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

Peekbank team[1], Martin Zettersten[2], & Michael C. Frank[1]

[1] Stanford University

[2] Princeton University

Author Note

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

Enter author note here.

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

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

Abstract

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

Keywords: keywords

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). A key part of the word learning process is children's ability to rapidly process words and link them to relevant meanings – often referred to as word recognition. Developing word recognition skills builds a foundation for children's language development and is predictive of later linguistic and general cognitive outcomes (Bleses, Makransky, Dale, Højen, & Ari, 2016; Marchman et al., 2018).

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. Studies using this design have contributed to our understanding of a wide range of questions in language development, including 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. Age-related changes and individual differences in speed of word recognition are thought to support children's subsequent language learning and predict later cognitive outcomes (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.

The Peekbank data framework consists of three components: (1) processing raw experimental datasets; (2) populating a relational database; and (3) providing an interface to the database (Fig XX). 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.

Data Format and Processing.

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. We developed a common, tidy format for the eye-tracking data in Peekbank to ease the process of conducting cross-dataset analyses (Wickham et al., 2019). The schema of the database is sufficiently general to handle heterogeneous datasets, including both manually coded and automated eye-tracking data.

During data import, raw eye-tracking datasets are processed to conform to the Peekbank data schema. The centerpiece of the schema is the aoi_timepoints table (Fig XX), which records whether participants looked to the target or the distractor stimulus at each timepoint of a given trial. Additional tables track information about data sources, participant characteristics, trial characteristics, stimuli, and raw eye-tracking data. In addition to unifying the data format, we conduct several additional pre-processing steps to facilitate analyses across datasets, including resampling observations to a common sampling rate (40 Hz) and normalizing time relative to the onset of the target label.

Current Data Sources.

The database currently includes 11 looking-while-listening datasets comprising N=1233 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

Peekbankr

Functions: connect_to_peekbank() get_datasets() get_subjects() get_administrations() get_stimuli() get_aoi_timepoints() get_trials() get_trial_types() get_xy_timepoints() get_aoi_region_sets()

OSF site

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

Peekbank in Action

General properties.

In general, participants demonstrated robust, above-chance word recognition in each dataset (with chance being 0.5). Table 2 shows the average proportion of target looking within a standard critical window of 300-2000ms after the onset of the label for each dataset (Swingley & Aslin, 2000). The number of unique target labels and their associated accuracy vary widely across datasets (Figure). Proportion target looking was generally higher for familiar words (M = 67.5%, 95% CI = [66.6%, 68.5%]) than for novel words learned during the experiment (M = 55.1%, 95% CI = [53.8%, 56.3%]).

Using peekbank to inform item selection

[something showing accuracy for specific items at different ages, kinda like an analog to AOA curves in Wordbank?]

Using peekbank to inform time window selection

In our second analysis, we address a common analytic decision facing researchers: how to summarize time course data into a single measure of accuracy. Taking a similar approach to that of Peelle & Van Engen (2020), we conducted a multiverse-style analysis considering possible time windows researchers might select (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). Our multiverse analysis focuses on the reliability of participants' response to familiar words by measuring the subject-level inter-item correlation (IIC) for proportion of looking at familiar targets. The time windows selected by researchers varies substantially in the literature, with some studies analyzing shorter time windows between 300ms and 1800-2000ms post-target onset (Fernald et al., 2008; Swingley & Aslin, 2000), and others using longer time windows extending to approximately 3000-4000ms (especially with younger infants; e.g., Bergelson & Swingley, 2012). We thus examined a broad range of window start times ranging from 300ms pre-target onset to 1500ms post-target onset and window end times ranging from 0ms to 4000ms post-target onset. For each combination of window start time and end time with a minimum window duration of 50ms, we calculated participants' average inter-item correlation for proportion of looking at familiar targets (mean IIC). Since observations were unevenly distributed across the age range, and because children likely show a varying response to familiar items as they age (often motivating different window choices), we split our data into four age bins (12-24, 24-36, 36-48, and 48-60 months). While it is an open question what space of possible windows will yield the greatest reliability, we expect to see low reliability (i.e. 0) in windows that start before target onset and in windows that end within 300ms post-target onset, before participants can execute a response.

Results from this multiverse analysis are shown in Figure , where each colored pixel represents the mean IIC for proportion of looking to familiar targets for a specific combination of window start and end time. The analysis shows that IIC is positive (red) under a wide range of sensible window choices. IIC is relatively low however, especially for the youngest age group, suggesting that individual items carry only limited shared signal regarding children's underlying ability. It may be the case that even averaging many such trials does not yield highly reliable measures of individual differences, although some multi-trial paradigms are exceptions to this generalization (Fernald et al., 2008).

Intriguingly, however, late end times and long overall window lengths show the greatest reliability. Shorter windows (e.g., 300-2000ms, as we used above) likely maximize absolute recognition performance by fitting the peak of the recognition curve, but simultaneously lower reliability by failing to include all relevant data. Especially for older children, the maximal IICs were found with windows that started between 500 and 1000ms and ended between 3000 and 4000ms, windows usually reserved for younger children. This finding is sensible from a psychometric perspective – averaging more timepoints (even if some contain limited signal) increases reliability and reduces variation. Thus, researchers interested in better measurement of individual variation or condition differences should consider using longer windows by default.

Reaction-time illustrations (?) Other methodological/ measure ideas? Compare first half/ second half reliability? /split-half reliability Meta-analytic estimate/ Comparison to Metalab Maybe inspect meta-analytic effect size at different ages/ for different moderators E.g. novel vs. familiar items. Integrating Peekbank + Wordbank Revisit AOA analysis? Integrating Peekbank + Childes-db Revisit frequency analysis? Teaching examples (as in peekbank) Could be useful training or teaching tool

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 4.0.2; R Core Team, 2020) and the R-packages *dplyr* (Version 1.0.2; Wickham et al., 2020), *forcats* (Version 0.5.0; Wickham, 2020a), *ggplot2* (Version 3.3.2; Wickham, 2016), *here* (Version 0.1; Müller, 2017), *papaja* (Version 0.1.0.9997; Aust & Barth, 2020), *png* (Version 0.1.7; Urbanek, 2013), *purrr* (Version 0.3.4; Henry & Wickham, 2020), *readr* (Version 1.3.1; Wickham, Hester, & Francois, 2018), *stringr* (Version 1.4.0; Wickham, 2019), *tibble* (Version 3.0.4; Müller & Wickham, 2020), *tidyr* (Version 1.1.2; Wickham, 2020b), and *tidyverse* (Version 1.3.0; Wickham, Averick, et al., 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

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools*. Retrieved from https://CRAN.R-project.org/package=purrr

Müller, K. (2017). *Here: A simpler way to find your files*. Retrieved from https://CRAN.R-project.org/package=here

Müller, K., & Wickham, H. (2020). *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 https://www.R-project.org/

Urbanek, S. (2013). *Png: Read and write png images*. Retrieved from https://CRAN.R-project.org/package=png

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H. (2020a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2020b). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., Hester, J., & Francois, R. (2018). *Readr: Read rectangular text data*. Retrieved from https://CRAN.R-project.org/package=readr

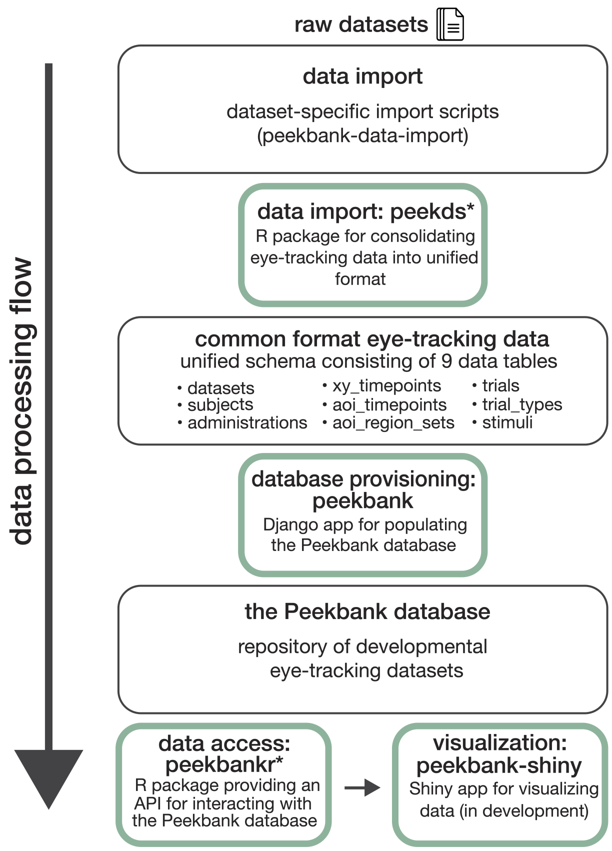

Figure 1. Overview of the Peekbank data ecosystem. Peekbank tools are highlighted in green. *custom R packages.

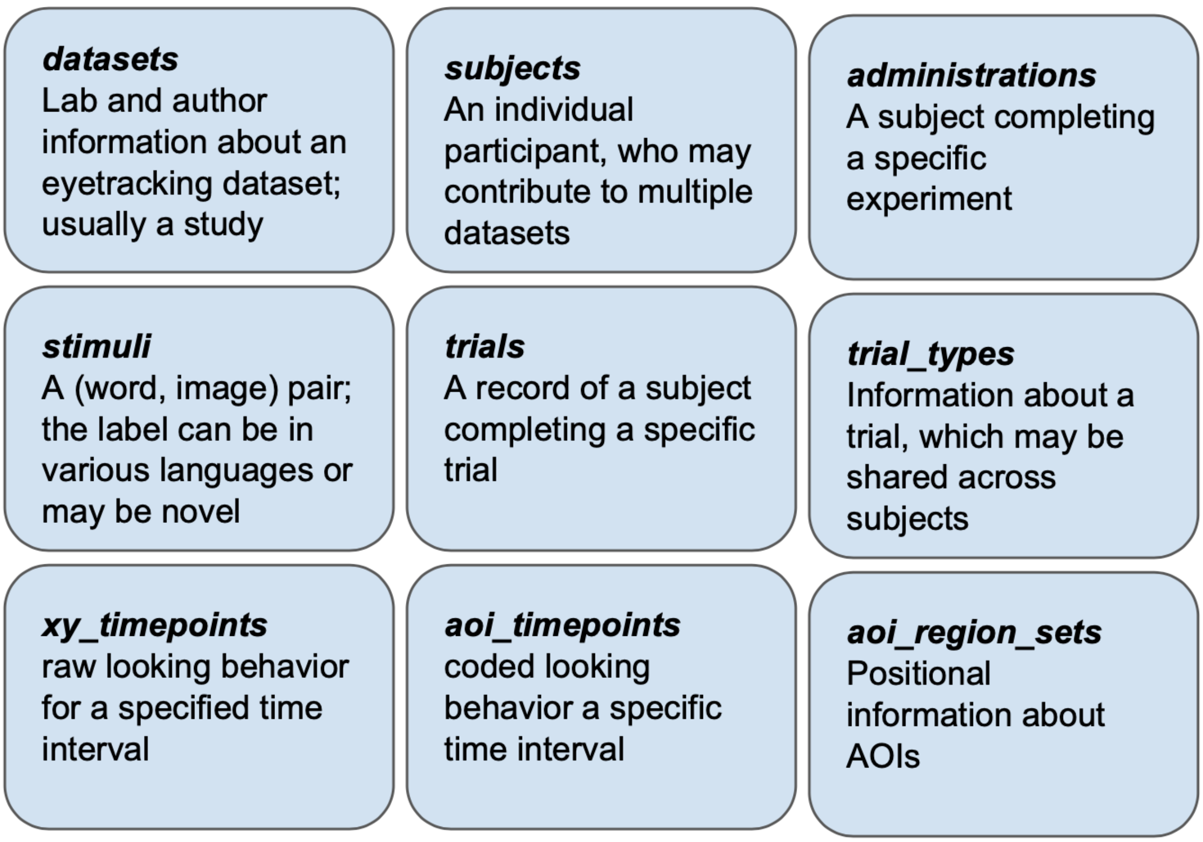

Figure 2. The Peekbank schema. Each square represents a table in the relational database.