLTK library in Python (Bird & Loper, 2004). Logic implemented in Python converts the linear, multi-tier parse into a tabular format appropriate for `childes-db`. This logic includes decisions that we review below regarding what information sources are captured in the current release of the database and which are left for future development.

The data imported into `childes-db` is subject to data integrity checks to ensure that our import of the corpora is accurate and preferable over ad-hoc parsers developed by many individual researchers. In order to evaluate our success in replicating CLAN parses, we compared unigram counts in our database with those outputted by CLAN, the command-line tool built specifically for analysis of transcripts coded in CHAT. We used the CLAN commands FREQ and MLU to compare total token counts and mean lengths of utterance for every speaker in every transcript and compared these these values to our own using the Pearson correlation coefficient. The results of the comparison were .99 and .98 for the unigram count and MLU data, respectively, indicating reliable parsing.

Versioning. The content of CHILDES changes as additional corpora are added or transcriptions are updated; as of time of writing, these changes are not systematically tracked in a public repository.[1] To facilitate reproducibility of past analyses, we introduce a simple versioning system by adding a new complete parse of the current state of CHILDES every six months or as warranted by changes in CHILDES. By default, users interact with the most recent version of the database available. To support reproduction of results with previous versions of the database, we continue to host recent versions (up to the last three years / six versions) through our `childesr` API so that researchers can run analyses against specific historical versions of the database. For versions more than three years old, we host compressed .sql files that users may download and serve using a local

[1] Specific versions of the database, tracked using the version control system Git, can be obtained by emailing the maintainers of the CHILDES project. While tracking line-level changes with Git provides detailed information about what has changed, our method allows researchers to access the relevant version programmatically by simply adding an argument to a function call.

installation of MySQL server (for which we provide instructions).

Current Annotation Coverage. The current implementation of `childes-db` emphasizes the computation of lexical statistics, and consequently focuses on reproducing the words, utterances, and speaker information in CHILDES transcripts. For this reason, we do not preserve all of the information available in CHILDES, such as:

- Sparsely annotated tiers, e.g. phonology (`%pho`) and situation (`%sit`)
- Media links
- Tone direction and stress
- Filled pauses
- Reformulations, word revision, and phrase revision, e.g. <what did you>[//] how can you see it ?
- paralinguistic material, e.g. [=! cries]

At present, `childes-db` focuses strictly on the contents of CHILDES, and does not include material in related TalkBank projects such as PhonBank, AphasiaBank, or DementiaBank. We will prioritize the addition of these information sources and others in response to community feedback.

Interfaces for Accessing `childes-db`

We first discuss the `childes-db` web apps and then introduce the `childesr R` package.

Interactive Web Apps

The ability to easily browse and explore the CHILDES corpora is a cornerstone of the `childes-db` project. To this end we have created powerful yet easy-to-use interactive web applications that enable users to visualize various dimensions of the CHILDES corpus: frequency counts, mean lengths of utterance, type-token ratios, and more. All of this is doable without the requirement of understanding command-line tools.[2]

Our web apps are built using Shiny, a software package that enables easy app construction using `R`. Underneath the hood, each web app is making calls to our `childesr` API and subsequently plots the data using the popular `R` plotting package `ggplot2`. A user's only task is to configure exactly what should be plotted through a series of buttons, sliders, and text boxes. The user may specify what collection, corpus, child, age range, caregiver, etc., should be included in a given analysis. The plot is displayed and updated in real-time, and the underlying data are also available for download alongside the plot. All of these analyses may also be reproduced using the `childesr` package, but the web apps are intended for the casual user who seeks to easily extract developmental indices quickly and without any technical overhead.

Frequency Counts. The lexical statistics of language input to children have long been an object of study in child language acquisition research. Frequency counts of words in particular may provide insight into the cognitive, conceptual, and linguistic experience of a young child (see e.g., Ambridge, Kidd, Rowland, & Theakston, 2015 for review). In this web app, inspired by ChildFreq (Bååth, 2010), we provide users the ability to search for any word spoken by a participant in the CHILDES corpora and track the usage of that word by a child or caregiver over time. Because of the various toggles available to the user that can subset the data, a user may view word frequency curves for a single child in the Brown corpus or all Spanish speaking children, if desired. In addition, users can plot frequency curves belonging to caregivers alongside their child for convenient side-by-side comparisons. A single word or multiple words may be entered into the input box.

Derived Measures. The syntactic complexity and lexical diversity of children's speech are similarly critical metrics for acquisition researchers (Miller & Chapman, 1981;

[2]The LuCiD toolkit (Chang, 2017) provides related functionality for a number of common analyses. In contrast to those tools, which focus on filling gaps not covered by CLAN – e.g., the use of n-gram models, incremental sentence generation, and distributional word classification – our web apps focus on covering the same common tasks as CLAN, but yielding visualizations for the web browser.

Watkins, Kelly, Harbers, & Hollis, 1995). There are a number of well-established measures of children's speech that operationalize complexity and diversity, and have many applications in speech-language pathology (SLP), where measures outside of the normal range may be indicative of speech, language, or communication disorders.

Several of the most common of these measures are available in the Derived Measures app, which plots these measures across age for a given subset of data, again specified by collection, corpora, children, and speakers. As with the Frequency Counts app, caregivers' lexical diversity measures can be plotted alongside children's. We have currently implemented the following measures:

- MLU-w (mean length of utterance in words),
- MLU-m (mean length of utterance in morphemes),
- TTR (type-token ratio, a measure of lexical diversity; Templin, 1957),
- MTLD (measure of textual lexical diversity; Malvern & Richards, 1997),
- HD-D (lexical diversity via the hypergeometric distribution; McCarthy & Jarvis, 2010)

As with the Frequency Counts app, a user may subset the data as they choose, compare measures between caregivers and children, and aggregate across children from different corpora.

Population Viewer. In many cases a researcher may want to view the statistics and properties of corpora (e.g., their size, number of utterances, number of tokens) before choosing a target corpus or set of corpora for an analysis. This web app is intended to provide a basic overview regarding the scale and temporal extent of various corpora in CHILDES, as well as give researchers insight into the aggregate characteristics of CHILDES. For example, examining the aggregate statistics reveals that coverage in CHILDES peaks at around 30 months.

The `childesr` Package

Although the interactive analysis tools described above cover some of the most common use cases of CHILDES data, researchers interested in more detailed and flexible analyses will want to interface directly with the data in `childes-db`. Making use of the `R` programming language (R Core Team, 2017), we provide the `childesr` package. `R` is an open-source, extensible statistical computing environment that is rapidly growing in popularity across fields and is increasing in use in child language research (Norrman & Bylund, 2015; e.g. Song, Shattuck-Hufnagel, & Demuth, 2015). The `childesr` package abstracts away the details of connecting to and querying the database. Users can take advantage of the tools developed in the popular `dplyr` package (Wickham, Francois, Henry, & Müller, 2017), which makes manipulating large datasets quick and easy. We describe the commands that the package provides and then give several worked examples of analyses using the package.

The `childesr` package is easily installed via CRAN, the comprehensive `R` archive network. To install, simply type: `install.packages("childesr")`. After installation, users have access to functions that can be used to retrieve tabular data from the database:

- `get_collections()` gives the names of available collections of corpora ("Eng-NA", "Spanish", etc.)
- `get_corpora()` gives the names of available corpora ("Brown", "Clark", etc.)
- `get_transcripts()` gives information on available transcripts (language, date, target child demographics)
- `get_participants()` gives information on transcript participants (name, role, demographics)
- `get_speaker_statistics()` gives summary statistics for each participant in each transcript (number of utterances, number of types, number of tokens, mean length of utterance)
- `get_utterances()` gives information on each utterance (glosses, stems, parts of

speech, utterance type, number of tokens, number of morphemes, speaker information, target child information)

- `get_types()` gives information on each type within each transcript (gloss, count, speaker information, target child information)
- `get_tokens()` gives information on each token (gloss, stem, part of speech, number of morphemes, speaker information, target child information)

Each of these functions take arguments that restrict the query to a particular subset of the data (e.g. by collection, by corpus, by speaker role, by target child age, etc.) and returns the output in the form of a table. All functions support the specification of the database version to use. For more detailed documentation, see the package repository (http://github.com/langcog/childesr).

Using `childes-db`: Worked Examples

In this section we give a number of examples of how `childes-db` can be used in both research and teaching, using both the web apps and the `R` API. Note that all of these examples use `dplyr` syntax (Wickham et al., 2017); several accessible introductions to this framework are available online (e.g., Wickham & Grolemund, 2016).

Research applications

Color frequency. One common use of CHILDES is to estimate the frequency with which children hear different words. These frequency estimates are used both in the development of theory (e.g., frequent words are learned earlier; Goodman et al., 2008), and in the construction of age-appropriate experimental stimuli. One benefit of the `childes-db` interface is that it allows for easy analysis of how the frequencies of words change over development. Many of our theories in which children learn the structure of language from its statistical properties implicitly assume that these statistics are *stationary*, i.e. unchanging over development (e.g., Saffran, Aslin, & Newport, 1996). However a number of recent analyses show that the frequencies with which infants encounter both linguistic and visual properties of their environment may change dramatically over development (Fausey, Jayaraman, & Smith, 2016), and these changing distributions may produce similarly dramatic changes in the ease or difficulty with which these regularities can be learned (Elman, 1993).

To demonstrate how one might discover such non-stationarity, we take as a case study the frequency with which children hear the color words of English (e.g. "blue", "green"). Color words tend to be learned relatively late by children, potentially in part due to the abstractness of the meanings to which they refer (see Wagner, Dobkins, & Barner, 2013). However, within the set of color words, the frequency with which these words are heard predicts a significant fraction of the variance in their order of acquisition (Yurovsky, Wagner, Barner, & Frank, 2015). But are these frequencies stationary – e.g. do children hear "blue" as often at 12 months as they do at 24 months? We answer this question in two ways – first using the web apps, and then using the `childesr` package.

Using web apps. To investigate whether the frequency of color words is stationary over development, a user can navigate to the Frequency app, and enter a set of color words into the `Word` selector separated by a comma: here "blue, red, green." Because the question of interest is about the frequency of words in the input (rather than produced by children), the `Speaker` field can be set to reflect this choice. In this example we select "Mother." Because children learn most of their basic color words by the age of 5, the age range 1–5 years is a reasonable choice for `Ages to include`. The results of these selections are shown in Figure **??**. We can also create a hyperlink to store these set of choices so that we can share these results with others (or with ourselves in the future) by clicking on the `Share Analysis` button in the bottom left corner.

From this figure, it seems likely that children hear "blue" more frequently early in development, but the trajectories of "red" and "green" are less clear. We also do not have a good sense of the errors of these measurements, are limited to just a few colors at a time before the plot becomes too crowded, and cannot combine frequencies across speakers. To perform this analysis in a more compelling and complete way, a user can use the `childesr` interface.

Using `childesr`. We can analyze these learning trajectories using `childesr` by breaking the process into five steps: (1) define our words of interest, (2) find the frequencies with which children hear these words, (3) find the proportion of the *total words* children hear that these frequencies account for, (4) aggregate across transcripts and children to determine the error in our estimates of these proportions, and (5) plot the results.

For this analysis, we will define our words of interest as the basic color words of English (except for gray, which children hear very rarely). We store these in the `colors` variable, and then use the `get_types()` function from `childesr` to get the type frequency of each of these words in all of the corpora in CHILDES. All other functions are provided by base `R` or the `tidyverse` package. For demonstration, we look only at the types produced by the speakers in each corpus tagged as Mother and Father. We also restrict ourselves to children from 1–5 years old (12–60 months), and look only at the North American English corpora.

```
colors <- c("black", "white", "red", "green", "yellow", "blue", "brown",
            "orange", "pink", "purple")

color_counts <- get_types(collection = "Eng-NA",
                          role = c("Mother", "Father"),
                          age = c(12, 60),
                          type = colors)
```

To normalize correctly (i.e., to ask what proportion of the input children hear consists of these color words), we need to know how many total words these children hear from their parents in these transcripts. To do this, we use the `get_speaker_statistics()` function, which will return a total number of tokens (`num_tokens`) for each of these speakers.

```
# Get the ids corresponding to all of the speakers we are interested in
parent_ids <- color_counts %>%
  distinct(collection_id, corpus_id, transcript_id, speaker_id)

# Find the total number of tokens produced by these speakers
parents <- parent_ids %>%
  left_join(get_speaker_statistics(collection = "Eng-NA")) %>%
  select(collection_id, corpus_id, transcript_id, speaker_id, num_tokens)
```

We now join these two pieces of information together – how many times each speaker produced each color word, and how many total words they produced. We then group the data into 6-month age bins, and compute the proportion of tokens that comprise each color for each child in each 6-month bin. For comparability with the web app analysis, these proportions are converted to parts per million words.

```
count_estimates <- color_counts %>%
  left_join(parents) %>%
  mutate(age_months = target_child_age,
         age_bin = as.integer(floor(age_months / 6) * 6),
         color = tolower(gloss)) %>%
  group_by(age_bin, color, target_child_id, transcript_id) %>%
  summarise(transcript_count = sum(count), transcript_num_tokens = sum(num_tokens)) %>%
  group_by(age_bin, color, target_child_id) %>%
  summarise(child_count = sum(transcript_count), child_num_tokens = sum(transcript_num_t
  mutate(parts = child_count / child_num_tokens * 1e6)
```

Finally, we use non-parametric bootstrapping to estimate 95% confidence intervals for our estimates of the parts per million words of each color term with the `tidyboot` package.

```
count_estimates_with_error <- count_estimates %>%
  tidyboot::tidyboot_mean(parts) %>%
  left_join(graph_colors) %>%
  mutate(color = factor(color, levels = colors))
```

Figure **??** shows the results of these analyses: Input frequency varies substantially over the 1–5 year range for nearly every color word.

Gender. Gender has long been known to be an important factor for early vocabulary growth, with girls learning more words earlier than boys (Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991). Parent-report data from ten languages suggest that female children have larger vocabularies on average than male children in nearly every language (Eriksson et al., 2012). Comparable cross-linguistic analysis of naturalistic production data has not been conducted, however, and these differences are easy to explore using `childesr`. By pulling data from the `transcript_by_speaker` table, a user has access to a set of derived linguistic measures that are often used to evaluate a child's grammatical development. In this worked example, we walk through a sample analysis that explores gender differences in early lexical diversity.

First, we use the `childesr` function call `get_speaker_statistics()` to pull data relating to the aforementioned derived measures for children and their transcripts. Note that we exclusively select the children's production data, and exclude their caregivers' speech.

```
speaker_stats <- get_speaker_statistics(role = "Target_Child")
```

This `childesr` call retrieves data from all collections and corpora, including those languages for which there are very sparse data. In order to make any substantial inferences from our analysis, we begin by filtering the dataset to include only languages for which there are a large number of transcripts (> 500). We also restrict our analysis to children under the age of four years.

```
number_of_transcripts_threshold <- 500
max_age <- 4

included_languages <- speaker_stats %>%
  filter(target_child_age < max_age * 12) %>%
  count(language) %>%
  filter(n > number_of_transcripts_threshold) %>%
  pull(language)
```

Our `transcript_by_speaker` table contains multiple derived measures of lexical diversity – here we use MTLD (McCarthy, 2005). MTLD is derived from the average length of orthographic words that are above a pre-specified type-token ratio, making it more robust to transcript length than simple TTR. We start by filtering to include only those children for which a sex was defined in the transcript, who speak a language in our subset of languages with a large number of transcripts, and who are in the appropriate age range. We then compute an average MTLD score for each child at each age point by aggregating across transcripts while keeping information about the child's sex and language. Note that one child in particular, "Leo" in the eponymous German corpus, contained transcripts that were a collection of his most complex utterances (as caregivers were instructed to record); this child was excluded from the analysis.

```
mtld_data <- speaker_stats %>%
  filter(!is.na(target_child_sex), target_child_name != "Leo",
         language %in% included_languages) %>%
  group_by(target_child_id, target_child_age, target_child_sex, language) %>%
  summarise(measure = mean(mtld)) %>%
  ungroup() %>%
  mutate(age_years = target_child_age / 12,
```

```
        target_child_sex = factor(target_child_sex,
                                          levels = c("male","female"))) %>%
  filter(age_years < max_age)
```

The data contained in CHILDES is populated from a diverse array of studies reflecting varying circumstances of data collection. This point is particularly salient in our gender analysis due to potential non-independence issues that may emerge from the inclusion of many transcripts from longitudinal studies. To account for non-indepenence, we fit a linear mixed effects model with a $gender * age$ (treated as a quadratic predictor) interaction as fixed effects, child identity as a random intercept, and $gender + age$ by language as a random slope, the maximal converging random effects structure (Barr, Levy, Scheepers, & Tily, 2013).[3] The plot below displays the average MTLD scores for various children at different ages, split by gender, with a line corresponding to the prediction of our fit mixed effects model.

This plot reveals a slight gender difference in linguistic productivity in young children, replicating the moderate female advantage found by Eriksson et al. (2012). The goal of this analysis was to showcase an example of using `childesr` to explore the CHILDES dataset. We also highlighted some of the potential pitfalls – sparsity and non-independence – that emerge in working with a diverse set of corpora, many of which were collected in longitudinal studies.

Teaching with childes-db

In-class demonstrations. Teachers of courses on early language acquisition often want to illustrate the striking developmental changes in children's early language. One method is to present static displays that show text from parent-child conversations extracted from CHILDES or data visualizations of various metrics of production and input (e.g., MLU or Frequency), but one challenge of such graphics is that they cannot be

[3] All code and analyses are available at https://github.com/langcog/childes-db-paper

modified during a lecture and thus rely on the instructor selecting examples that will be compelling to students. In contrast, in-class demonstrations can be a powerful way to explain complex concepts while increasing student engagement with the course materials.

Consider the following demonstration about children's first words. Diary studies and large-scale studies using parent report show that children's first words tend to fall into a fairly small number of categories: people, food, body parts, clothing, animals, vehicles, toys, household objects, routines, and activities or states (Clark, 2009; Fenson et al., 1994; Tardif et al., 2008). The key insight is that young children talk about what is going on around them: people they see every day, e.g., toys and small household objects they can manipulate or food they can control. To illustrate this point, an instructor could:

1. introduce the research question (e.g., What are the types of words that children first produce?),
2. allow students to reflect or do a pair-and-share discussion with their neighbor,
3. show the trajectory of a single lexical item while explaining key parts of the visualization (see Panel A of Figure **??**),
4. elicit hypotheses from students about the kinds of words that children are likely to produce,
5. make real-time queries to the web application to add students' suggestions and talk through the updated plots (Panels B and C of Figure **??**), and
6. finish by entering a pre-selected set of words that communicate the important takeaway point (Panel D of Figure **??**) .

Tutorials and programming assignments. One goal for courses on applied natural language processing (NLP) is for students to get hands-on experience using NLP tools to analyze real-world language data. A primary challenge for the instructor is to decide how much time should be spent teaching the requisite programming skills for accessing and formatting language data, which are typically unstructured. One pedagogical strategy is to abstract away these details and avoid having students deal with obtaining data and formatting text. This approach shifts students' effort away from data cleaning and towards programming analyses that encourage the exploration and testing of interesting hypotheses. In particular, the `childesr` API provides instructors with an easy-to-learn method for giving students programmatic access to child language data.

For example, an instructor could create a programming assignment with the specific goal of reproducing the key findings in the case studies presented above – color words or gender. Depending on the students' knowledge of R, the instructor could decide how much of the `childesr` starter code to provide before asking students to generate their own plots and write-ups. The instructor could then easily compare students' code and plots to the expected output to measure learning progress. In addition to specific programming assignments, the instructor could use the `childes-db` and `childesr` workflow as a tool for facilitating student research projects that are designed to address new research questions.

Conclusion

We have presented `childes-db`, a database formatted mirror of the CHILDES dataset. This database – together with the `R` API and web apps – facilitates the use of child language data. For teachers, students, and casual explorers, the web apps allow browsing and demonstration. For researchers interested in scripting more complex analyses, the API allows them to abstract away from the details of the CHAT format and easily create reproducible analyses of the data. We hope that these functionalities broaden the set of users who can easily interact with CHILDES data, leading to future insights into the process of language acquisition.

`childes-db` addresses a number of needs that have emerged in our own research and teaching, but there are still a number of limitations that point the way to future improvements. For example, `childes-db` currently operates only on transcript data, without links to the underlying media files; in the future, adding such links may facilitate further computational and manual analyses of phonology, prosody, social interaction, and other phenomena by providing easy access to the video and audio data. Further, we have focused on including the most common and widely-used tiers of CHAT annotation into the database first, but our plan is eventually to include the full range of tiers. Finally, a wide range of further interactive analyses could easily be added to the current suite of web apps. We invite other researchers to join us in both suggesting and contributing new functionality as our system grows and adapts to researchers' needs.

References

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*(2), 239–273.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bååth, R. (2010). ChildFreq: An online tool to explore word frequencies in child language. *Lucs Minor*, *16*, 1–6.

Bird, S., & Loper, E. (2004). NLTK: The natural language toolkit. In *Proceedings of the acl 2004 on interactive poster and demonstration sessions* (p. 31). Association for Computational Linguistics.

Brown, R. (1973). *A first language: The early stages.* Harvard U. Press.

Chang, F. (2017). The lucid language researcher's toolkit [computer software]. Retrieved from http://www.lucid.ac.uk/resources/for-researchers/toolkit/

Clark, E. V. (2009). *First language acquisition.* Cambridge University Press.

Demuth, K., Culbertson, J., & Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of english. *Language and Speech*, *49*(2), 137–173.

Donoho, D. L. (2010). An invitation to reproducible computational research. *Biostatistics*, *11*(3), 385–388.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

Eriksson, M., Marschik, P. B., Tulviste, T., Almgren, M., Pérez Pereira, M., Wehberg, S., … Gallego, C. (2012). Differences between girls and boys in emerging language skills: Evidence from 10 language communities. *British Journal of Developmental Psychology*, *30*(2), 326–343.

Fausey, C. M., Jayaraman, S., & Smith, L. B. (2016). From faces to hands: Changing visual input in the first two years. *Cognition*, *152*, 101–107.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., … Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, i–185.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(3), 515–531.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*(2), 236.

Kline, M. (2012). CLANtoR. http://github.com/mekline/CLANtoR/; GitHub. doi:10.5281/zenodo.1196626

MacWhinney, B. (2000). *The childes project: The database* (Vol. 2). Psychology Press.

MacWhinney, B. (2014). *The childes project: Tools for analyzing talk, volume ii: The database*. Psychology Press.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*(2), 271–295.

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. *British*

Studies in Applied Linguistics, *12*, 58–71.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, i–178.

McCarthy, P. M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (mtld). *Dissertation Abstracts International*, *66*, 12.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392.

Meylan, S. C., Frank, M. C., Roy, B. C., & Levy, R. (2017). The emergence of an abstract grammatical category in children's early speech. *Psychological Science*, *28*(2), 181–192.

Miller, J. F., & Chapman, R. S. (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech, Language, and Hearing Research*, *24*(2), 154–161.

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The Words Children Hear: Picture Books and the Statistics for Language Learning. *Psychological Science*, *26*(9), 1489–1496.

Norrman, G., & Bylund, E. (2015). The irreversibility of sensitive period effects in language development: Evidence from second language acquisition in international adoptees. *Developmental Science*.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*(4), 425–469.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Snyder, W. (2007). *Child language: The parametric approach*. Oxford University Press.

Song, J. Y., Shattuck-Hufnagel, S., & Demuth, K. (2015). Development of phonetic variants (allophones) in 2-year-olds learning american english: A study of alveolar stop/t, d/codas. *Journal of Phonetics*, *52*, 152–169.

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., … Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science*, *354*(6317), 1240–1241.

Tardif, T., Fletcher, P., Liang, W., Zhang, Z., Kaciroti, N., & Marchman, V. A. (2008). Baby's first 10 words. *Developmental Psychology*, *44*(4), 929.

Templin, M. (1957). Certain language skills in children: Their development and interrelationships (monograph series no. 26). *Minneapolis: University of Minnesota, the Institute of Child Welfare*.

Wagner, K., Dobkins, K., & Barner, D. (2013). Slow mapping: Color word learning as a gradual inductive process. *Cognition*, *127*(3), 307–317.

Watkins, R. V., Kelly, D. J., Harbers, H. M., & Hollis, W. (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech, Language, and Hearing Research*, *38*(6), 1349–1355.

Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc."

Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Yang, C. (2013). Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences*, *110*(16), 6324–6327.

Yurovsky, D., Wagner, K., Barner, D., & Frank, M. C. (2015). Signatures of domain-general categorization mechanisms in color word learning. In *CogSci*.