Peekbank: Exploring child lexical processing through a large-scale open-source database of developmental eye-tracking datasets

Anonymous CogSci submission

Abstract

Lexical processing skills – the ability to rapidly process words and link them to referents in context – are central to children's early language development. Children's lexical processing is typically studied in the looking-while-listening paradigm, which measures infants' fixation of a target object (vs. a distracter) after hearing a target label. We present a large-scale open-source database of infant and toddler eye-tracking data from looking-while-listening tasks. The goal of this database is to address theoretical and methodological challenges in measuring infant vocabulary development. We present two analyses of the current database (N=1,233): (1) models capturing age-related changes in infants' lexical processing while generalizing across item-level variability and (2) an analysis of how a central methodological decision – selecting the time window of analysis – impacts measure reliability. 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: lexical processing; eye-tracking; database; 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 part of the word learning process is children's ability to rapidly process words and link them to relevant meanings - often referred to as lexical processing. Developing lexical processing 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).

Lexical processing 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' lexical processing 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 (Golinkoff, Ma, Song, & Hirsh-Pasek, 2013), including infants' early noun knowledge (Bergelson & Swingley, 2012), phonological representations of words (Swingley & Aslin, 2000), prediction during language processing (Lew-Williams & Fernald, 2007), and individual differences in language development (Marchman et al., 2018).

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 lexical processing ability and the nature of the method itself. One central question relates to teasing apart variability due to participants and variability due to specific items across development. Drawing inferences about item-level variability is key to many questions in how word learning unfolds, such as how properties of the language input influence lexical development (Goodman, Dale, & Li, 2008; Roy, Frank, Decamp, Miller, & Roy, 2015). Most studies of infant lexical processing focus on generalizing performance across participants, and are constrained in their ability to provide generalizations across the item level - the level of specific words. Generalizing behavior on the level of both participants and items simultaneously is difficult in the context of a solitary study, especially given practical constraints on the number of trials (and consequently items) tested within a given infant. One key to meeting this challenge is having sufficiently large datasets to account for and explain variability in lexical processing on the item level.

A second question relates to evaluating methodological best-practices. In particular, many fundamental analytic decisions vary substantially across studies. For example, researchers vary in their decisions regarding how to select time windows for analysis, modeling how lexical processing unfolds over time, and the appropriate transformations to perform on the dependent measure of target fixations (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 likely 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 lexical processing studies varying across such factors, in order to independently test the potential consequences of a variety of methodological decisions on the interpretation of study results.

What these questions share is that they are difficult to answer at the scale of a single looking-while-listening study. In order to address this challenge, we introduce *Peekbank*, a flexible and reproducible interface to an open database of

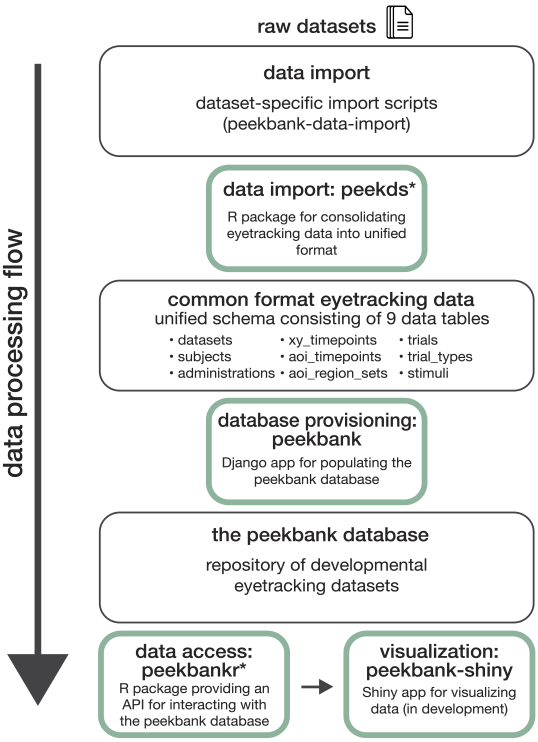

Figure 1: Overview of the peekbank data ecosystem.

developmental eye-tracking studies. Here, we give a brief overview of the key components of the project and some initial demonstrations of its utility in advancing theoretical and methodological questions in the study of children's lexical processing. The *Peekbank* project (a) collects a large set of eye-tracking datasets on children's lexical processing, (b) introduces a data format and processing tools for standardizing eye-tracking data across data sources, and (c) provides an API for accessing and analyzing the database. We report two analyses using the database and associated tools (N=1,233): (1) a growth-curve analysis modeling age-related changes in infants' lexical processing 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 measure reliability.

Methods

Database Framework

The *Peekbank* data framework consists of three libraries that help to populate and query a relational database (Fig. 1). The `peekds` library (for the R language) helps researchers convert and validate existing datasets to use the relational format used by the database. The `peekbank` library (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 provides high-level abstractions to help researchers run common analysis tasks on the database.

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 among labs conducting individual experiments. Researcher conventions for exporting and structuring data vary, 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. 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 distracter 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 information. 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

Dataset Name	N	Mean Age	Method
canine	36	23.8	manual coding
coartic	29	20.8	eyetracking
cowpig	45	20.5	manual coding
ft_pt	69	17.1	manual coding
reflook_socword	435	33.6	eyetracking
reflook_v4	347	37.2	eyetracking
salientme	44	40.1	manual coding
switchingCues	60	44.3	manual coding
tablet	110	33.8	eyetracking
tseltal	23	31.3	manual coding
yoursmy	35	14.5	eyetracking

Table 1: Overview over the datasets in the current database.

The database currently includes 11 datasets comprising N=1233 total participants, with 23 to 435 participants per dataset (Table 1). The vast majority of datasets (10 out of 11 total) consist of monolingual native English speakers. The datasets 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 studies in the current database vary across a number of dimensions related to design and methodology. The database includes studies using both manually coded video recordings or aumotated eye-tracking methods (e.g., Tobii, EyeLink) to measure children's gaze behavior. Most studies focused on testing familiar items, but the database also includes studies with novel pseudowords.

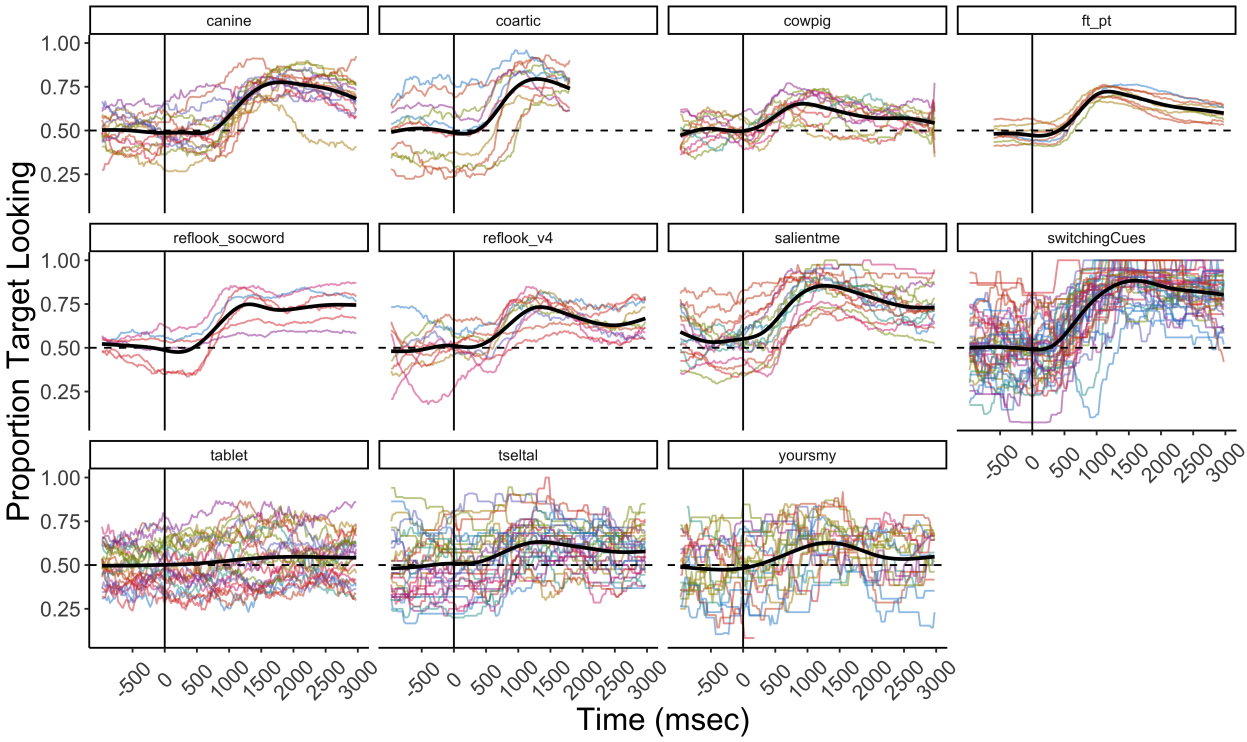

Figure 2: Item-level variability in proportion target looking within each dataset. Colored lines represent specific target labels. Black lines represent smoothed average fits based on a general additive model using cubic splines.

Results

General descriptives and item variability

Dataset Name	Unique Items	Prop. Target	95% CI
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]
reflook_socword	6	0.61	[0.6, 0.63]
reflook_v4	10	0.63	[0.61, 0.64]
salientme	16	0.73	[0.71, 0.75]
switchingCues	40	0.77	[0.75, 0.79]
tablet	24	0.53	[0.51, 0.56]
tseltal	30	0.59	[0.54, 0.63]
yoursmy	87	0.60	[0.55, 0.64]

Table 2: Average proportion target looking in each dataset.

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 target labels (M = 67.5%, 95% CI = [66.6%, 68.5%]) than for novel target labels learned during the experiment (M = 55.1%, 95% CI = [53.8%, 56.3%]).

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 creates a tricky issue. 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 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 tests and their particular items.

Peekbank provides the appropriate data 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.

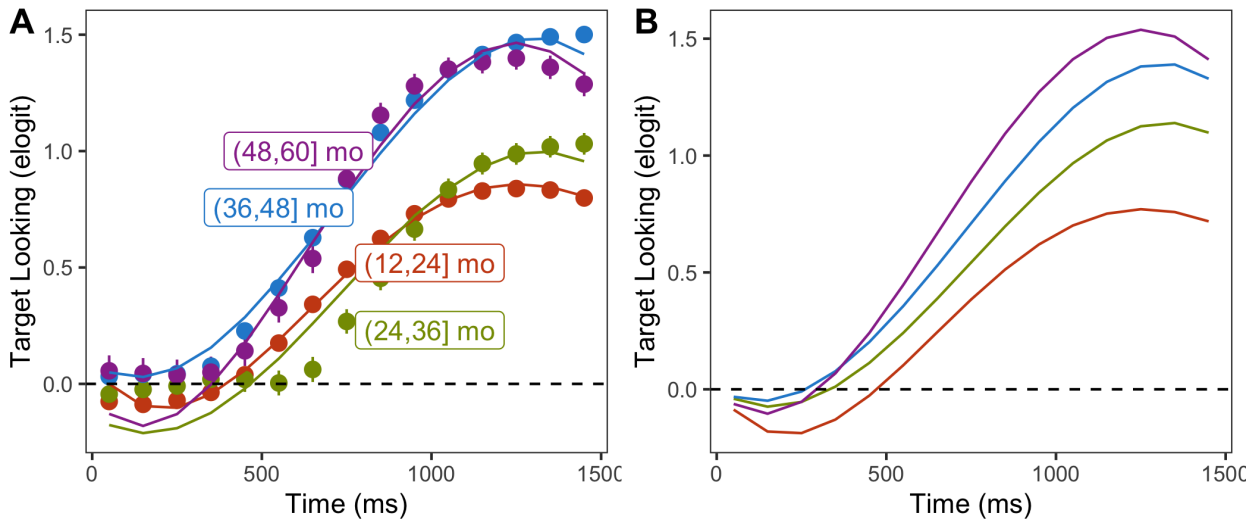

Figure 3: Growth curve models of proportion target looking during the critical target window at each age range (in months).

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 timecourse of recognition. Specifically, we predicted children's proportion of target looking during an early window of time (0 - 1500ms), using an empirical logit transform on the proportion of target looking to allow the use of linear (rather than logistic) regression models. 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 a larger random effect structure could be justified by the data, the size of the dataset precluded fitting these.

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 irregular - for example, 12–24 month-olds show slightly earlier recognition than 24–36 month-olds. This pattern is an artifact of averaging across datasets with substantially different items and structures. 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

Taking a similar approach to that of Peelle & Van Engen (2020), we conducted a multiverse-style analysis considering possible time windows researchers might select for analyzing their data. 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 300 ms and 1800-2000 ms post-target onset (Fernald et al., 2008; Swingley & Aslin, 2000), while longer time windows extending to approximately 3000-4000ms post-target onset are often also used (e.g., Bergelson & Swingley, 2012). We thus examined a broad range of window start and end times: we calculated subjects' mean IIC across from 300 ms pre-target onset to 1500 ms post-target word onset) and window end times (ranging from 0 ms to 4000 ms). While it is an open question what space of possible windows will yield the greatest reliability, we expect to see very low reliability (i.e. 0) in windows that start before target onset, and likely for any windows that end within 300 ms post-target onset, before participants have had a chance to execute a response. Since observations were unevenly distributed across the age range, and because children likely show a varying response to familiar items as they age, we split our data into four age bins (12-24, 24-36, 36-48, and 48-60 months). For each combination of window start time and end time with a minimum window duration of 50 ms, participants' average inter-item correlation for proportion of looking at familiar targets was calculated.

The resulting correlations of this multiverse analysis are shown in Figure 4, where subjects' mean ICC for proportion of looking to familiar targets for each combination of window start time and end time is shown as a colored pixel. The analysis shows that ICC is positive (red) under a wide range

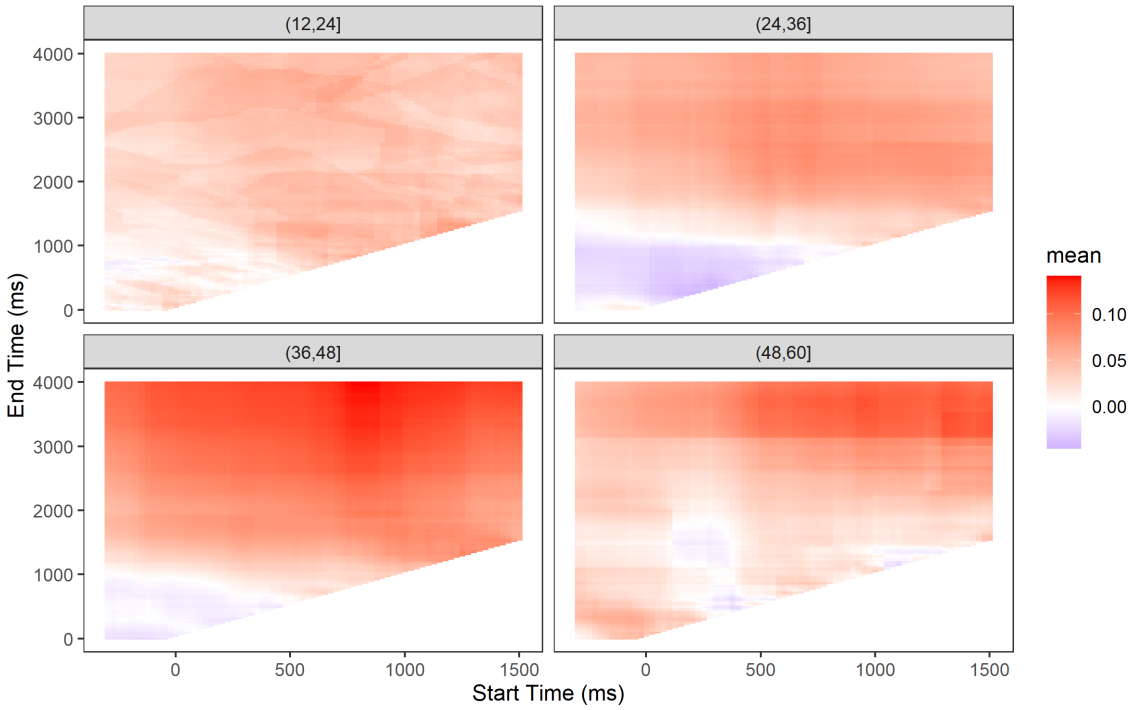

Figure 4: Participants' average inter-item correlation for proportion of looking time to familiar targets, as a function of window start time and end time, with each facet showing a different age group. More positive (red) correlations are more desirable, and blue/white represent start/end time combinations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

of window choices, and generally only negative (blue) or 0 (white) when both the start time is less than ˜500 ms and the end time is less than ˜1000 ms – a window choice researchers rarely consider for analyzing word recognition. Intriguingly, late end times and long overall window lengths–generally not used by researchers–show the greatest reliability, suggesting that researchers may consider keeping this data for analysis rather than discarding it. Moreover, there is some variation by age group in where the strongest ICCs are found (and in the overall strength of ICCs).

What general recommendations follow from this analysis? We minimally consider which start times and window lengths result in IICs of at least .01, noting that this is still a rather low target. A window length of at least 1500 ms eliminated 94% of low ICCs, and this threshold combined with a start time of at least 300 ms eliminated all but 0.5% of low ICCs. A start time of 500 ms and a window length of at least 1500 ms resulted in no IICs $< .01$, and an average IIC of 0.08. However, the overall strength of the IICs are generally weaker than might be desired (median = .05), with maximum values of 0.15 (reached only in 3-year-olds).

Discussion

Many central questions in developmental science face a fundamental data collection challenge: Studying effects of interest requires a large amount of observations, but collecting infant data is difficult, time-intensive, and often limited to a small number of observations per participant. Recent years have seen a growing effort to build open source tools and pooling research efforts to meet the challenge of building a cumulative developmental science (Bergmann et al., 2018; The ManyBabies Consortium, 2020). *Peekbank* expands on these efforts by building an infrastructure for aggregating eye-tracking data across studies, with a particular focus on the looking-while-listening paradigm. This paper presents a preliminary illustration of some of the key theoretical and methodological questions *Peekbank* aims to address: generalizing across item-level variability in children's lexical processing and providing data-driven guidance on methodological choices.

In our first analysis, we show that *Peekbank* can be used to model item-independent changes in the speed and accuracy of lexical processing across development. Children showed steady age-related increases in the speed of lexical processing across one to five years of age, replicating and extending past foundational work (e.g., Fernald et al., 1998), while demonstrating that these word processing gains generalize across items and are not only attributable word-specific gains in processing speed. The second analysis provides a demonstration of how *Peekbank* can be used to inform principled, data-driven analytic decisions, specifically the decision of how to choose time windows for analyzing eye-tracking data. In looking-while-listening studies, researchers often choose a relatively short time window of roughly 300-1800 or 2000

ms (Fernald et al., 2008), with the justification that eye movements occurring after this window may no longer reflect fixations related to the target label (Swingley & Aslin, 2000). Intriguingly, we find that longer time windows that extend as much as 4000 ms after the target label typically continue to increase the reliability of proportion target looking. Our results recommend that researchers consider increasing the size of the time window for analyzing target fixations (at least for familiar words) to maximize the consistent signal present in children's target fixations.

There are a number of critical limitations surrounding the current scope of the database. A key priority in future work will be to expand the size of the database. With 11 datasets currently available in the dataset, 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, allowing for more robust generalizations regarding participant- and 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 environments; Muthukrishna et al., 2020). Increasing the diversity of participant backgrounds and languages will increase the scope of the generalizations we can form about child lexical processing.

Finally, while the current database is mainly focused on studies of lexical processing, the tools and infrastructure developed in the project can in principle be expanded to accommodate any eye-tracking paradigm used with infants and toddlers, opening up new avenues for insights into infant development. Infant looking has been at the core of many of the key advances in our understanding 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

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.

Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early productive vocabulary predicts academic achievement 10 years later. *Applied Psycholinguistics*, *37*(6), 1461–1476.

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., ... others. (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 Associates.

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. (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.

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 Psychological Science*, *8*(3), 316–339.

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

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.

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. *Jour-*

nal of Developmental & Behavioral 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

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.

Roy, B. C., Frank, M. C., Decamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *PNAS*, *112*(41), 12663–12668.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*(2), 147–66.

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.