

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

Martin Zettersten (martincz@princeton.edu)¹, Claire Bergey², Naiti S. Bhatt³, Veronica Boyce⁴, Mika Braginsky⁵, Alexandra Carstensen⁴, Benny deMayo¹, George Kachergis⁴, Molly Lewis⁶, Bria Long⁴, Kyle MacDonald⁷, Jessica Mankewitz⁴, Stephan Meylan^{5,8}, Annissa N. Saleh⁹, Rose M. Schneider¹⁰, Angeline Sin Mei Tsui⁴, Sarp Uner⁸, Tian Linger Xu¹¹, Daniel Yurovsky⁶, Michael C. Frank (mcfrank@stanford.edu)⁴

¹Dept. of Psychology, Princeton University, ²Dept. of Psychology, University of Chicago, ³Scripps College, ⁴Dept. of Psychology, Stanford University, ⁵Dept. of Brain and Cognitive Sciences, MIT, ⁶Dept. of Psychology, Carnegie Mellon University, ⁷Core Technology, McD Tech Labs, ⁸Dept. of Psychology and Neuroscience, Duke University, ⁹Dept. of Psychology, UT Austin
¹⁰Dept. of Psychology, UC San Diego, ¹¹Dept. of Psychological and Brain Sciences, Indiana University

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. We present two analyses of the current database ($N=1,320$): (1) capturing age-related changes in infants' word recognition while generalizing across item-level variability and (2) assessing how a central methodological decision – selecting the time window of analysis – impacts the reliability of measurement. Future efforts will expand the scope of the current database to advance our understanding of participant-level and item-level variation in children's vocabulary development.

Keywords: word recognition; eye-tracking; vocabulary development; looking-while-listening

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 facet of children's growing word knowledge is their ability to rapidly process words and link them to relevant meanings – often termed word recognition. Developing word recognition skills builds a foundation for language development; for example, the speed and accuracy with which infants process words are predictive of later linguistic and cognitive outcomes (e.g., Marchman et al., 2018).

Word recognition is traditionally studied in the “looking-while-listening” paradigm (or alternatively in a version of the intermodal preferential looking paradigm; 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).

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 how to improve measurement of children's word recognition in the first place. One central question is how to measure developmental change in the speed and accuracy of word recognition. In language development research, processing speed - the ability to quickly link a word to its referent - has been of key interest because it is thought to both reflect past language learning and to support subsequent learning. Age-related changes in speed of processing are argued to accelerate infants' learning: as infants begin to process incoming speech input faster, they become better equipped to learn from their language environment (Fernald & Marchman, 2012). Consistent with this hypothesis, 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 research, 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). One way to overcome this challenge is to measure word recognition across development in a large-scale dataset with a wide range of items. A sufficiently large dataset would allow researchers to estimate developmental change in word recognition speed and accuracy while generalizing across changes related to specific words.

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.

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* (<https://peekbank.stanford.edu/>), 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 tools for standardizing eye-tracking data across data sources, and (c) provides an interface for accessing and analyzing the database. In creating the Peekbank database, the project pursues two main aims: to answer central theoretical questions about lexical development and to provide data-driven guidance on methodological decisions. In the current paper, we give an overview of the key components of the project and demonstrate its utility in advancing theoretical and methodological insights. We report two analyses using the database and associated tools ($N=1,320$): (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.

Methods

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. 1). 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

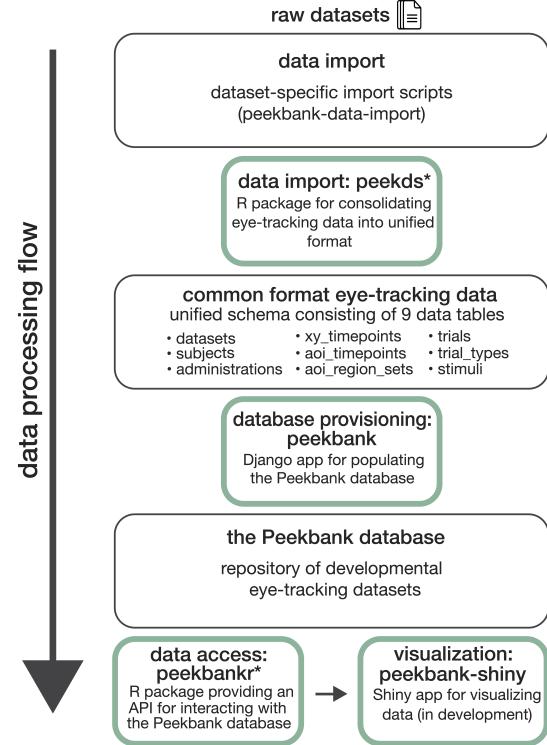


Figure 1: Overview of the Peekbank data ecosystem. Peekbank tools are highlighted in green. *custom R packages.

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. 1), 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, participants, trials, stimuli, and raw eye-tracking data. In addition to unifying the data format, we conduct several additional preprocessing steps to facilitate analyses across datasets. First, we normalize time relative to the onset of the target label, since the main goal is to assess participants’ fixation of the target image in response to hearing its corresponding label. Second, we resample observations to a common sampling rate (40 Hz), in order to ease data visualization and analysis. Where necessary (e.g., if the original data was sampled at 30 Hz), observations are interpolated by selecting the look location at the nearest time point in the original data.

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 15 looking-while-listening datasets comprising $N=1320$ total participants (Table 1; Pomper & Saffran, 2016, 2019; Adams et al., 2018; Byers-Heinlein, Morin-Lessard, & Lew-Williams, 2017; Casillas, Brown, & Levinson, 2017; Frank, Braginsky, Yurovsky, & Marchman, 2017; Garrison, Baudet, Breitfeld, Aberman, & Bergelson, 2020; Mahr, McMillan, Saffran, Ellis Weismer, & Edwards, 2015; Perry & Saffran, 2017; Potter, Fourakis, Morin-Lessard, Byers-Heinlein, & Lew-Williams, 2019; Swingley & Aslin, 2002; Yurovsky & Frank, 2017; Yurovsky, Wade, & Frank, 2013). Most datasets (12 out of 15 total) consist of data from monolingual native English speakers. They span a wide age spectrum with participants ranging from 9 to 70 months of age, and are balanced in terms of gender (46% female). The datasets vary across a number of design-related dimensions, and include studies using manually coded video recordings and automated eye-tracking methods (e.g., Tobii, EyeLink) to measure gaze behavior. All studies tested familiar items, but the database also includes 5 datasets that tested novel pseudo-words in addition to familiar words. All data are openly available on the Open Science Framework (<https://osf.io/pr6wu/>).

Results

General descriptives and Item Variability

Analysis scripts for all results are openly available at <https://github.com/langcog/peekbank-paper>. In general, participants demonstrated robust, above-chance word recognition in each dataset (chance=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 2). Proportion target looking was generally higher for familiar words ($M = 0.66$, 95% CI = [0.65, 0.67], $n = 1269$) than for novel words learned during the experiment ($M = 0.59$, 95% CI = [0.58, 0.61], $n = 822$).

Dataset Name	Unique Items	Prop. Target	95% CI
attword	6	0.63	[0.61, 0.64]
canine	16	0.64	[0.61, 0.67]
coartic	10	0.70	[0.67, 0.73]
cowpig	12	0.60	[0.58, 0.63]
ft.pt	8	0.64	[0.63, 0.66]
mispron	22	0.57	[0.55, 0.59]
mix	6	0.55	[0.52, 0.58]
reflook_socword	6	0.61	[0.6, 0.63]
reflook_v4	10	0.61	[0.57, 0.65]
remix	8	0.62	[0.58, 0.66]
salientme	16	0.73	[0.71, 0.75]
switchingCues	40	0.77	[0.75, 0.79]
tablet	24	0.63	[0.6, 0.67]
tseltal	30	0.59	[0.54, 0.63]
yoursmy	87	0.60	[0.56, 0.64]

Table 2: Average proportion target looking in each dataset.

Predicting Age-Related Changes While Generalizing Across Items

Developmental changes in word recognition have been a central issue since early investigations of eye-tracking techniques (Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998). Children’s speed and accuracy of word recognition increases across early childhood, yet measuring these increases presents an item selection puzzle for researchers: Words that are appropriate for an 18-month-old will be too easy for a three-year-old; those that are appropriate for a three-year-old will be difficult for the 18-month-old. Failure to choose appropriate test items can even lead to spurious conclusions about development (Peter et al., 2019).

This issue is familiar in psychometrics: test developers interested in measuring across a wide range of a particular latent ability must choose items appropriate for different abilities. One solution is to use data from a bank of questions that have been taken by test-takers of a range of abilities, and then use item-response theory models to create different test versions appropriate for different ability ranges (Embretson & Reise, 2000). Such tests can then be used to extract estimates of developmental change that are independent of individual

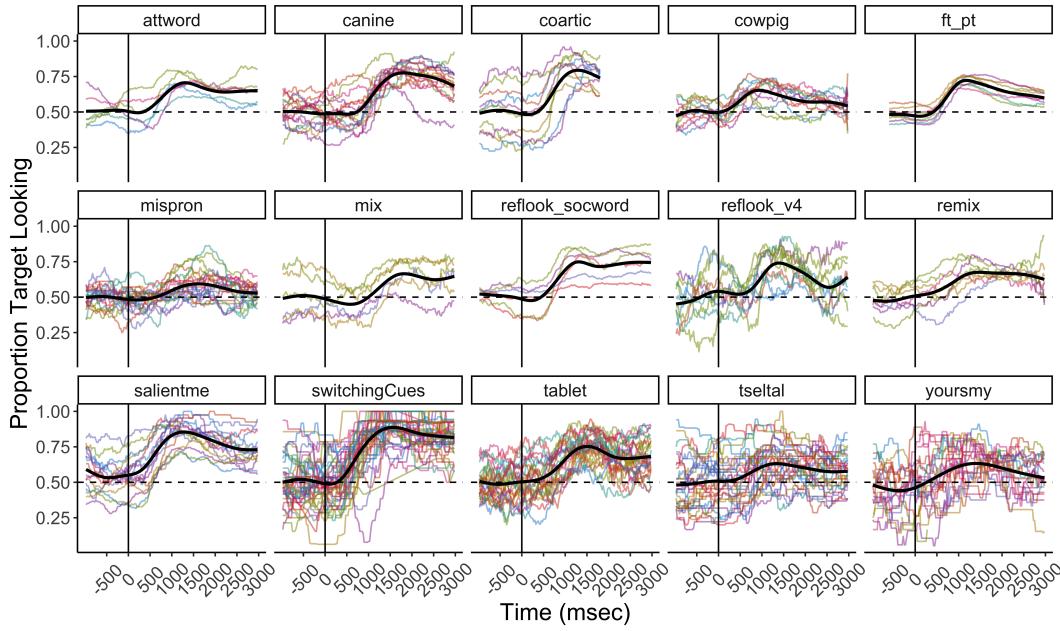


Figure 2: Item-level variability in proportion target looking within each dataset (chance=0.5). Time is centered on the onset of the target label (vertical line). Colored lines represent specific target labels. Black lines represent smoothed average fits based on a general additive model using cubic splines. See Table 1 for details on each dataset.

tests and their particular items.

Peekbank provides the appropriate large-scale database for estimating these item-independent developmental changes and designing age-appropriate tests in the future. Here we show a proof of concept by providing an estimate of the item-independent growth of word recognition accuracy across development. We take advantage of the equivalence between item response theory and linear mixed-effects models (LMMs; De Boeck et al., 2011), using LMMs to model the trajectory of word recognition across age. We follow the approach of Mirman (2014) and use growth curve LMMs to predict the time course of recognition. Specifically, we predicted children’s proportion of target looking during an early window of time (0–1500ms, chosen to avoid modeling declines in looking later in the trial), using an empirical logit transform on the proportion of target looking to allow the use of linear (rather than logistic) regression models. We focused our analysis on familiar words and included only datasets testing infants with monolingual English language exposure (12 datasets, see Table 1). Our predictors were time after word onset and age, and we additionally included polynomial functions of time (up to fourth order) and quadratic effects of age, as well as their interactions. We subtracted all intercepts to force fits to start at a baseline of 0 (chance performance) at time zero. As a random effect structure, we included by-item, by-subject, and by-dataset random intercepts; though larger random effect structures could be justified by the data, the size of the dataset precluded fitting these. The random effect structure allows us to model word recognition while generalizing across items and participants.

Figure 3 depicts the results of this analysis. Panel A shows the mean empirical word recognition curves for four age groups, along with fitted model performance. Although model fits are acceptable, developmental change appears ir-

regular – for example, 12- to 24-month-olds show slightly earlier recognition than 24- to 36-month-olds. This pattern is likely an artifact of averaging across datasets with substantially different items. Panel B shows model predictions for the population level of each random effect – our best estimates of the latent ability structure. Here we see continuous increases in both speed (point at which the curve rises) and accuracy (asymptote of the curve) across ages, though this developmental trend decelerates (consistent with other work on reaction time development; Frank, Lewis, & MacDonald, 2016; Kail, 1991). This proof of concept suggests that Peekbank can be used to model developmental change over multiple years, overcoming the limitations of individual datasets.

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

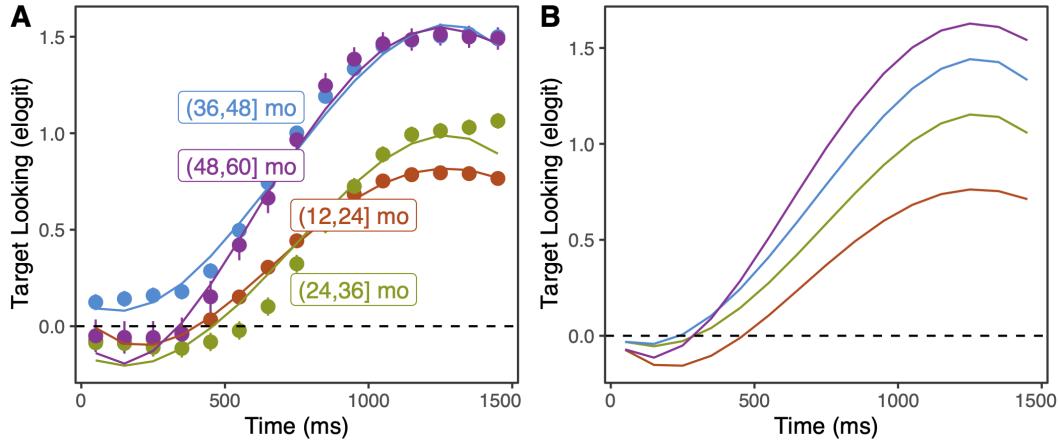


Figure 3: Growth curve models of proportion target looking during the critical target window at each age range (in months). (A) Mean empirical word recognition fit. (B) Population-level estimates.

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) within each dataset and subsequently averaged inter-item correlations across datasets. 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). We restricted the analysis within a given age bin to include only datasets that contributed at least 5 unique test sessions. 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 4, 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 tend to show higher reliability than shorter windows. 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 may lower reliability by failing to include all relevant data. Especially for older children, the maximal IICs tended to be found with windows that started between 300 and 1000ms and ended between 3000 and 4000ms. This finding is sensible from a psychometric perspective – averaging more time-

points (even if some contain limited signal) increases reliability and reduces variation. Thus, researchers interested in better measurement of individual variation or condition differences could consider using longer windows by default.

Discussion

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 et al., 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. The goals of the database are to ask theory-driven questions about the development of word knowledge and to establish methodological best-practices in infant eye-tracking methods. 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 choosing windows of analysis.

Our first analysis shows that Peekbank can be used to model item-independent changes in the speed and accuracy of word recognition across development. Children showed age-related increases in the speed of word recognition across one to five years of age, extending past foundational work (e.g., Fernald et al., 1998) by showing that these word processing gains generalize across items and are not only attributable to word-specific gains in processing speed. The second analysis demonstrates how Peekbank can be used to make data-driven analytic decisions, focusing on the choice of time windows for analysis. In looking-while-listening studies, researchers often choose a relatively short time window of roughly 300-1800 or 2000ms (Fernald et al., 2008), with the justification that eye movements occurring after this window may no longer be related to the target label (Swingley & Aslin, 2000). Our results suggest that researchers could consider increasing

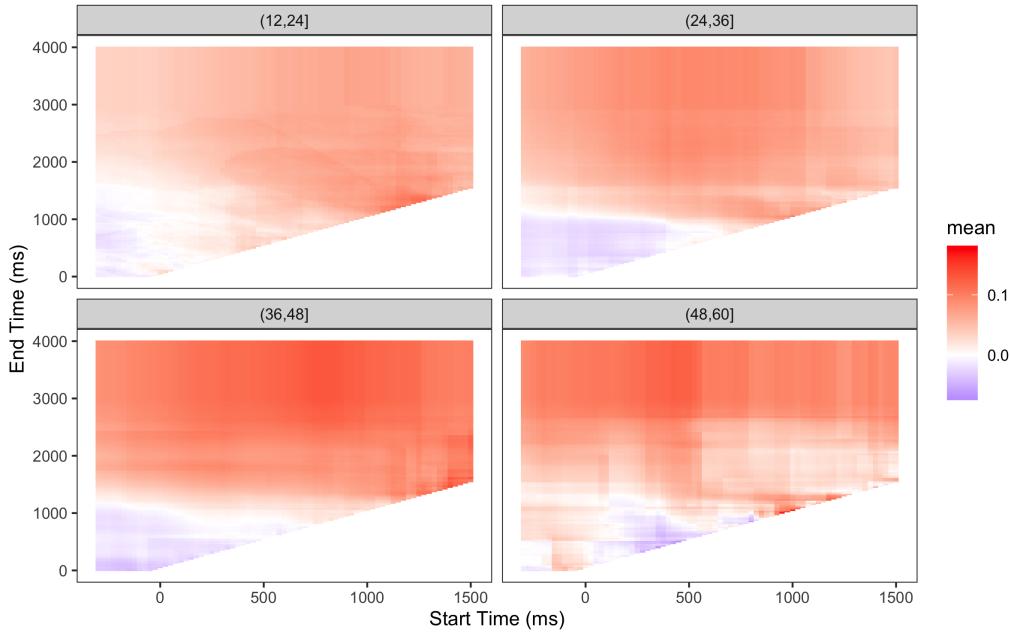


Figure 4: Participants' average inter-item correlation for proportion of looking time to familiar targets, as a function of window start time and end time, with each facet showing a different age group. More positive (red) correlations are more desirable, and blue/white represent start/end time combinations that researchers should avoid.

the size of the time window for analyzing target fixations (at least for familiar words) if their goal is to maximize consistent signal in children's target fixations.

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 15 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 lead 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 (mostly 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.

References

- Adams, K. A., Marchman, V. A., Loi, E. C., Ashland, M. D., Fernald, A., & Feldman, H. M. (2018). Caregiver talk and medical risk as predictors of language outcomes in full term and preterm toddlers. *Child Development*, 89(5), 1674–1690.
- Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *PNAS*, 109(9), 3253–3258.
- Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development*, 89(6), 1996–2009.
- Byers-Heinlein, K., Morin-Lessard, E., & Lew-Williams, C. (2017). Bilingual infants control their languages as they listen. *PNAS*, 201703220.
- Casillas, M., Brown, P., & Levinson, S. (2017). Casillas HomeBank Corpus. <http://doi.org/10.21415/T51X12>
- Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Lengyel, M. (2016). Statistical treatment of looking-time data. *Developmental Psychology*, 52(4), 521–536.
- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in r. *Journal of Statistical Software*, 39(12), 1–28.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Fernald, A., & Marchman, V. A. (2012). Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child Development*, 83(1), 203–22.
- Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*, 9(3), 228–231.

- Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In I. A. Sekerina, E. M. Fernandez, & H. Clahsen (Eds.), *Developmental psycholinguistics: On-line methods in children's language processing* (pp. 97–135). Amsterdam: John Benjamins.
- Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, 44(3), 677–694.
- Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and Consistency in Early Language Learning: The Wordbank Project*. Cambridge, MA: MIT Press.
- Frank, M. C., Lewis, M., & MacDonald, K. (2016). A performance model for early word learning. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2610–2614). Austin, TX: Cognitive Science Society.
- Garrison, H., Baudet, G., Breitfeld, E., Aberman, A., & Bergelson, E. (2020). Familiarity plays a small role in noun comprehension at 12–18 months. *Infancy*, 25(4), 458–477.
- Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: What have we learned? *Perspectives on Psychol. Science*, 8(3), 316–339.
- Hirsh-Pasek, K., Cauley, K. M., Golinkoff, R. M., & Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of Child Language*, 14(1), 23–45.
- Huang, Y., & Snedeker, J. (2020). Evidence from the visual world paradigm raises questions about unaccusativity and growth curve analyses. *Cognition*, 200, 104251.
- Kail, R. (1991). Developmental change in speed of processing during childhood and adolescence. *Psychological Bulletin*, 109(3), 490.
- Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, 18(3), 193–198.
- Mahr, T., McMillan, B. T. M., Saffran, J. R., Ellis Weismer, S., & Edwards, J. (2015). Anticipatory coarticulation facilitates word recognition in toddlers. *Cognition*, 142, 345–350.
- Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science*, 11(3), F9–F16.
- Marchman, V. A., Loi, E. C., Adams, K. A., Ashland, M., Fernald, A., & Feldman, H. M. (2018). Speed of language comprehension at 18 months old predicts school-relevant outcomes at 54 months old in children born preterm. *Journal of Dev. & Behav. Pediatrics*, 39(3), 246–253.
- Mirman, D. (2014). *Growth curve analysis and visualization using R*. Boca Raton, FL: CRC Press.
- Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, educated, industrial, rich, and democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science*, 31(6), 678–701.
- Peelle, J. E., & Van Engen, K. J. (2020). Time stand still: Effects of temporal window selection on eye tracking analysis. *PsyArXiv*. <http://doi.org/https://doi.org/10.31234/osf.io/pc3da>
- Perry, L. K., & Saffran, J. R. (2017). Is a pink cow still a cow? Individual differences in toddlers' vocabulary knowledge and lexical representations. *Cognitive Science*, 41(4), 1090–1105. <http://doi.org/10.1111/cogs.12370>
- Peter, M. S., Durrant, S., Jessop, A., Bidgood, A., Pine, J. M., & Rowland, C. F. (2019). Does speed of processing or vocabulary size predict later language growth in toddlers? *Cognitive Psychology*, 115, 101238.
- Pomper, R., & Saffran, J. R. (2016). Roses are red, socks are blue: Switching dimensions disrupts young children's language comprehension. *Plos One*, 11(6), e0158459.
- Pomper, R., & Saffran, J. R. (2019). Familiar object salience affects novel word learning. *Child Development*, 90(2), e246–e262.
- Potter, C. E., Fourakis, E., Morin-Lessard, E., Byers-Heinlein, K., & Lew-Williams, C. (2019). Bilingual toddlers' comprehension of mixed sentences is asymmetrical across their two languages. *Developmental Science*, 22(4), 1–9.
- R Development Core Team. (2020). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
- Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychol. Science*, 11(5), 702–712.
- Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, 76(2), 147–66.
- Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science*, 13(5), 480–484.
- The ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed speech preference. *Advances in Methods and Practices in Psychological Science*, 3(1), 24–52.
- 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.
- Yurovsky, D., & Frank, M. C. (2017). Beyond naïve cue combination: Salience and social cues in early word learning. *Developmental Science*, 20(2), 1–17.
- Yurovsky, D., Wade, A., & Frank, M. C. (2013). Online processing of speech and social information in early word learning., 1641–1646.