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

Anonymous CogSci submission

Abstract

Developing lexical processing skills – the ability to rapidly process words and link them to referents in context – is 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 (as opposed to a distracter) after hearing a target label. We present a large-scale open-source database of data from infant and toddler looking-while-listening studies. The goal of this database is to address theoretical and methodological challenges in measuring infant vocabulary development that go beyond the scope of the individual studies. We present three analyses of the current database (N=XYZ): (1) models capturing item-level variability in infants' lexical processing across age; (2) an analysis of how a central methodological decision – selecting the time window of analysis – impacts measure reliability; and (3) an analysis demonstrating the link between the age of acquisition of specific words and children's ability to rapidly and accurately link those words to their referents. 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; eyetracking; database; vocabulary development; looking-while-listening

Introduction

Across their first years of life, children learn words in their native tongues at a rapid pace (Braginsky, Yurovsky, Marchman, & Frank, 2019). A key part of the word learning process is children's ability to rapidly process words and link them to relevant meanings in context - often referred to as lexical processing. Developing lexical processing skills builds a foundation for children's language development and is predictive of both linguistic and more general cognitive outcomes later in life (Bleses, Makransky, Dale, Højen, & Ari, 2016; Marchman et al., 2018).

Lexical processing is traditionally studied in "looking-while-listening" studies (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 infants subsequently fixate the correct target image after hearing its label. Studies using this basic 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 nature of children's early word knowledge and the nature of the method itself. One central question relates to understanding word-specific variability across development, and generalizing lexical processing on the level of specific words. 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 often 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. However, drawing inferences about item-level variability is key to many questions in how word learning unfolds, including how properties of the language input influence lexical development (Goodman, Dale, & Li, 2008; Roy, Frank, Decamp, Miller, & Roy, 2015). One key to meeting this challenge is having sufficiently large datasets to interrogate 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, Zangl, Portillo, & Marchman, 2008; Huang & Snedeker, 2020). Establishing best practices regarding analytic decisions of this kind requires a large database of infant lexical processing studies, in order to independently test the potential consequences of a variety of methodological decisions on the interpretation of study results.

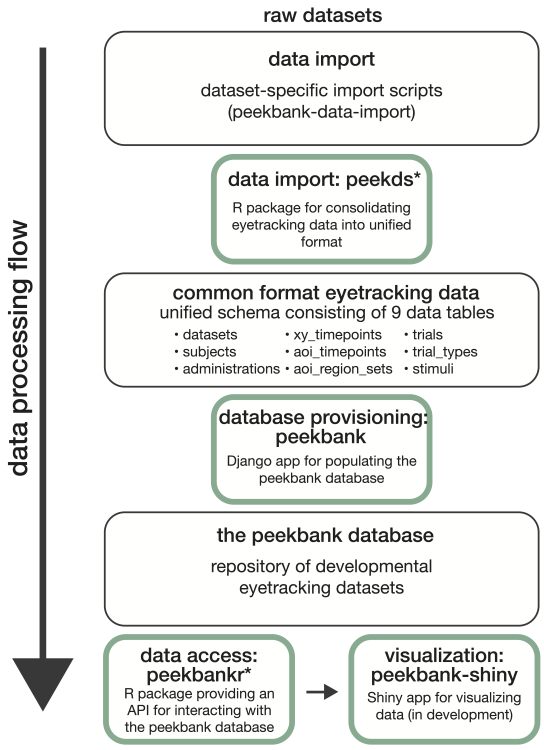

Figure 1: Overview of the peekbank data ecosystem.

Peekbank: A large-scale database of looking-while-listening-studies

What these questions and challenges share is that they are difficult to answer at the scale of a single looking-while-listening study. In order to address these questions, we introduce peekbank, a flexible and reproducible interface to an open database of developmental eye-tracking studies. Here, we give a brief overview over the key components of the peekbank 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 eyetracking data across different data sources, and (c) provides an API for quickly accessing and analyzing the database.

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

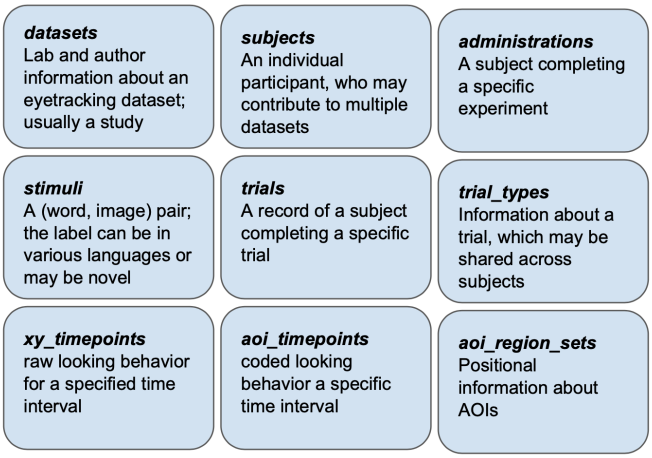

Figure 2: Data schema for the peekbank database.

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 eyetracking dataset is the lack of a shared re-usable data format among labs conducting individual experiments. Eyetracking methods and researcher teams vary in their conventions for exporting and structuring data, rendering the task of integrating datasets from different labs and data sources difficult. We developed a common, tidy format for the eyetracking data in peekbank to ease the process of conducting cross-dataset analyses. The schema of the database (Fig. 2) is sufficiently general to handle heterogeneous datasets from many studies from many labs, including both manually coded and automated eyetracking data.

During data import, raw eyetracking datasets are processed to conform to the peekbank data schema. The centerpiece of the schema is the aoi_timepoints table (Fig. 2), which records whether participants looked to the target or distracter stimulus at each timepoint of a given trial. Additional tables track information about data sources (datasets), participant characteristics (subjects, administrations), trial characteristics (trials, trial_types), stimuli (stimuli), and raw eyetracking data information (xy_timepoints, aoi_region_sets). In addition to unifying the data format, we conduct several additional preprocessing 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 consist of monolingual native English speakers, with the exception of XX. 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 eyetracking methods to measure children's gaze behavior. Most studies focused on testing familiar items, but the database also includes studies in which both familiar words and novel pseudowords were tested.

Results

General descriptives

(Fig. 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. For example, (**fernald1998?**) famously showed ...

Following the approach of Mirman (2014), we used growth curve modeling to assess the timecourse of children's fixations to the target object at different ages, generalizing across items. Specifically, we predicted children's proportion of target looking during the critical window from the interaction between age and four orthogonal polynomial time terms (linear, quadratic, cubic, and quartic). We included by-item and by-dataset random effects. Figure 4 depicts the model fit at four different age bins (though not that age was analyzed continuously in the model).

We found that XXXXX.

Predicting Target Fixation from Word AOA

In the next analysis, we asked whether properties of a specific item - in particular, the age of acquisition (AOA) for a particular item - predict children's lexical processing. Using estimates of the age of acquisition derived from Wordbank (Frank, Braginsky, Yurovsky, & Marchman, 2017) for the target and the distractor word, we modeled whether earlier-acquired target words are more likely to be fixated accurately. XX.

Time Window Selection

Taking a similar approach to that of Peelle & Van Engen (2020), we conducted a multiverse-style analysis considering possible time windows which researchers might select analyze 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. Although researchers typically consider windows in the range of 300 ms post-target onset to approximately 1500-1800 ms post-target onset, we examined a much broader 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 the greatest reliability, we expect to see very low reliability (i.e. 0) in windows that start and end 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 5, 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 of window choices, and generally only negative (blue) or 0 (white) when the start time is less than ˜500 ms and the end time is less than ˜1000 ms – far lower for both parameters than researchers generally consider. 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 can we make? 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 much

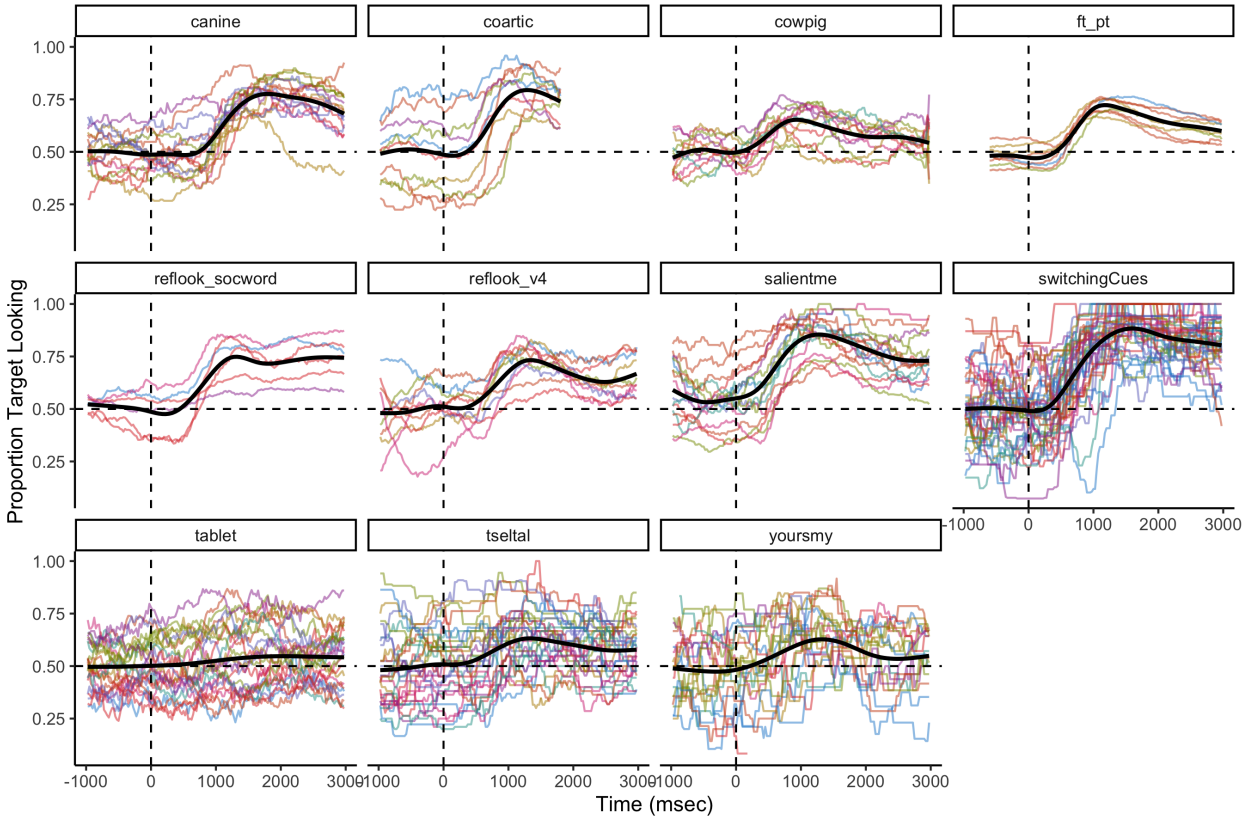

Figure 3: Item-level variability in proportion target looking within each dataset. Colored lines represent specific target labels.

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 data collection and aggregation in developmental science (Bergmann et al., 2018; The ManyBabies Consortium, 2020). Peekbank expands on these efforts by building an infrastructure for aggregating eyetracking 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 the peekbank database aims to address: understanding item-level variability in children's lexical processing and providing data-driven guidance on methodological choices.

Diving into more specifics

Future directions and limitations

limitations in language background (almost entirely English, monolingual)

limitations in participant background (almost entirely WEIRD participants)

growing the database will address these questions while increasing our power to answer the key generalization questions of interest

expanding beyond child lexical processing: tools and infrastructure can in principle be expanded to accommodate any eyetracking paradigm used with infants and toddlers

Acknowledgements

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

References

10 Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(9), 3253–8. http://doi.org/10.1073/pnas.1113380109

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. http://doi.org/10.1111/cdev.13079

Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early productive vocab-

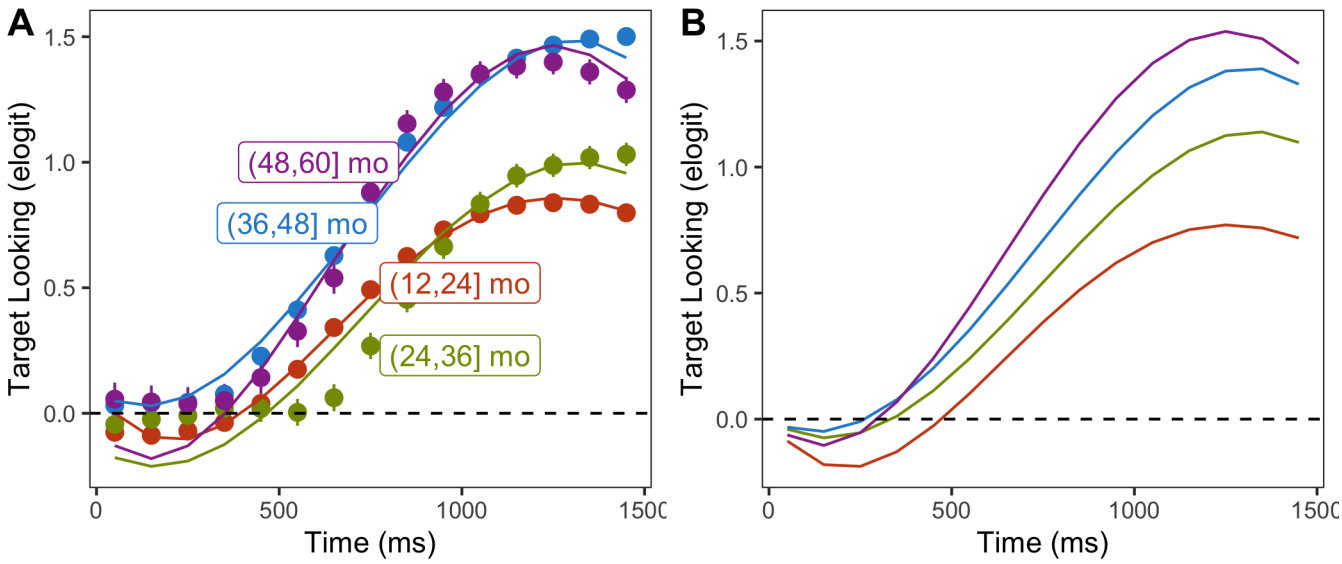

Figure 4: Growth curve models of proportion target looking during the critical target window at each age range (in months).

ulary predicts academic achievement 10 years later. *Applied Psycholinguistics*, *37*(6), 1461–1476. http://doi.org/10.1017/S0142716416000060

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, *3*, 52–67. http://doi.org/10.1162/opmi_a_00026

Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Lengyel, M. (2016). Statistical treatment of looking-time data. *Developmental Psychology*, *52*(4), 521–536. http://doi.org/10.1037/dev0000083

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. http://doi.org/10.1017/S0305000916000209

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. http://doi.org/10.1177/1745691613484936

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. http://doi.org/10.1017/S0305000907008641

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. http://doi.org/10.1017/S030500090001271X

Huang, Y., & Snedeker, J. (2020). Evidence from the visual world paradigm raises questions about unaccusativity and growth curve analyses. *Cognition*, *200*, 104251. http://doi.org/10.1016/j.cognition.2020.104251

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. http://doi.org/10.1111/j.1467-9280.2007.01871.x

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 Developmental & Behavioral Pediatrics*, 1. http://doi.org/10.1097/DBP.0000000000000541

Mirman, D. (2014). *Growth curve analysis and visualization using R*. CRC Press.

Peelle, J. E., & Van Engen, K. J. (2020). Time stand still: Effects of temporal window selection on eye tracking analysis. http://doi.org/`https://doi.org/10.31234/osf.io/pc3da`

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. http://doi.org/10.1073/pnas.1419773112

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*. http://doi.org/10.1177/2515245919900809

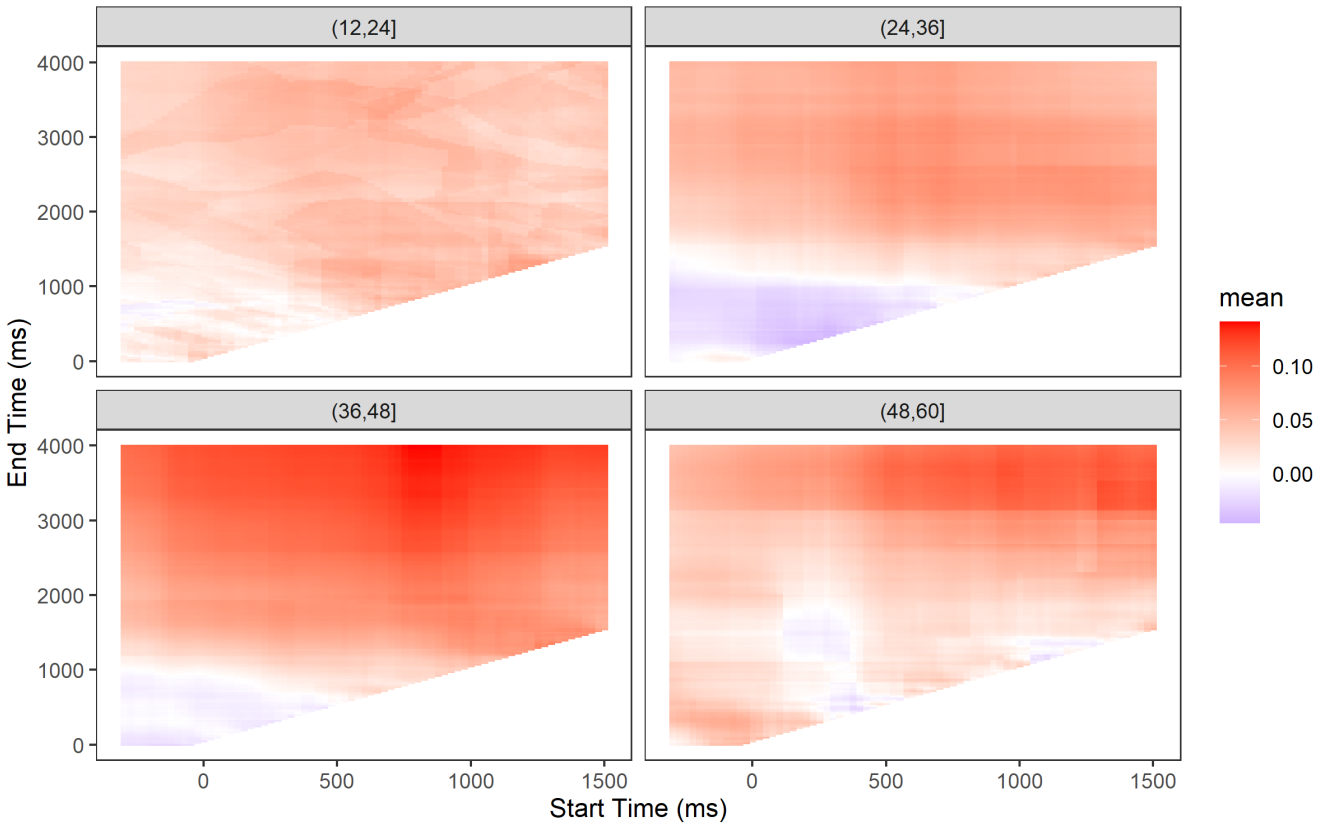

Figure 5: Participants' average inter-item correlation for proportion of looking time to familiar targets, as a function of window start time and end time, with each facet showing a different age group. More positive (red) correlations are more desirable, and blue/white represent start/end time combinations that researchers should avoid.