

1 Peekbank: Exploring children's word recognition through an open, large-scale repository for
2 developmental eye-tracking data

3 Peekbank team², Martin Zettersten¹, Claire Bergey², Naiti S. Bhatt³, Veronica Boyce⁴, Mika
4 Braginsky⁵, Alexandra Carstensen⁴, Benny deMayo¹, George Kachergis⁴, Molly Lewis⁶, Bria
5 Long⁴, Kyle MacDonald⁷, Jessica Mankewitz⁴, Stephan Meylan^{5,8}, Annissa N. Saleh⁹, Rose
6 M. Schneider¹⁰, Angeline Sin Mei Tsui⁴, Sarp Uner⁸, Tian Linger Xu¹¹, Daniel Yurovsky⁶, &
7 Michael C. Frank¹

8 ¹ Dept. of Psychology, Princeton University

9 ² Dept. of Psychology, University of Chicago

10 ³ Scripps College

11 ⁴ Dept. of Psychology, Stanford University

12 ⁵ Dept. of Brain and Cognitive Sciences, MIT

13 ⁶ Dept. of Psychology, Carnegie Mellon University

14 ⁷ Core Technology, McD Tech Labs

15 ⁸ Dept. of Psychology and Neuroscience, Duke University

16 ⁹ Dept. of Psychology, UT Austin

17 ¹⁰ Dept. of Psychology, UC San Diego

18 ¹¹ Dept. of Psychological and Brain Sciences, Indiana University

Author Note

19

20 Add complete departmental affiliations for each author here. Each new line herein
21 must be indented, like this line.

22 Enter author note here.

23 The authors made the following contributions. Peekbank team: Conceptualization,
24 Writing - Original Draft Preparation, Writing - Review & Editing.

25 Correspondence concerning this article should be addressed to Peekbank team, Postal
26 address. E-mail: my@email.com

Abstract

27

28 The ability to rapidly recognize words and link them to referents in context is central to
29 children's early language development. This ability, often called word recognition in the
30 developmental literature, is typically studied in the looking-while-listening paradigm, which
31 measures infants' fixation on a target object (vs. a distractor) after hearing a target label.
32 We present a large-scale, open database of infant and toddler eye-tracking data from
33 looking-while-listening tasks. The goal of this effort is to address theoretical and
34 methodological challenges in measuring vocabulary development. [tools; processing; analysis/
35 usage examples]

36 *Keywords:* keywords

37 Word count: X

Peekbank: Exploring children’s word recognition through an open, large-scale repository for developmental eye-tracking data

Introduction

Across their first years of life, children learn words in their native tongues at a rapid pace (Frank, Braginsky, Yurovsky, & Marchman, 2021). [notes about the size/ pace] A key part of the word learning process is children’s emerging ability to rapidly process words and link them to relevant meanings – often referred to as word recognition. Measuring early word recognition offers insight into children’s early word representations and the processes supporting early language comprehension (Bergelson, 2020). Word recognition skills are also thought to build a foundation for children’s subsequent language development. Past research has found that early word recognition efficiency is predictive of later linguistic and general cognitive outcomes (Bleses, Makransky, Dale, Højen, & Ari, 2016; Marchman et al., 2018). While word recognition is a central part of children’s language development, mapping the trajectory of word recognition skills has remained elusive. Studies investigating children’s word recognition are typically limited in scope to experiments in individual labs involving small samples tested on a limited set of items. This limitation in scale makes it difficult to understand developmental changes in children’s word knowledge at a broad scale. Peekbank provides an openly accessible database of eye-tracking data of children’s word recognition, with the primary goal of facilitating the study of developmental changes in children’s word knowledge and recognition speed.

The “Looking-While-Listening” Paradigm

Word recognition is traditionally studied in the “looking-while-listening” paradigm (alternatively referred to as the intermodal preferential looking procedure; Fernald, Zangl, Portillo, & Marchman, 2008; Hirsh-Pasek, Cauley, Golinkoff, & Gordon, 1987). In such studies, infants listen to a sentence prompting a specific referent (e.g., Look at the dog!) while viewing two images on the screen (e.g., an image of a dog – the target image – and an

image of a duck – the distractor image). Infants’ word recognition is measured in terms of how quickly and accurately they fixate on the correct target image after hearing its label. Past research has used this same basic method to study a wide range of questions in language development. For example, the looking-while-listening paradigm has been used to uncover early knowledge of nouns in infants’ early noun knowledge, phonological representations of words, prediction during language processing, and individual differences in language development (Bergelson & Swingley, 2012; Golinkoff, Ma, Song, & Hirsh-Pasek, 2013; Lew-Williams & Fernald, 2007; Marchman et al., 2018; Swingley & Aslin, 2000).

Measuring developmental change in word recognition

While the looking-while-listening paradigm has been highly fruitful in advancing understanding of early word knowledge, fundamental questions remain both about the trajectory of children’s word recognition ability and the nature of the method itself. One central question is how to measure developmental change in word recognition. A key idea in the language learning literature is that processing speed - the ability to quickly link a word with its referent - supports language learning. Age-related changes in speed of processing are thought to accelerate infants’ subsequent language learning: the faster infants are able to process incoming speech input, the better able they become to learn from their language environment. Similarly, longitudinal analyses have found that individual differences in word recognition speed predict linguistic and cognitive outcomes later in childhood (e.g., Marchman & Fernald, 2008). However, measuring increases in the speed and accuracy of word recognition faces the challenge of distinguishing developmental changes in word recognition skill from changes in knowledge of specific words. This problem is particularly thorny in child development, since the number of items that can be tested within a single session is limited and items must be selected in an age-appropriate manner (Peter et al., 2019). Measuring developmental change therefore requires large-scale datasets with a range of items, in order to generalize age-related changes across words.

Developing methodological best-practices

A second question relates to evaluating methodological best practices. In particular, many fundamental analytic decisions vary substantially across studies, and different decisions may lead to different inferences about children’s word recognition. For example, researchers vary in how they select time windows for analysis, transform the dependent measure of target fixations, and model the time course of word recognition (Csibra, Hernik, Mascaro, Tatone, & Lengyel, 2016; Fernald et al., 2008; Huang & Snedeker, 2020). This problem is made more complex by the fact that many of these decisions depend on a variety of design-related and participant-related factors (e.g., infant age). Establishing best practices therefore requires a large database of infant word recognition studies varying across such factors, in order to test the potential consequences of methodological decisions on study results.

Peekbank: An open database of developmental eye-tracking studies.

What these two questions share is that they are difficult to answer at the scale of a single study. To address this challenge, we introduce Peekbank, a flexible and reproducible interface to an open database of developmental eye-tracking studies. The Peekbank project (a) collects a large set of eye-tracking datasets on children’s word recognition, (b) introduces a data format and processing tools for standardizing eye-tracking data across data sources, and (c) provides an interface for accessing and analyzing the database. In the current paper, we give an overview of the key components of the project and some initial demonstrations of its utility in advancing theoretical and methodological insights. We report two analyses using the database and associated tools (N=1,233): (1) a growth curve analysis modeling age-related changes in infants’ word recognition while generalizing across item-level variability; and (2) a multiverse-style analysis of how a central methodological decision – selecting the time window of analysis – impacts inter-item reliability.

Design and Technical Approach

Database Framework.

One of the main challenges in compiling a large-scale eye-tracking dataset is the lack of a shared re-usable data format across individual experiments. Researcher conventions for structuring data vary, as do the technical specifications of different devices, rendering the task of integrating datasets from different labs and data sources difficult. Therefore, our first effort was developing a common, tidy format for the eye-tracking data in Peekbank to ease the process of conducting cross-dataset analyses (Wickham et al., 2019). All incoming datasets will be processed into this unified format and then ingested into the core database.

As illustrated in Figure XX, the Peekbank framework consists of three main components: (1) processing eye-tracking experimental datasets into a unified format; (2) populating a relational database; and (3) providing an interface to the database. These components are supported by three libraries. The peekds library (for the R language; R Development Core Team, 2020) helps researchers convert and validate existing datasets to use the relational format of the database. The peekbank module (Python) creates a database with the relational schema and populates it with the standardized datasets produced by peekds. The database is implemented in MySQL, an industry standard relational database, which may be accessed by a variety of programming languages over the internet. The peekbankr library (R) provides an application programming interface, or API, that offers high-level abstractions for accessing data in Peekbank.

In the following sections, we will begin by providing the details on the database's scheme design and technical implementation on peekds. For users who are primarily interested in accessing the database can skip these details and focus on access through the peekbankr API and the web apps.

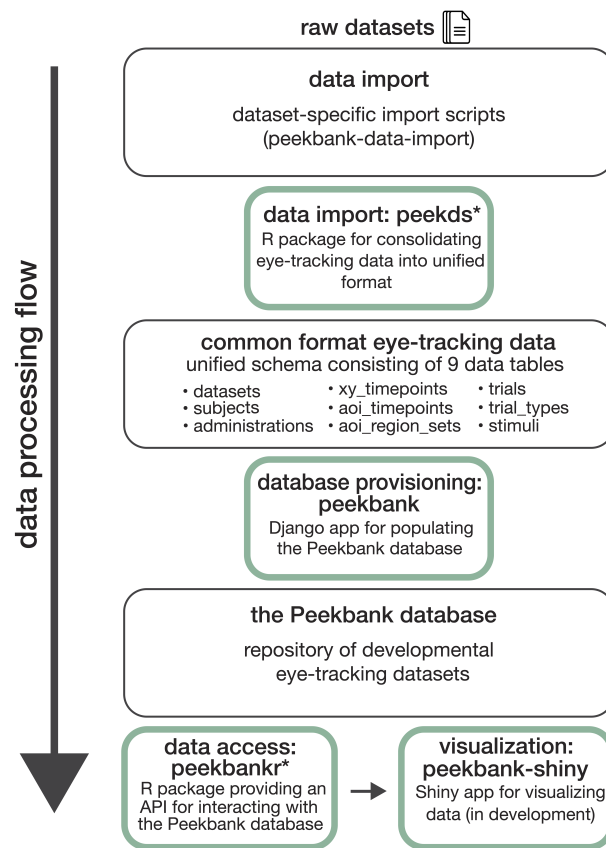


Figure 1. Overview of the Peekbank data ecosystem. Peekbank tools are highlighted in green.

*custom R packages.

Database Schema

The peekbank database has two main sets of data: (1) metadata on the experiment design and participants and (2) timecourse looking data (Fig XX). Here, we will give an outline of the tables encoding this data, starting with metadata.

Metadata can be separated into two parts: (1) subject-level demographics and (2) trial structure information. Subject-level information includes the `subjects` table, information about individuals who participated in the study, and the `administrations` table, information about runs of the experiment. In the case of longitudinal studies, multiple administrations of an experiment can be linked to a single subject or participant. The `subjects` table contains information at the subject level, such as a subject's sex and native

language, and the **administrations** table includes information specific to a particular run of the experiment, such as the subject’s age at the time of the experiment and the eyecoding method for that administration.

The **stimuli** and **trial_types** tables store information about the experiment design. Stimuli are (label, image) mappings that are seen in the experiment. When the experiment includes multiple potential labels per image (e.g., in designs testing understanding of multiple languages or synonymy), images can have multiple rows in the **stimuli** table for unique labels as well as a row with no label to be used when the image appears solely as a distractor (and thus its label is ambiguous). The **trial_types** table encodes information about each trial of the experiment including the target stimulus and location, the distractor stimulus and location, and the point of disambiguation for that trial. If this dataset used automatic eyetracking rather than manual coding, each trial type is additionally linked to a set of area of interest (x, y) coordinates, encoded in the **aoi_region_sets** table.

Because individual trial types can be repeated multiple times within an administration, the order of the trials is encoded in the **trials** table. Each unique ordering that occurred in the experiment is encoded—for example, if every participant saw the same ordering, the trials table would only have as many rows as there were trials in the experiment; if there were many different orderings, the trials table would represent each ordering. The **trial_id**, which links a trial type to the order it was presented in an administration, is attached to the time course looking data.

Timecourse looking data is encoded in two tables: **aoi_timepoints** and **xy_timepoints**. The **aoi_timepoints** table encodes time-stamped looks to areas of interest: looks to the target, looks to the distractor, looks to other places, and missing looks. All datasets have this table, and it is the centerpiece of the schema where looking behavior is recorded. Additionally, for eyetracking data that is automatically rather than manually coded, the **xy_timepoints** table encodes the (x, y) coordinates of looks over the course of

the experiment. Both the `aoi_timepoints` and `xy_timepoints` tables are resampled to a consistent sampling rate, as described below.

Import

During data import, raw eye-tracking datasets are processed to conform to the Peekbank data schema. The following section is a description of the import process for peekbank. It serves as both a description of our method in importing the datasets already in the database, as well as a high-level overview of the import process for researchers looking to import their data in the future. First, we will describe the import of metadata, and second, we will describe import of the timecourse looking data, including processing functions in `peekds` for normalizing and resampling looking behavior.

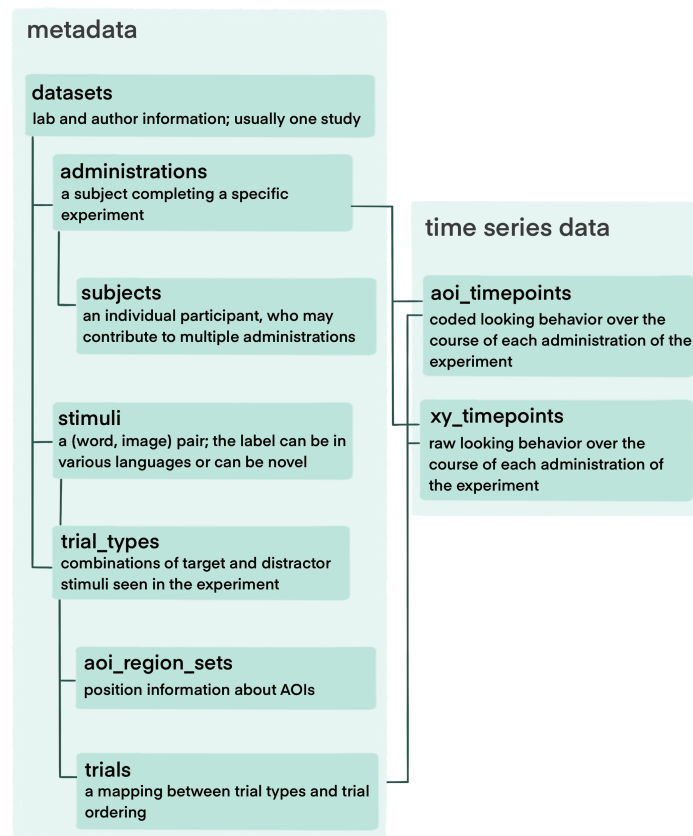


Figure 2. The Peekbank schema. Each square represents a table in the relational database.

Dataset Name	Citation	N	Mean Age (mos.)	Age Range (mos.)	Method	Language
attword	(Yurovsky & Frank, 2017)	288	25.5	13 - 59	eye-tracking	English
canine	unpublished	36	23.8	21 - 27	manual coding	English
coartic	(Mahr et al., 2015)	29	20.8	18 - 24	eye-tracking	English
cowpig	(Perry et al., 2017)	45	20.5	19 - 22	manual coding	English
ft_pt	(Adams et al., 2018)	69	17.1	13 - 20	manual coding	English
mispron	(Swingley & Aslin, 2002)	50	15.1	14 - 16	manual coding	English
mix	(Byers-Heinlein et al., 2017)	48	20.1	19 - 21	eye-tracking	English, French
reflook_socword	(Yurovsky et al., 2013)	435	33.6	12 - 70	eye-tracking	English
reflook_v4	unpublished	45	34.2	11 - 60	eye-tracking	English
remix	(Potter et al., 2019)	44	22.6	18 - 29	manual coding	Spanish, English
salientme	(Pomper & Saffran, 2019)	44	40.1	38 - 43	manual coding	English
switchingCues	(Pomper & Saffran, 2016)	60	44.3	41 - 47	manual coding	English
tablet	(Frank et al., 2016)	69	35.5	12 - 60	eye-tracking	English
tseltal	(Casillas et al., 2017)	23	31.3	9 - 48	manual coding	Tseltal
yoursmy	(Garrison et al., 2020)	35	14.5	12 - 18	eye-tracking	English

Table 1

Overview over the datasets in the current database.

Current Data Sources.

The database currently includes 11 looking-while-listening datasets comprising $N=1320$ total participants (Table XX). Most datasets (10 out of 11 total) consist of data from monolingual native English speakers. They span a wide age spectrum with participants ranging from 8 to 84 months of age, and are balanced in terms of gender (48% female). The datasets vary across a number of dimensions related to design and methodology, and include studies using manually coded video recordings and automated eye-tracking methods (e.g., Tobii, EyeLink) to measure gaze behavior. Most studies focused on testing familiar items, but the database also includes studies with novel pseudowords. All data (and accompanying references) are openly available on the Open Science Framework (https://osf.io/pr6wu/?view_only=07a3887eb7a24643bdc1b2612f2729de).

How selected? Language coverage? More details about lab and design variation?

Versioning + Expanding the database

Information about versioning approach/ regularity of updates Steps for extending the database?

Interfacing with peekbank

Shiny App

One goal of the Peekbank project is to allow a wide range of users to easily explore and learn from the database. We therefore have created an interactive web application – `peekbank-shiny` – that allows users to quickly and easily create informative visualizations of individual datasets and aggregated data. `peekbank-shiny` is built using Shiny, a software package for creating web apps using R. The Shiny app allows users to create commonly used visualizations of looking-while-listening data, based on data from the Peekbank database. Specifically, users can visualize

1. the time course of looking data in a profile plot depicting infant target looking across trial time,
2. overall accuracy (proportion target looking) within a specified analysis window,
3. reaction times (speed of fixating the target image) in response to a target label, and
4. an onset-contingent plot, which shows the time course of participant looking as a function of their look location at the onset of the target label.

Users are given various customization options for each of these visualizations, e.g., choosing which datasets to include in the plots, controlling the age range of participants, splitting the visualizations by age bins, and controlling the analysis window for time course analyses. Plots are then updated in real time to reflect users' customization choices, and users are given options to share the visualizations they created. The Shiny app thus allows users to quickly inspect basic properties of Peekbanks datasets and create reproducible visualizations without incurring any of the technical overhead required to access the

221 database through R.

222 Peekbankr

223 Functions: `connect_to_peekbank()`: `get_datasets()` `get_subjects()`
 224 `get_administrations()` `get_stimuli()` `get_aoi_timepoints()` `get_trials()` `get_trial_types()`
 225 `get_xy_timepoints()` `get_aoi_region_sets()`

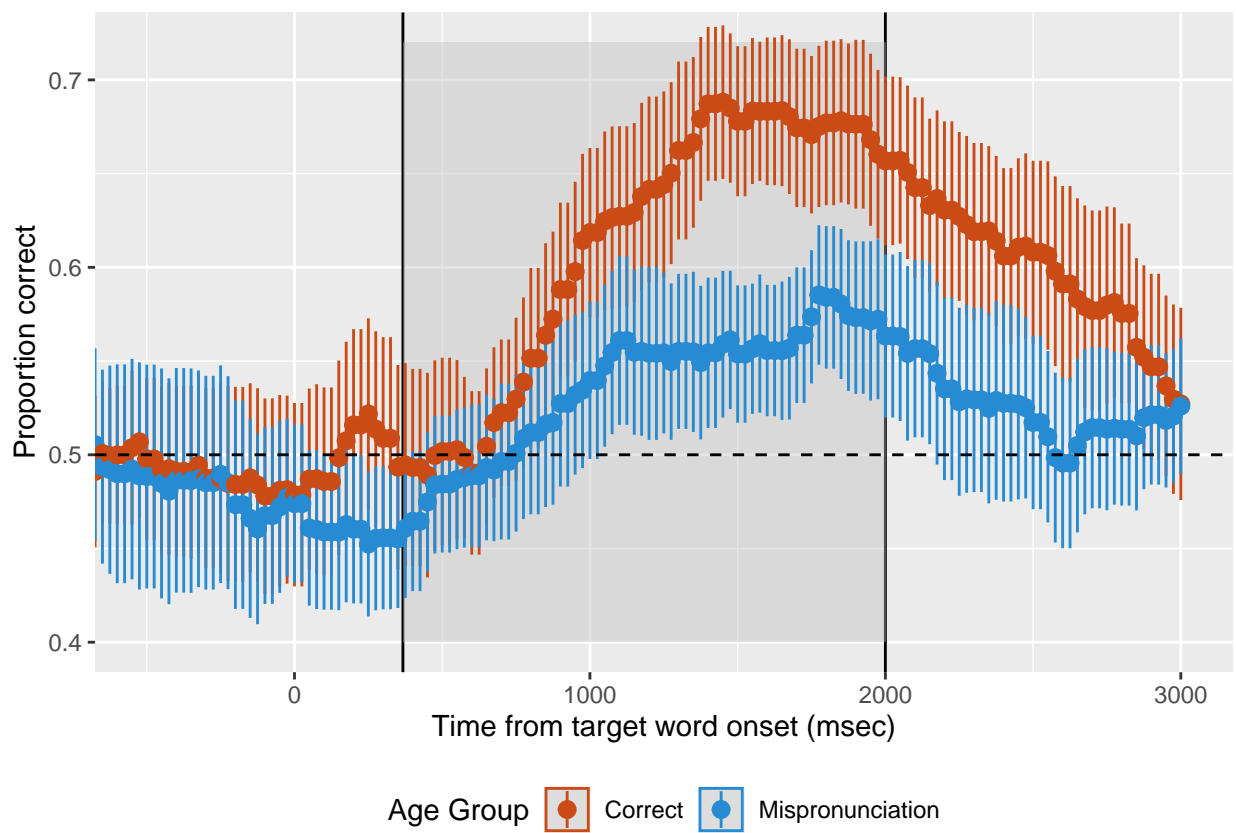
226 OSF site

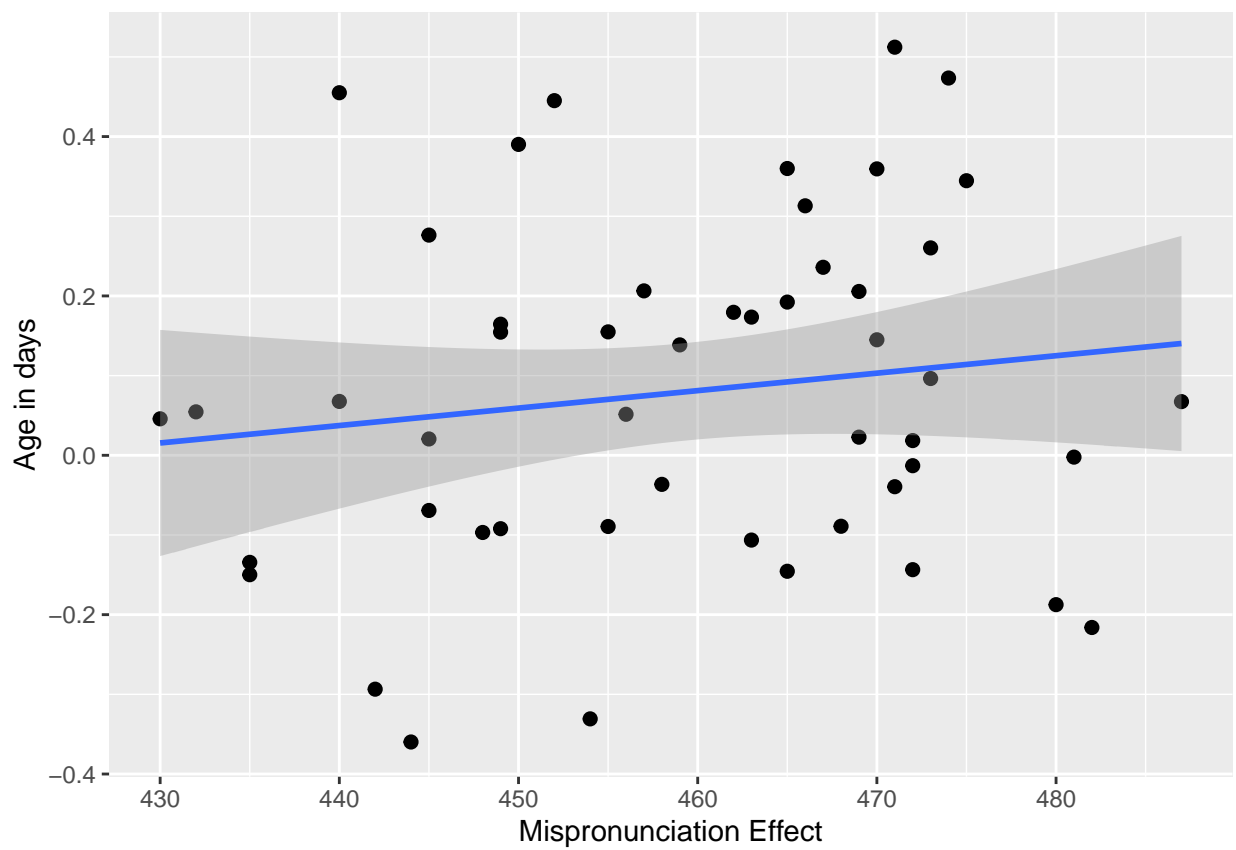
227 Stimuli Data in raw format (if some additional datum needed, e.g. pupil size?)

228 Peekbank in Action

229 We provide two potential use-cases for Peekbank data. In each case, we provide sample
 230 code so as to model how easy it is to do simple analyses using data from the database. Our
 231 first example shows how we can replicate the analysis for a classic study. This type of
 232 computational reproducibility can be a very useful exercise for teaching students about best
 233 practices for data analysis (e.g., Hardwicke et al., 2018) and also provides an easy way to
 234 explore looking-while-listening timecourse data in a standardized format. Our second
 235 example showss an in-depth exploration of developmental changes in the recognition of
 236 particular words. Besides its theoretical interest (which we will explore more fully in
 237 subsequent work), this type of analysis can be used for optimizing the stimuli for new
 238 experiments.

239 Computational reproducibility example: Swingley & Aslin (2002)





241

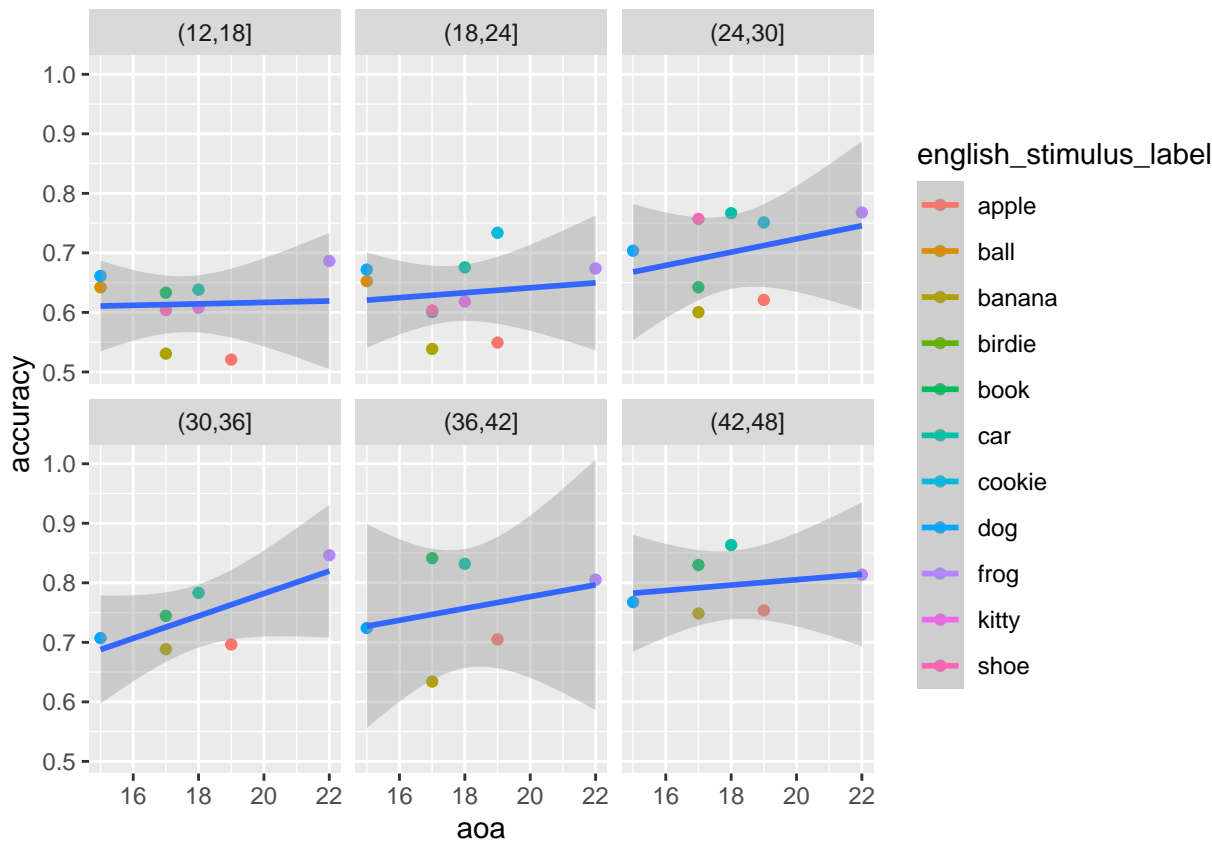
242 Item analyses

243 To illustrate the power of aggregating data across multiple datasets, we,

244 aspirational goal -

245 Also, item selection but maybe not yet?

Links to parent report vocabulary data



Discussion/ Conclusion

Theoretical progress in understanding child development requires rich datasets, but collecting child data is expensive, difficult, and time-intensive. Recent years have seen a growing effort to build open source tools and pool research efforts to meet the challenge of building a cumulative developmental science (Bergmann et al., 2018; Frank, Braginsky, Yurovsky, & Marchman, 2017; The ManyBabies Consortium, 2020). The Peekbank project expands on these efforts by building an infrastructure for aggregating eye-tracking data across studies, with a specific focus on the looking-while-listening paradigm. This paper presents an illustration of some of the key theoretical and methodological questions that can be addressed using Peekbank: generalizing across item-level variability in children’s word recognition and providing data-driven guidance on methodological choices.

There are a number of limitations surrounding the current scope of the database. A priority in future work will be to expand the size of the database. With 11 datasets currently available in the database, idiosyncrasies of particular designs and condition manipulations still have substantial influence on modeling results. Expanding the set of distinct datasets will allow us to increase the number of observations per item across datasets, leading to more robust generalizations across item-level variability. The current database is also limited by the relatively homogeneous background of its participants, both with respect to language (almost entirely monolingual native English speakers) and cultural background (all but one dataset comes from WEIRD populations; Muthukrishna et al., 2020). Increasing the diversity of participant backgrounds and languages will expand the scope of the generalizations we can form about child word recognition. Finally, while the current database is focused on studies of word recognition, the tools and infrastructure developed in the project can in principle be used to accommodate any eye-tracking paradigm, opening up new avenues for insights into cognitive development. Gaze behavior has been at the core of many of the key advances in our understanding of infant cognition. Aggregating large datasets of infant looking behavior in a single, openly-accessible format promises to bring a fuller picture of infant cognitive development into view.

Acknowledgements

We would like to thank the labs and researchers that have made their data publicly available in the database.

We used R [Version 4.0.3; R Core Team (2020)] and the R-packages *dplyr* [Version 1.0.5; Wickham, François, Henry, and Müller (2021)], *extrafont* [Version 0.17; Winston Chang (2014)], *forcats* [Version 0.5.1; Wickham (2021a)], *ggplot2* [Version 3.3.3; Wickham (2016)], *here* [Version 1.0.1; Müller (2020)], *papaja* [Version 0.1.0.9997; Aust and Barth (2020)], *peekbankr* [Version 0.1.1.9000; Braginsky, MacDonald, and Frank (2021)], *plyr* [Version 1.8.6; Wickham, François, Henry, and Müller (2021); Wickham (2011)], *png* [Version

285 0.1.7; Urbanek (2013)], *pso* (Bendtsen., 2012), *purrr* [Version 0.3.4; Henry and Wickham
286 (2020)], *readr* [Version 1.4.0; Wickham and Hester (2020)], *stringr* [Version 1.4.0; Wickham
287 (2019)], *tibble* [Version 3.1.1; Müller and Wickham (2021)], *tidyr* [Version 1.1.3; Wickham
288 (2021b)], *tidyverse* [Version 1.3.0; Wickham et al. (2019)], and *xtable* [Version 1.8.4; Dahl,
289 Scott, Roosen, Magnusson, and Swinton (2019)] for all our analyses.

References

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*.

Retrieved from <https://github.com/crsh/papaja>

Bendtsen., C. (2012). *Pso: Particle swarm optimization*. Retrieved from

<https://CRAN.R-project.org/package=pso>

Braginsky, M., MacDonald, K., & Frank, M. (2021). *Peekbankr: Accessing the*

peekbank database. Retrieved from <http://github.com/langcog/peekbankr>

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., & Swinton, J. (2019). *Xtable:*

Export tables to LaTeX or HTML. Retrieved from

<https://CRAN.R-project.org/package=xtable>

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C.,

Kidwell, M. C., . . . Frank, M. C. (2018). Data availability, reusability, and

analytic reproducibility: Evaluating the impact of a mandatory open data policy

at the journal Cognition. *Royal Society Open Science*, 5(8).

<https://doi.org/10.1098/rsos.180448>

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools*. Retrieved

from <https://CRAN.R-project.org/package=purrr>

Müller, K. (2020). *Here: A simpler way to find your files*. Retrieved from

<https://CRAN.R-project.org/package=here>

Müller, K., & Wickham, H. (2021). *Tibble: Simple data frames*. Retrieved from

<https://CRAN.R-project.org/package=tibble>

R Core Team. (2020). *R: A language and environment for statistical computing*.

Vienna, Austria: R Foundation for Statistical Computing. Retrieved from

313 <https://www.R-project.org/>

314 Urbanek, S. (2013). *Png: Read and write PNG images*. Retrieved from

315 <https://CRAN.R-project.org/package=png>

316 Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of*
317 *Statistical Software*, 40(1), 1–29. Retrieved from

318 <http://www.jstatsoft.org/v40/i01/>

319 Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag

320 New York. Retrieved from <https://ggplot2.tidyverse.org>

321 Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string*

322 *operations*. Retrieved from <https://CRAN.R-project.org/package=stringr>

323 Wickham, H. (2021a). *Forcats: Tools for working with categorical variables (factors)*.

324 Retrieved from <https://CRAN.R-project.org/package=forcats>

325 Wickham, H. (2021b). *Tidyr: Tidy messy data*. Retrieved from

326 <https://CRAN.R-project.org/package=tidyr>

327 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., . . .

328 Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*,

329 4(43), 1686. <https://doi.org/10.21105/joss.01686>

330 Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of*

331 *data manipulation*. Retrieved from <https://CRAN.R-project.org/package=dplyr>

332 Wickham, H., & Hester, J. (2020). *Readr: Read rectangular text data*. Retrieved from

333 <https://CRAN.R-project.org/package=readr>

334 Winston Chang. (2014). *Extrafont: Tools for using fonts*. Retrieved from

335

<https://CRAN.R-project.org/package=extrafont>