J. Child Lang., Page 1 of 18. © Cambridge University Press 2016
doi:10.1017/S0305000916000209

Wordbank: an open repository for developmental vocabulary data*

MICHAEL C. FRANK, MIKA BRAGINSKY, DANIEL YUROVSKY AND VIRGINIA A. MARCHMAN
Stanford University, USA

(*Received* 30 *July* 2015 – *Revised* 29 *December* 2015 – *Accepted* 13 *March* 2016)

ABSTRACT

The MacArthur-Bates Communicative Development Inventories (CDIs) are a widely used family of parent-report instruments for easy and inexpensive data-gathering about early language acquisition. CDI data have been used to explore a variety of theoretically important topics, but, with few exceptions, researchers have had to rely on data collected in their own lab. In this paper, we remedy this issue by presenting Wordbank, a structured database of CDI data combined with a browsable web interface. Wordbank archives CDI data across languages and labs, providing a resource for researchers interested in early language, as well as a platform for novel analyses. The site allows interactive exploration of patterns of vocabulary growth at the level of both individual children and particular words. We also introduce `wordbankr`, a software package for connecting to the database directly. Together, these tools extend the abilities of students and researchers to explore quantitative trends in vocabulary development.

INTRODUCTION

Learning language is one of the most impressive and intriguing human accomplishments, and understanding the processes by which vocabulary grows can provide a window into mechanisms of linguistic and cognitive development more generally (e.g. Bloom, 2002). The MacArthur-Bates

[*] This work supported by a John Merck Scholars award and NSF BCS-1528526. Thanks to Ranjay Krishna for contributions to the initial development of the site, to Rune Nørgaard Jørgensen for helping port data from CLEX, to all of the contributors listed at <http://wordbank.stanford.edu/contributors> for generously sharing their data, and to the Advisory Board of the MacArthur-Bates Communicative Development Inventories, especially Philip Dale and Larry Fenson, for their support. Address for correspondence: Michael C. Frank, Department of Psychology, Jordan Hall (Bldg. 420), 450 Serra Mall, Stanford, CA 94305; tel: (650) 724-4003; e-mail: mcfrank@stanford.edu

Communicative Development Inventories[1] (Fenson *et al*., 1994, 2007) are a widely used family of parent-report instruments for easy and inexpensive data-gathering about early language acquisition. CDI data have been used to explore many theoretically rich topics, including variation in early word production (Fenson *et al*., 1994), vocabulary composition (Bates *et al*., 1994), the relationship between lexical and grammatical development (Bates & Goodman, 1999), and the growth of lexical networks (Hills, Maouene, Maouene, Sheya & Smith, 2009). With few exceptions, however, researchers have had to rely on data collected in their own lab. While CDI norms are available (Fenson *et al*., 2007; Jørgensen, Dale, Bleses & Fenson, 2010), no public resource offers researchers the opportunity to share and access raw, cross-linguistic data at the scale necessary to address questions about demographic variation, vocabulary composition, relations with grammatical development, and other important issues.

To remedy this issue, we introduce Wordbank (<http://wordbank.stanford.edu>), a structured database of developmental vocabulary data. Building on previous tools like Cross Linguistic Lexical Norms (CLEX; Jørgensen *et al*., 2010), Wordbank archives raw CDI data across languages and labs, providing a large-scale database of information about children's vocabulary knowledge. The site hosts an interactive and expandable set of in-depth analyses that can be explored by interested researchers, students, and members of the public. Wordbank lowers the cost of new, exploratory analyses by facilitating the productive reuse of data.

The current paper presents the Wordbank site in detail. We begin by discussing the motivations for constructing such a site. The bulk of the paper then describes the Wordbank site, including its database architecture, its web-based front-end, and its extensibility. In particular we highlight two analysis functions that are provided by the online interface: vocabulary growth norms across individuals, and trajectories of acquisition for individual words. These broad analyses allow a very wide range of targeted investigations. Throughout the paper, we use an exploration of gender differences in production vocabulary as a worked case study that illustrates the various features of the site. We end by presenting `wordbankr`, a package for the `R` statistical programming language that allows research users to access the database directly.

MOTIVATION AND BACKGROUND

The nature and course of early word learning is an important window into children's growing understanding of the world. Early words cross-cut a variety of linguistic categories, but generally consist of names for

[1] We use the umbrella abbreviation 'CDI' to refer to the broader class of parent-report instruments adapted from the original English version.

caregivers (e.g. *mama*), common objects (e.g. *bottle, shoe*), social expressions (e.g. *bye-bye*), and actions or routines (e.g. *peekaboo, throw*) (Nelson, 1973; Tardif *et al*., 2008). New words enter children's expressive vocabularies slowly at first, but this process accelerates over the second year such that children reach an average of 300 words by 24 months and more than 60,000 by the time they graduate from high school (Fenson *et al*., 2007). At the same time, there are significant individual differences in language acquisition. For example, according to detailed observational studies, although some 18-month-olds already produce 50–75 words, others produce no words at all, and will not do so until they are 22 months or older (e.g. Brown, 1973; Bloom, 2002; Clark, 2003). How can such differences be measured accurately and efficiently? And can we promote early detection of differences in vocabulary growth that will be clinically significant later in development?

Measuring early vocabulary

Traditional studies of language development typically apply a combination of observational assessment and structured tests, frequently relying on short samples of interactions and small samples of children. Discerning both the universal features and natural variation of early lexical development has been greatly facilitated by the development of parent-report instruments like the MacArthur-Bates CDI (Fenson *et al*., 1994, 2007) and the Language Development Survey (LDS; Rescorla, 1989). The CDIs in particular were developed across a period of more than forty years. Originally designed for use in a research study (Bates, 1976), the instruments have evolved from a structured interview to the current paper-and-pencil format and are now increasingly administered online (e.g. Kristoffersen *et al*., 2013, for Norwegian or <http://laboratorium.detskarec.sk/> for Slovak). While other assessment tools exist for slightly older children, to our knowledge, no other measure allows cost-effective global language assessment for children in the critical age ranges between the emergence of language and the period when children become more able to engage in structured, face-to-face activities (around 30 months).

Naturalistic observations are the other leading candidate for measurement of early language, but such observations are extremely costly and time-consuming to transcribe and annotate. These difficulties lead to a trade-off where most studies either include dense data about a small number of children or smaller amounts of data with a larger sample size. Dense datasets currently provide the best method for in-depth study of the interaction between learning mechanisms and language input in individuals (e.g. Lieven, Salomo & Tomasello, 2009; Roy, Frank,

DeCamp, Miller & Roy, 2015), although the generality of these studies is necessarily limited by their small sample sizes. At the other end of the spectrum, assessment of many individual language samples can yield information about individual variability (e.g. Dickinson & Tabors, 2001; Cartmill *et al*., 2013; Weisleder & Fernald, 2013), but at some cost in terms of depth.

In addition, naturalistic observations do not measure children's language comprehension, a variable of interest for many early language researchers. Estimates of production vocabulary from naturalistic observation are highly correlated with the CDI within studies (e.g. Bornstein & Haynes, 1998), but affected substantially by length of the session, context, and interlocutor when comparing across studies. And although there exist methods to extract insights about global vocabulary from naturalistic observation, these statistical extrapolations are relatively new and have not been validated extensively (Hidaka, 2016). Other comprehension vocabulary measures are also available across some range of languages (e.g. the Peabody Picture Vocabulary Test 4; Dunn & Dunn, 2007), but these assessments are tailored for substantially older children.

Parent-report measures like the CDI and LDS take advantage of the fact that parents are expert observers of their child. CDI instruments ask about use of communicative gestures, grammar, and symbolic play, as well as vocabulary, which is measured using checklists consisting of representative samples of words. Parents choose the words their child currently 'understands' (comprehension, measured for younger children) or 'says' (production, measured for both younger and older children). The checklists contain words from many different semantic (e.g. animal names, household items) and syntactic (e.g. action words, connectives) categories, resulting in broader samples of lexical knowledge than are available from other methods. In their English and Spanish instantiations, the instruments come in two versions: Words & Gestures (8–18 months) and Words & Sentences (16–30 months). Originally designed for English, parallel instruments have now been adapted for more than sixty languages (Dale & Penfold, n.d.).

Limitations of parent report

Although the standardization of parent reports using the CDI contributes to the availability of large amounts of data in a comparable format, there are significant limitations to the parent-report methodology as well (Tomasello & Mervis, 1994; Feldman *et al*., 2000). First, parents may be biased observers; some may overestimate, while others likely underestimate their children's abilities. There is also some evidence that some variability may be due to reporting biases linked to factors such as socioeconomic status

(Feldman *et al*., 2000, 2005; Fenson *et al*., 2000). Second, parent reports of comprehension for younger children likely suffer from a number of biases and are probably substantially more accurate for content words than function words. Third, the items on the original CDI instruments were chosen to be a representative sample of vocabulary items for the appropriate age and language (Fenson *et al*., 1994), not with the intention that they would be a complete set of words that could be compared across instruments, or that they would be individually reliable and license the conclusion that a particular child knows a particular word. Fourth, although the length of the CDI may give the impression that it yields an estimate of the child's full vocabulary, in fact it likely understates the size of a child's vocabulary substantially, especially for older children (Mayor & Plunkett, 2011).

Despite these limitations, when used appropriately the CDI instruments are an important tool. The instruments were designed to minimize bias by targeting current behaviors and asking parents about highly salient features of their child's abilities. They yield reliable and valid estimates of total vocabulary size, with dozens of studies demonstrating concurrent and predictive relations with naturalistic and observational measures, in both typically developing and at-risk populations (e.g. Dale & Fenson, 1996; Thal, Jackson-Maldonado & Acosta, 2000; Marchman & Martínez-Sussmann, 2002). In addition, a variety of recent work has shown that individual item-level responses can yield exciting new insights, for example about the growth patterns of semantic networks (Hills *et al*., 2009; Hills, Maouene, Riordan & Smith, 2010). Such analyses have the potential to be even more powerful when applied to larger samples and across languages.

WORDBANK

To take advantage of the opportunity posed by the broad use of CDI instruments in the child language community, we have constructed Wordbank, an open repository for CDI data that allows for interactive analysis and visualization. The main page of the site at time of writing is shown in Figure 1. In this section, we begin by describing technical details of the site's database architecture. We then describe the two primary analysis tools that form the heart of the site's interactive functionality. We give a worked example of how to use these, and then end by discussing the extensibility of the Wordbank framework, highlighting opportunities for contributing data and for building new analyses.

Our inspiration for Wordbank comes from two successful projects for sharing data on children's language acquisition. The first is the Child Language Data Exchange System (CHILDES; MacWhinney, 2000). A database of transcripts of children's speech and speech to children, CHILDES has grown into a robust and important tool for the

Fig. 1. Screenshot of the Wordbank main page. Visitors can navigate from this page to the interactive reports, as well as to a statistics page that shows the database composition, a contributors page that shows citation information, and a blog that highlights recent updates.

community, with many contributors and affiliated projects. The second is the Cross Linguistic Lexical Norms site (CLEX; <www.cdi-clex.org>; Jørgensen *et al*., 2010), which is closer in content to Wordbank, and effectively our precursor. CLEX archives normative data from a range of CDI adaptations across languages, allowing browsing of acquisition trajectories for individual items or age groups.

Wordbank builds on CLEX, offering the same functionality but allowing flexible and interactive visualization and analysis, as well as direct database access and data download. In addition, Wordbank's goal is to extend beyond the norming data provided by the developers of individual CDIs by dynamically incorporating data from many different researchers and projects of varying sizes and scopes. While the resulting datasets in Wordbank are likely more heterogeneous, they nevertheless have the potential to be considerably larger and more representative than the individual norming datasets. Wordbank provides tools that enable more powerful, flexible, and nuanced analyses of general trends and comparisons across sub-populations in a variety of different languages.

While the general Wordbank architecture enables a huge variety of analyses in principle, some illustrative examples are helpful for

understanding the site. Consider an experimenter constructing a new set of stimuli for a word recognition experiment: the appropriate tool for this task would be the Item Trajectories analysis, which shows the trajectory of acquisition for individual words. The experimenter could explore different combinations of items using this tool and match them for age of acquisition. Or consider a researcher interested in gender effects on vocabulary growth: the appropriate tool would be the Vocabulary Norms analysis, which shows percentile curves for a particular instrument. (We walk through detailed instructions for how such an analysis would be conducted below.)

Database architecture

Why use a database to store vocabulary data? Consider the standard format of raw CDI data. Figure 2 shows a small slice of the original CDI norming data (Fenson *et al*., 1994, 2007). Each row is a child, each column gives a variable – either a demographic variable or the result of a particular word being administered to a particular child. Although this format is useful for homogeneous administrations of a single instrument, it cannot accommodate multiple instruments, multiple languages, or datasets with different sources or kinds of demographic information. Consolidating data across different instruments is very difficult in this format, and tracking data on children with multiple longitudinal administrations of a single instrument must also be done in an ad-hoc manner. The move to a database format allows far more flexible and programmatic handling of heterogeneous data structures from different sources.

A relational database such as Wordbank is at its heart a series of tables linked by unique identifiers. There are two primary groups of tables in Wordbank. The COMMON tables store data that is shared between CDI instruments, including information about children, administrations (individual instances of a form being filled out for a child), and items (words and other questions on a form). The INSTRUMENT tables store response data for particular CDI instruments. We currently include all items on CDI instruments, including questions about communication, gesture, morphology, and grammar (though in many of the datasets that we archive these non-vocabulary questions have not been digitized so data on them are sparse at present).

One strength of the Wordbank framework is that it allows the storage of subsidiary information about the words that are included in a particular instrument, so that this information can be used in future analyses. For example, information about grammatical and semantic categories or norms like concreteness and imageability could all be appended to particular words. This functionality is not yet present in Wordbank, however. The difficulty of compiling this kind of information for a particular set of

Fig. 2. Example data from the CDI norming sample (Fenson *et al*., 2007). Each row has a unique child identifier, demographics, and word-by-word checklist data.

words is compounded by the large number of languages that the database includes. We hope that in future this functionality will allow the gradual accumulation of information about the words included in the database.

Technical details. Wordbank is constructed using free, open-source tools. The database is a standard MySQL database, managed using Python and Django. Analysis apps are constructed using the Shiny package for R, an open-source statistical programming language. The code is hosted in a GitHub repository (<http://github.com/langcog/wordbank>) where interested users can browse, leave comments, and contribute modifications.

All data uploaded to Wordbank are open and freely available for download, both through the site itself and through the GitHub repository. The site includes only de-identified data that cannot be linked to the parents and children who provided it. Because of these features, the Stanford Institutional Review Board has determined that the Wordbank project does not constitute human subjects research.

Cross-linguistic and cross-instrument architecture. The general philosophy of creating CDIs for new languages has been summarized as "adaptation, not translation" (Dale, n.d.). In other words, CDIs are a useful tool for many languages, but the forms differ between languages – words and even whole sections are added, dropped, and modified to ensure that the form captures the details of the particular language for which it is designed. To date, more than sixty adaptations of the original English CDI have been documented (Dale and Penfold, n.d.). These forms vary widely, including differences in length and intended age range. Some forms include hundreds of items more than the original 680 words on the English Words & Sentences form; others are so-called 'short forms' and include only a hundred or a few hundred carefully selected words. Some are designed to capture development from the emergence of language through ages three to four years, while others are focused on very early development (like the English Words & Gestures form, designed for ages 8–18 months). All of these differences make it problematic to compare scores and score distributions across forms, even using percentile ranks, since some instruments will have more or more difficult items than others.

Wordbank is designed so that it can accommodate data from a wide variety of instruments, both within and across languages. Indeed, at the time of

writing, the site includes data from more than 42,000 administrations of the CDI across fourteen different languages and twenty-four different instruments. But because of the difficulties in comparison across instruments, our approach to cross-linguistic and cross-instrument data is to provide standardized analyses within each instrument and language, without assuming equivalence across words, instruments, or populations. Thus, our primary exploratory visualization tools in general do not allow comparison across languages, and we urge users to interpret cross-linguistic and cross-instrument differences with caution.[2] Developing statistical techniques to facilitate these comparisons is a current focus of our research.

Interactive analysis tools

The primary method for users to interact with Wordbank is through interactive analysis tools that are hosted on the website. These tools allow for fast and flexible exploration of the dataset, the results of which can be exported in tabular and graphical formats for further analysis and presentation.

Vocabulary Norms. One of the primary purposes of the CDI instruments is to provide percentile ranks for vocabulary growth across ages, both for visualizing the variability of early vocabulary growth and for examining differences in these growth patterns due to individual differences and demographic variables. Accordingly, Wordbank provides a Vocabulary Norms analysis, pictured in Figure 3. The inset plot shows all administrations of a particular CDI instrument within the instrument's valid age range. Dots show individual children, with age binned by month and jittered to avoid overplotting. Lines on the plot indicate estimates of percentiles, fit using quantile regression with monotonic polynomial splines as the base function (using the `gcrq` function of the `quantregGrowth` package; Muggeo, Sciandra, Tomasello & Calvo, 2013). An important feature of the norms app is that it can be split by any demographic field in the data, so that comparisons on variables like gender, birth order, or maternal education can be conducted.

The original and updated norming studies (Fenson *et al*., 1993, 2007) gathered data from a diverse (though not nationally representative) sample and used these data to construct normative curves from which percentile ranks could be derived. In contrast to these studies, Wordbank is not explicitly designed to provide stable, clinically relevant norms. Wordbank's sample is heterogeneous and continually growing, and its analyses are subject to revision and update. Thus, Wordbank does not currently generate percentile ranks, and we do not recommend that

[2] The only exception to this policy currently is that we allow users to see responses across instruments for individual words, in the Item Trajectories analysis (e.g., the proportion of children who say the word *cat* on both Words & Gestures and Words & Sentences forms).

Fig. 3. A screenshot of the Vocabulary Norms analysis tool, showing 10th, 25th, 50th, 75th, and 90th percentiles (default) for English production scores. Dots show individual administrations, jittered slightly to avoid overplotting. Curves show polynomial spline fits. (See text for more details; color online).

Wordbank-generated norming values be used for research or clinical purposes in which the goal is to evaluate children's performance in reference to an established normative standard. For these types of applications, users should refer to the published norms in the appropriate language.[3]

Item Trajectories. A second function of the CDI instruments is to provide aggregate data on the proportion of children at a particular age who know a specific word (Dale & Fenson, 1996; Jørgensen *et al*., 2010). Such analyses can be extremely helpful for the design and evaluation of materials for young children, including experimental stimuli. Accordingly, the second major interactive visualization in Wordbank is the Item Trajectories analysis tool.

[3] Users can always generate percentile ranks themselves, and this may be desirable or necessary for research purposes, but we caution against the clinical use of such ad-hoc norms.

This tool allows exploration of growth curves for individual words on a CDI form. Users can select a language and instrument (and choose production or comprehension where available), and then select or input a list of words whose trajectories are plotted (Figure 4). The 'both' measure option shows data from multiple forms for the same language, with different markers for each item. In general, our exploration suggests that there are only small differences across different instruments for the same item and age. Lines on the plot show a local polynomial regression smoothing line (`loess` in `R`).

Other features: static reports and tabular data download. In addition to the interactive analysis tools described above, Wordbank also includes a number of non-interactive but continuously updated reports on features like vocabulary composition across languages, links between grammar and the lexicon (Braginsky, Yurovsky, Marchman & Frank 2015), and gender differences in vocabulary growth (see below). On the Analyses page (<http://wordbank.stanford.edu/analyses>), we provide a gallery of both interactive and non-interactive analyses.

Wordbank also allows raw tabular data to be browsed and downloaded for subsequent analysis in all popular statistical packages. Using the same basic interface as the Vocabulary Norms and Item Trajectory tools, users can browse raw data aggregated across children (similar to the Vocabulary Norms tool), across items (similar to the Items Trajectory tool), or even view the raw subject-by-item data. All data in these 'standard' reports can be downloaded in CSV format.

A worked example: gender differences. Imagine a student interested in gender[4] differences in production vocabulary size, perhaps for a class project. Gender differences in language production are commonly found in individual studies (e.g. Fenson *et al*., 1994; Huttenlocher, Haight, Bryk, Seltzer & Lyons, 1991; see Wallentin, 2009, for review), and one large-scale previous study found differences in production vocabulary in ten languages (Eriksson *et al*., 2012).

To explore these differences using Wordbank, the student would navigate from the home page to the Vocabulary Norms report. English is the default language for the report, but the student could in principle select any language in the database. Similarly, she could select her desired instrument in the 'Forms' menu (Words & Sentences is the default). She would then select 'Gender' as a split variable for the data (in the 'Split Variable' menu) to see normative curves and sample sizes for each part of the dataset. Or, to

[4] The distinction between sex (biological characteristic) and gender (social characteristic) is complex, and not well understood in early childhood. We defer discussion of this issue; since the CDI is a parent-report form, we do not have access to either sex or gender information directly.

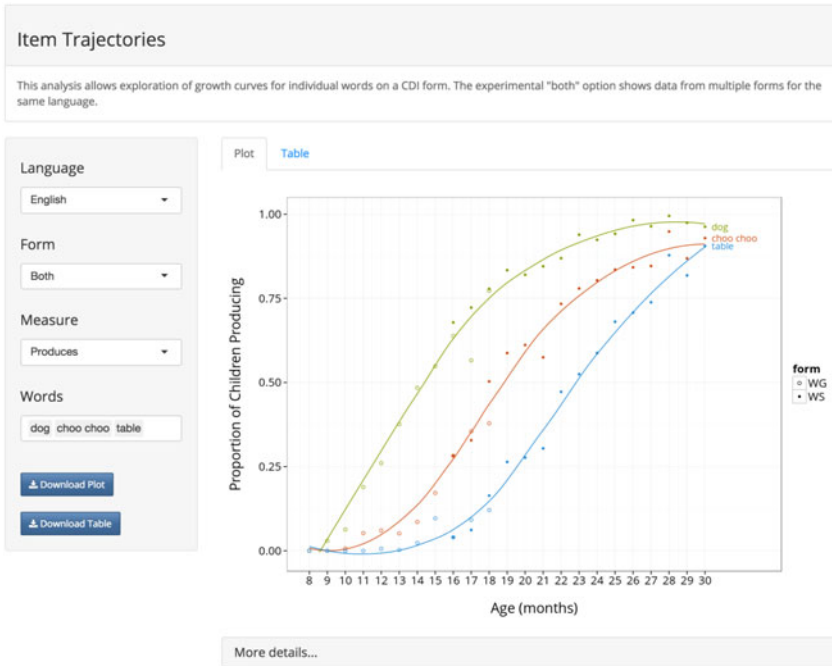

Fig. 4. A screenshot of the Item Trajectories analysis tool, showing a visualization of the developmental trajectory of production for three words (*dog, choo choo,* and *table*) across both Words & Gestures and Words & Sentences forms.

make a plot that enabled comparison of the median level of production vocabulary, she could select 'Median' in the 'Quantiles' menu.

Selecting 'Download Plot' would result in the plot shown in Figure 5. Or she could navigate to the 'Table' tab of the display window to see tabular form data showing the 50th percentile (median) for both females and males, by age. These tabular summary data are available for download via the 'Download Table' button, and the raw data (with a row for each one of the 4072 children represented in the plot) are available via the 'Download Raw Data' button. In sum, this graphical workflow allows interested users to manipulate and download individual parts of the dataset as well as to create visualizations of basic analyses.

Extensibility

Extensibility is one of the major strengths of Wordbank. Although programming knowledge is not necessary for interacting with Wordbank, interested researchers with programming skills can contribute to the

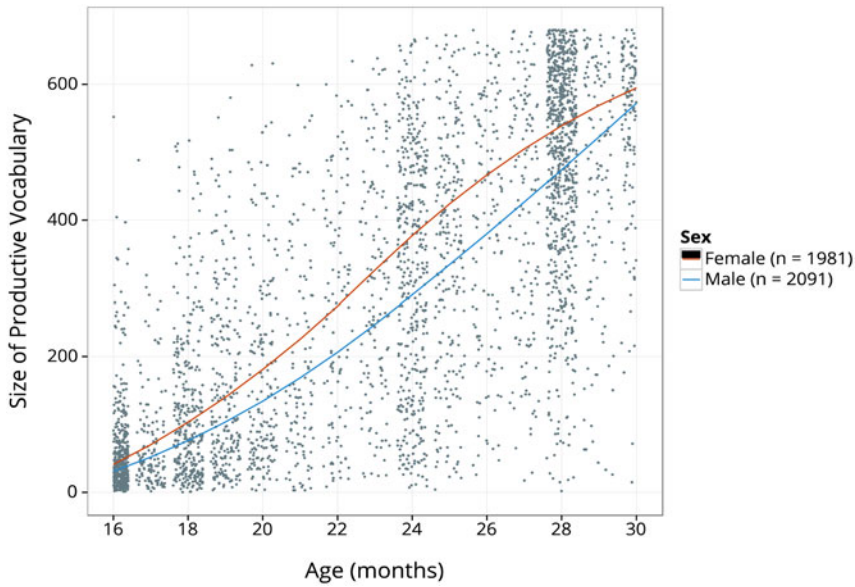

Fig. 5. A downloaded plot of gender differences in production language for English-speaking children (color online).

development effort by adding new analyses. Each Wordbank analysis app is constructed as a standalone script or set of scripts in the `R` language. Constructing an interactive analysis requires specifying a visualization and some interactive functionality using Shiny. Non-interactive analyses can be constructed as R Markdown documents that execute scripts using the Wordbank database. Both of these have the virtue of rerunning on the newest version of the database whenever they are opened, so they do not go out of date as new data are added.

In addition, we encourage contributions of individual datasets. Wordbank currently imports data from Excel and CSV formats via automated import scripts. Individuals or labs interested in contributing should consult with the authors for advice about data formatting and upload.

WORDBANKR: AN R PACKAGE FOR ACCESSING WORDBANK

Although the analysis tools described above suffice for many needs, researchers interested in detailed quantitative or cross-linguistic analyses may wish to connect directly to the Wordbank database and manipulate the data directly. Making use of the R programming language (R Foundation for Statistical Computing, 2014), we provide the `wordbankr` package to help researchers accomplish this task. 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 (e.g. Norrman & Bylund, 2016; Song, Shattuck-Hufnagel & Demuth, 2015). The `wordbankr` package abstracts away the details of connecting to the database. Users can take advantage of the SQL tools developed in the popular `dplyr` package (Wickham & Francois, 2014), which make manipulating large datasets quick and easy. We describe the commands that the package provides and then give a worked example of using the package for a simple analysis.

Package details

The `wordbankr` package is easily installed via CRAN, the comprehensive R archive network. To install, simply type: `install.packages("wordbankr")`. After installation, users can use the three main data loading functions provided by `wordbankr::get_administration_data` to retrieve information about each CDI administration, including the child's demographics and vocabulary sizes; `get_item_data` to retrieve information about each CDI item, including its text and categories; and `get_instrument_data` to retrieve administration-by-item response values. Each of these can be run in remote mode, which loads data from the Wordbank server, or in local mode if the user has a copy of the database set up on their local machine. For more detailed documentation, see the package repository (<http://github.com/langcog/wordbankr>).

Worked example, part 2: gender differences across languages

We next demonstrate the analytic potential of direct manipulation of the Wordbank database using `wordbankr`, by using the package to extend the worked example of gender differences above. This section also replicates a large-scale analysis by Eriksson *et al*. (2012). To perform the analysis, we first begin by using `wordbankr` to load the data from Wordbank and connect to the tables we need:

```
admins <- get_administration_data()
items <- get_item_data()
```

We next use a series of dplyr calls to compute the number of words in each language, select the appropriate subset of the data, and calculate the proportion of words produced for this data subset:

```
num_words <- items %>%
  filter(form == "WS", type == "word") %>%
  group_by(language) %>%
  summarise(n = n())
```

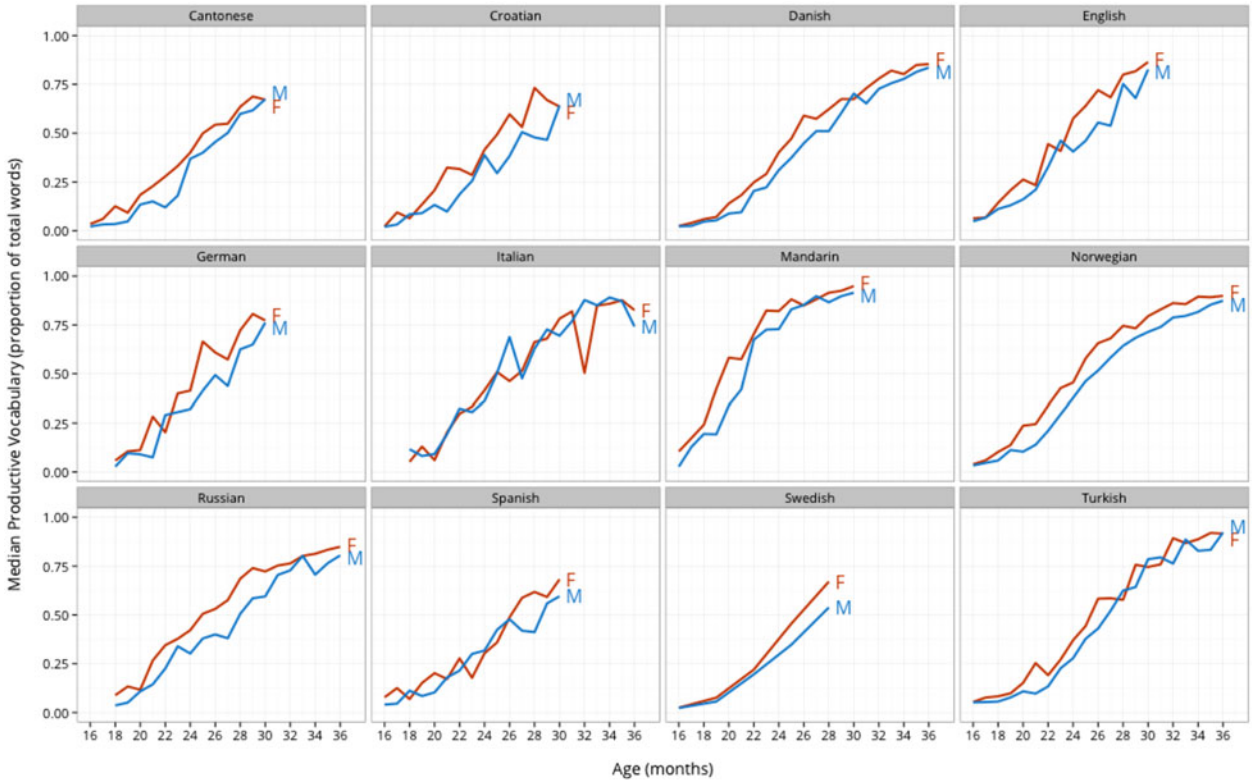

Fig. 6. Median production vocabulary as a proportion of total words on an instrument, plotted by age in months. Red and blue lines show females and males, respectively (color online).

```
vocab_admins <- admins %>%
    filter(form == "WS", !is.na(sex)) %>%
    select(data_id, language, form, age, sex, production)
vocab_data <- vocab_admins %>%
    group_by(language, sex, age) %>%
    left_join(num_words) %>%
    mutate(production = production / n) %>%
    summarise(median = median(production))
```

We then plot the `vocab_data` data frame using the `ggplot2` package (Wickham, 2009). Full code for the analysis as a whole (including the plot) is available at <http://mikabr.github.io/demo-vocab/gender.html>.

The results of this analysis are shown in Figure 6. As expected, we replicate the gender differences found in previous work (Eriksson *et al.*, 2012): females showed a small but highly reliable advantage in early production. This effect is highly consistent and clearly visible in eleven out of twelve languages, with Italian being the only exception. For comparison, the previous work found a positive female effect for all ten out of ten languages, but the size of the effect was close to zero for two of these. Observational data such as those contained in Wordbank allow us only to speculate about the origins of this gender difference or the sources of cross-linguistic variation (for some discussion, see Eriksson *et al.*, 2012). But the Wordbank platform dramatically facilitates the formulation and testing of analyses of this sort, allowing hypotheses to be tested quickly and easily against large datasets.

CONCLUSION

In this paper, we have presented Wordbank, an open repository for parent-report vocabulary data from the MacArthur-Bates CDI. The interactive analysis tools available on the Wordbank site allow interested researchers to explore a wide variety of phenomena in vocabulary development quickly and easily, exporting data and downloading presentation-quality graphics that document their analysis. In addition, users can contribute new analyses and data to the site and connect to it directly using an `R` package for data loading. These functions all facilitate greater sharing and reuse of existing data on children's vocabulary, enabling new discoveries in the future.

REFERENCES

Bates, E. (1976). *Language and context: the acquisition of pragmatics* (Vol. 13). New York, NY: Academic Press.

Bates, E. & Goodman, J. (1999). On the emergence of grammar from the lexicon. In B. MacWhinney (ed.), *The emergence of language* (pp. 29–79). Mahwah, NJ: Lawrence Erlbaum Associates.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., … Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language* **21**, 85–123.

Bloom, P. (2002). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Bornstein, M. H. & Haynes, O. M. (1998). Vocabulary competence in early childhood: measurement, latent construct, and predictive validity. *Child Development* **69**, 654–71.

Braginsky, M., Yurovsky, D., Marchman, V. A. & Frank, M. C. (2015). Developmental changes in the relationship between grammar and the lexicon. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Brown, R. (1973). *A first language: the early stages*. Cambridge, MA: Harvard University Press.

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N. & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences* **110**, 11278–83.

Clark, E. (2003). *First language acquisition*. Cambridge: Cambridge University Press.

Dale, P. S. (n.d.). Adaptations, not translations! Online: <http://mb-cdi.stanford.edu/adaptations.html> (last accessed 2015).

Dale, P. S. & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments & Computers* **28**, 125–7.

Dale, P. S. & Penfold, M. (n.d.). Adaptations of the MacArthur-Bates CDI into non-US English languages. Online: <http://mb-cdi.stanford.edu/documents/AdaptationsSurvey7-5-11Web.pdf> (last accessed 2011).

Dickinson, D. K. & Tabors, P. O. (2001). *Beginning literacy with language: young children learning at home and school*. Baltimore, MD: Paul H. Brookes Publishing.

Dunn, L. M. & Dunn, L. M. (2007). *Peabody Picture Vocabulary Test*, 4th ed. Parsippany, NJ: AGS Publishing / Pearson Assessments.

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**, 326–43.

Feldman, H. M., Dale, P. S., Campbell, T. F., Colborn, D. K., Kurs-Lasky, M., Rockette, H. E. & Paradise, J. L. (2005). Concurrent and predictive validity of parent reports of child language at ages 2 and 3 years. *Child Development* **76**, 856–68.

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E. & Paradise, J. L. (2000). Measurement properties of the MacArthur Communicative Development Inventories at ages one and two years. *Child Development* **71**, 310–22.

Fenson, L., Bates, E., Dale, P., Goodman, J., Reznick, J. S. & Thal, D. (2000). Reply: measuring variability in early child language: don't shoot the messenger. *Child Development* **71**, 323–8.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Hartung, J. P., Pethick, S. & Reilly, J. (1993). *MacArthur Communicative Development Inventories: user's guide and technical manual*. Baltimore, MD: Paul H. Brookes Publishing Co.

Fenson, L., Dale, P., Reznick, J., Bates, E., Thal, D., Pethick, S., … Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development* **59**.

Fenson, L., Marchman, V. A., Thal, D., Dale, P., Reznick, J. S. & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: user's guide and technical manual*, 2nd ed. Baltimore, MD: Brookes Publishing Company.

Hidaka, S. (2016). Estimating the latent number of types in growing corpora with reduced cost–accuracy trade-off. *Journal of Child Language* **43**, 1–28.

Hills, T. T., Maouene, J., Riordan, B. & Smith, L. B. (2010). The associative structure of language: contextual diversity in early word learning. *Journal of Memory and Language* **63**, 259–73.
Hills, T. T., Maouene, M., Maouene, J., Sheya, A. & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science* **20**, 729–39.
Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M. & Lyons, T. (1991). Early vocabulary growth: relation to language input and gender. *Developmental Psychology* **27**, 236–48.
Jørgensen, R. N., Dale, P. S., Bleses, D. & Fenson, L. (2010). CLEX: a cross-linguistic lexical norms database. *Journal of Child Language* **37**, 419–28.
Kristoffersen, K. E., Simonsen, H. G., Bleses, D., Wehberg, S., Jørgensen, R. N., Eiesland, E. A. & Henriksen, L. Y. (2013). The use of the Internet in collecting CDI data – an example from Norway. *Journal of Child Language* **40**, 567–85.
Lieven, E., Salomo, D. & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: a usage-based analysis. *Cognitive Linguistics* **20**, 481–507.
MacWhinney, B. (2000). *The CHILDES Project: tools for analyzing talk*, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates.
Marchman, V. A. & Martínez-Sussmann, C. (2002). Concurrent validity of caregiver/parent report measures of language for children who are learning both English and Spanish. *Journal of Speech, Language, and Hearing Research* **45**, 983–97.
Mayor, J. & Plunkett, K. (2011). A statistical estimate of infant and toddler vocabulary size from CDI analysis. *Developmental Science* **14**, 769–85.
Muggeo, V. M., Sciandra, M., Tomasello, A. & Calvo, S. (2013). Estimating growth charts via nonparametric quantile regression: a practical framework with application in ecology. *Environmental and Ecological Statistics* **20**, 519–31.
Nelson, K. (1973). Structure and strategy in learning to talk. *Monographs of the Society for Research in Child Development* **38**, 1–135.
Norrman, G. & Bylund, E. (2016). The irreversibility of sensitive period effects in language development: evidence from second language acquisition in international adoptees. *Developmental Science* **19**, 513–20.
R Foundation for Statistical Computing (2014). *R: a language and environment for statistical computing*. Software, online: <http://www.r-project.org>.
Rescorla, L. (1989). The language development survey: a screening tool for delayed language in toddlers. *Journal of Speech and Hearing Disorders* **54**, 587–99.
Roy, B. C., Frank, M. C., DeCamp, P., Miller, M. & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences* **112**, 12663–68.
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**, 152–69.
Tardif, T., Fletcher, P., Liang, W., Zhang, Z., Kaciroti, N. & Marchman, V. A. (2008). Baby's first 10 words. *Developmental Psychology* **44**, 929–38.
Thal, D., Jackson-Maldonado, D. & Acosta, D. (2000). Validity of a parent-report measure of vocabulary and grammar for Spanish-speaking toddlers. *Journal of Speech, Language, and Hearing Research* **43**, 1087–100.
Tomasello, M. & Mervis, C. B. (1994). The instrument is great, but measuring comprehension is still a problem. *Monographs of the Society for Research in Child Development* **59**, 174–9.
Wallentin, M. (2009). Putative sex differences in verbal abilities and language cortex: a critical review. *Brain and Language* **108**, 175–83.
Weisleder, A. & Fernald, A. (2013). Talking to children matters: early language experience strengthens processing and builds vocabulary. *Psychological Science* **24**, 2143–52.
Wickham, H. (2009). *Ggplot2: elegant graphics for data analysis*. New York, NY: Springer Science & Business Media.
Wickham, H. & Francois, R. (2014). *Dplyr: a grammar of data manipulation*. R package version 0·3·0·2. Online: <https://cran.r-project.org/web/packages/dplyr>.