A performance model for early word learning

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Molly L. Lewis
mll@stanford.edu
Department of Psychology
Stanford University

Kyle MacDonald
kyle.macdonald@stanford.edu
Department of Psychology
Stanford University

Abstract

The emergence of language around a child's first birthday is one of the greatest transformations in human development. Does this transition require a fundamental shift in the child's knowledge or beliefs, or could it instead be attributable to more gradual changes in processing abilities? We present a simple model of cognitive performance that supports the second conclusion. The premise of this model is that any cognitive operation requires multiple steps, each of which require some time to complete and have some probability of failure. We use meta-analysis to estimate these parameters for two components of simple ostensive word learning: social cue use and word recognition. When combined in our model, these estimates suggest that learning should be very difficult for children younger than around a year, especially with gaze alone. This model takes a first step towards quantifying performance limitations for cognitive development and may be broadly applicable to other developmental changes.

Keywords: Speed of processing; development; word learning; meta-analysis

Introduction

Human beings begin their lives as helpless infants and yet rapidly become children who are able to perceive, act, and communicate. Infants who cannot communicate become toddlers who use words to share attention and indicate their desires. Toddlers who cannot follow the trajectory of a ball become preschoolers who can. A fundamental question of developmental psychology is how these external behavioral differences come about via internal processes of developmental change.

One possibility is that these external transitions are a product of radical internal shifts, such as the discovery of the communicative function of language, or the emergence of a theory of others' minds. Such shifts have been a centerpiece of constructivist theories of development from Piaget (1969) onward. These theories have obvious appeal, at least in part because the outward changes in children's cognitive abilities are so dramatic. Yet several decades of work with infants has revealed a surprising amount of detectable knowledge about cognitive domains, often months or even years prior to these external manifestations (Carey, 2009). How could these two sets of observations co-exist?

Perhaps children's intense performance limitations—basically, difficulties *using* knowledge or representations that they nevertheless possess—limit our abilities to observe or even to measure their competence (Chomsky, 1965). Perhaps planning to reach for an object is difficult and time-consuming enough that toddlers lose track of what they were looking for (Keen, 2003). And perhaps infants are trying to learn the meanings of words, they are just too slow and error-prone to make much progress in this task. In some sense, this hypothesis constitutes a strong null model of development: speed and accuracy *must* change, even if no internal representations do.[1]

In the current paper, we take up the challenge of building such a null model, using early word learning as a case study for exploring the role of performance limitations. The emergence of language is one area where theoretical views have differed widely. Must children master a particular insight about the role of language in communication to begin learning words in earnest (Hollich, Hirsh-Pasek, & Golinkoff, 2000), or are they pursuing the same activity throughout early childhood, but with more success later on (McMurray, 2007)?

Some empirical data support the possibility of early communicative competence. At their first birthday, infants have some expectations about the function of words in communication and show longer looking times when those expectations are violated (Vouloumanos, Onishi, & Pogue, 2012). And 6- to 9-month-olds perform above chance in word-object mapping tasks (Bergelson & Swingley, 2012). But the level of performance they show compared with older children is so limited that the data also seem to provide *prima facie* evidence for some kind of shift in representation.

We suggest instead that continuous developmental processes might be responsible, specifically increases in the speed and reliability of internal cognitive processes. We pursue this suggestion by creating a performance model for early word-object mapping. Our starting point is the idea that even the simplest word learning input for object referents involves following some kind of attentional cue (e.g., gaze or pointing) to a distal target and then processing some kind of link between a word and the target referent. Each of these abili-

[1] Note that this viewpoint does not entail any sort of nativism at all. It merely suggests that the relevant competence emerges significantly earlier than is typically supposed.

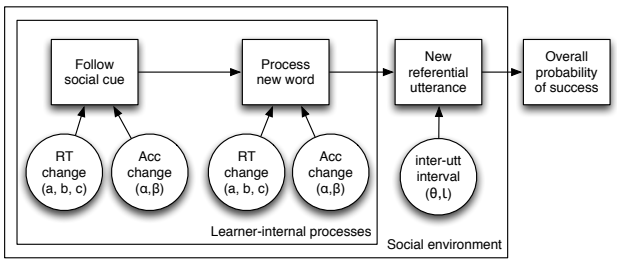

Figure 1: A schematic visualization of our model of early word learning, with the stages of processing (squares) along with the relevant parameters for each stage (circles).

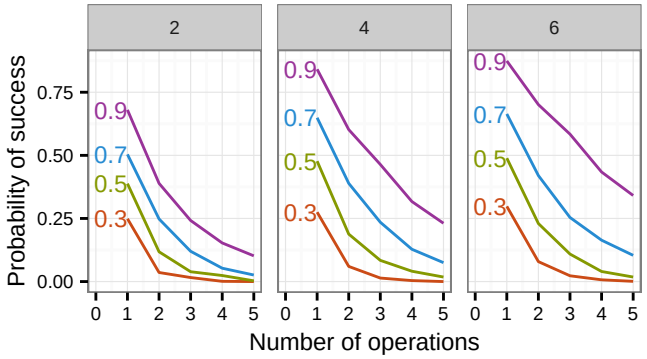

Figure 2: Probability of successsfully executing chains of cognitive operations—each with their own speed and reliability—with different numbers of steps (shown by different colors). Facets show diferent temporal thresholds.

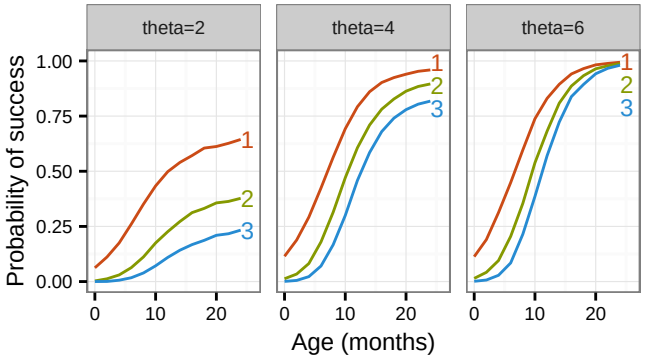

Figure 3: Probability of success for one set of developmental parameters. Different colored lines indicate chains of differing lengths, panels show different temporal thresholds.

ties has been shown to develop dramatically over the first two years and beyond. So it stands to reason that any achievement that depends on both should develop even *more* dramatically in the same period.

Our goal is to create a quantitative model that allows us to formalize this intuition. Inspired by recent meta-analyses of developmental phenomena (e.g., Cristia, Seidl, Junge, Soderstrom, & Hagoort, 2014), we conduct systematic literature reviews of the literature on social cueing (e.g., gaze following; Scaife & Bruner, 1975) and word recognition (Fernald, Pinto, Swingley, Weinbergy, & McRoberts, 1998). These meta-analytic surveys, in combination with parametric models of development (e.g., Kail, 1991) allow us to estimate the speed and accuracy for a two-component model of word learning.

The outline of the paper is as follows. We begin by describing the basic model and how it captures developmental changes. We next estimate the development of speed and accuracy independently for social cueing and word recognition. We then estimate the pace of referential utterances from a corpus, and compute children's predicted learning rate based on these parameter estimates. The conclusion of our analysis is that even if young infants were trying to learn in precisely the same way as older toddlers, they would be too slow and too fallible to extract much signal from their input data.

Model

The basic assumptions of our performance model are familiar from cognitive architectures that attempt to capture specifics of cognitive processes (e.g., ACT-R; Anderson, 1996), namely, every cognitive operation has a processing time and a probability of failure. Each complex cognitive operation is decomposable into a chain of simpler operations, any one of which can fail. And if a single link in the chain fails, then the overall operation fails as well. Thus, the probability of failure is the product of the individual failures. Similarly for timing, the total processing time for a chain of operations is the sum of the processing times for the parts.

Complex actions are describable at many different granularities. For example, word learning from an ostensive cue (e.g., parent says "doggie!" while pointing at a dog) can be decomposed into 1) social cue following and 2) word recognition/mapping. But social cue following can be further decomposed into 1a) attending to the cue, 1b) processing the directionality of the cue, and 1c) executing an eye-movement to the cue's target. Each of these could easily be broken down further. There is no one decomposition of a task, but we view this feature as a strength rather than a weaknesses of the basic framework, which can be applied to units at any grain size for which speed and reliability of processes can be measured. The overall model architecture is shown in Figure 1.

Chains of mental processes

Consider a sequence of interacting mental processes. We assume that each of these has a Bernoulli success probability s_p. Thus, the probability of a sequence of failures is exponential such that $p_{success} = \prod_p s_p$. And each operation also takes some time to complete t_p, which we assume is distributed log-normally. Thus, the total time of the chain is $rt = \sum_p t_p$.[2] Finally, consider that this operation is time-sensitive, and must be completed within a temporal threshold, also sampled from a log-normal distribution with mean θ and SD ι.

We can now approximate the probability that a chain is successful within a particular threshold. A representative set of simulations are shown in Figure 2. For these and many other parameter settings, long chains of operations are unlikely to succeed unless individual operations are very fast and very accurate.[3]

Development within the model

The two posited capacities in our model are speed and accuracy (probability of a successful operation). Both of these

[2]Note that there is not a known parametric form for the sum of multiple lognormals. A variety of analytic approximations for these sums exist (Fenton, 1960), but they have some limitations, so instead we use numerical simulations here.

[3]All code and data for the simulations reported here is available at `http://github.com/mcfrank/sop`.

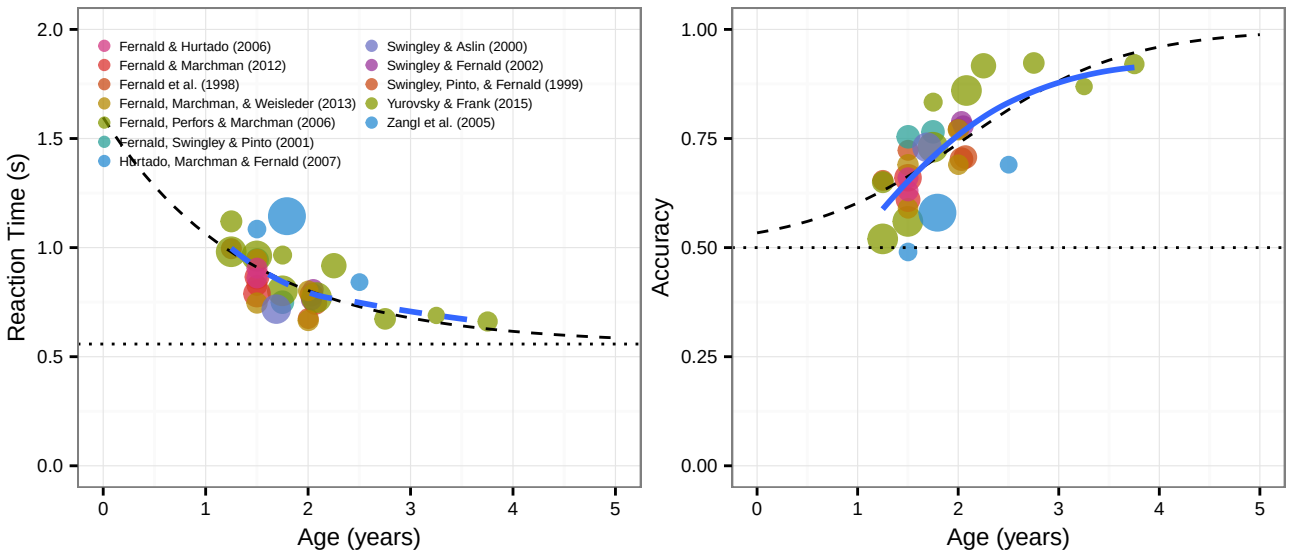

Figure 4: Reaction times (left) and accuracies (right) for two-alternative forced-choice word recognition paradigms. Each circle shows an experiment, with areas scaled by the number of participants. A generic loess smoothing function is shown in blue. For reaction times, dotted line shows adult reaction time and dashed line shows best-fitting Kail function. For accuracy, dotted line shows chance, and dashed lines show best-fitting half-logit function.

should change across development for any constituent cognitive operation, leading to dramatic changes in the cumulative speed and accuracy of chains of operations across development. To estimate these changes, we use parametric models of developmental change.

Pioneering work by Kail (1991) describes the developmental trajectory of reaction times for complex tasks, via aggregation across the published literature. Empirically, the slope of these reaction times follows an exponential, such that $Y(i) = a + be^{-ci}$, where Y is the predicted variable, a is the eventual (adult) asymptote, b is the multiplier for the (infant) intercept, c is the rate of development, and i is age. The Kail (1991) model is a model of RT multipliers. Since operations are additive, these multipliers should apply to individual operations or to chains of operations equivalently: if the multiplier is constant then it can be factored out.

Next we turn to accuracy. For simplicity, we consider the probability of success on a single operation changing across time as a simple logistic function where $Y(i) = \frac{1}{1+e^{\alpha+\beta i}}$. α sets the intercept and β marks the developmental multiplier, as in a standard logistic regression.

Preliminary simulations

We can combine these functions with the basic operation-chain simulations defined above and examine the probability of a successful chain of operations across ages. Results for one parameter set are shown in Figure 3. These simulations show that sharp developmental transitions from failure to success can be the product of relatively broad underlying functions. But the difficulty is constraining the model's predictions requires substantial information about speed and accuracy. In the next section we turn to the estimation of these parameters via meta-analysis.

Case Study: Early Word Learning

Why do children begin to show evidence of word learning around their first birthday? Although many accounts have been proposed (e.g. Tomasello, 1995's "nine-month revolution"), our null-model framework provides a simple explanation. Children may be trying to learn words from very early in development, but the basic cognitive components may be too slow and too challenging to allow for consistent learning (and consistent measurement of that learning by psychologists). The recent literature on early word learning gives some support for this contention, as careful measurement has revealed some aspects of receptive language prior to the first birthday (Bergelson & Swingley, 2012).

We focus here on learning a word that is presented ostensively via a social cue like gaze or pointing. For simplicity, we decompose the task of social word learning into two abilities: 1) social cue following, and 2) word recognition. This task analysis is an approximation: pointing is not the same as gaze following (and neither is it always necessary), and recognition is not the same as learning and retention. But it nevertheless captures some aspects of the task, namely following a social cue to a distal target and processing some language associated with that target. And it has the major benefit for our purposes of providing data on development, since each of these tasks is well-studied.

Word recognition

We first estimated developmental changes in the speed of processing for word recognition. Fernald et al. (1998) introduced the method of using eye-movements to measure children's ac-

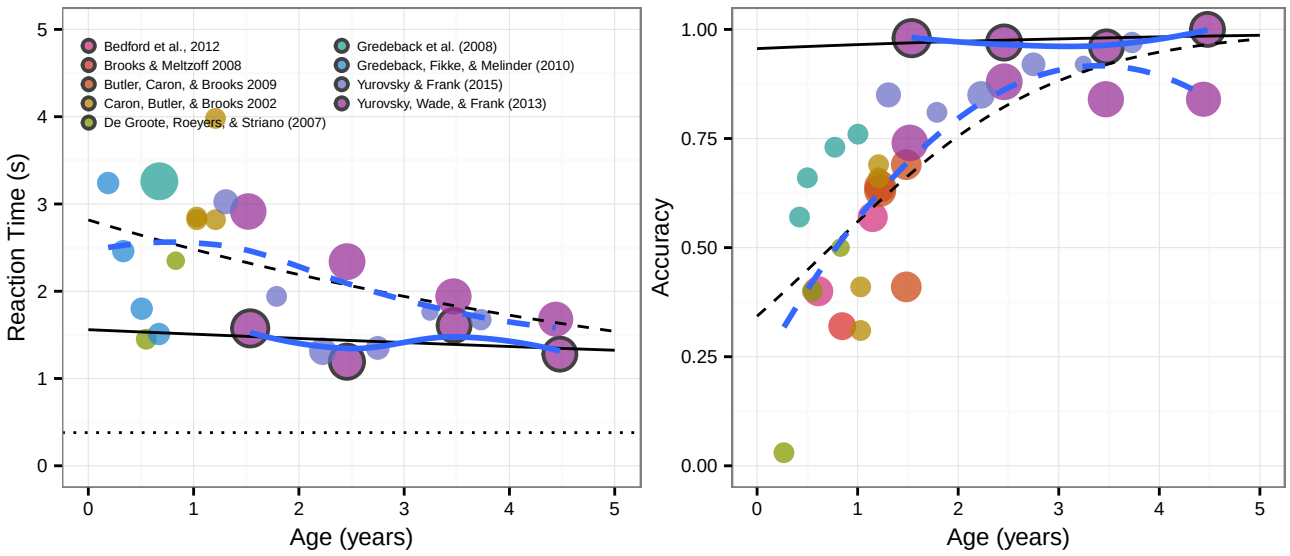

Figure 5: Reaction times (left) and accuracies (right) for social cue following. Plotting conventions are as above. Results for gaze following alone are shown with no border; results for gaze plus pointing are shown with a black border. Dashed and solid lines show fits for gaze and gaze plus pointing, respectively.

curacy and reaction time across development. Subsequent investigations have identified many factors involved in speed of word recognition (e.g., the frequency and familiarity of the target word). Nevertheless, such research generally attempts to select simple, easy words that should be accessible to most children, so we can use this literature to derive rough estimates of accuracy and reaction time across development.

We conducted a systematic literature review by using Google Scholar to identify peer-reviewed papers citing Fernald et al. (1998). We screened this sample manually to find the subsample of 12 papers that reported both accuracy and reaction time with sufficient detail to permit coding. Figure 4 shows reaction times from this sample of papers, plotted by the mean age of the children in the reported studies. The dashed line shows an exponential function fit to these data (with $a = 0.56$, $b = 1.04$, and $c = 0.72$). The intercept a is estimated as the mean reaction time for adults in control experiments.[4]

Accuracies can be estimated similarly. We fit a logistic regression to accuracy data, using a half-logit linking function ($.5 + .5 \times \frac{1}{1+e^{-x}}$) to bound accuracy between .5 and 1. Figure 4 shows the results of this analysis, with estimated parameters $\alpha = -2.62$ and $\beta = 1.27$. In both of these cases we see a qualitatively strong fit between the developmental model we assumed and the particulars of the experimental data (e.g., as estimated by a naive smoothing model, shown in blue).

[4]Note that Kail (1991)'s original analysis was of the slope of mental rotation speeds, so the exponential curve described a multiplier on the adult slope. This analysis is simpler, fitting a curve to the mean RTs directly.

Social cue following

For our analysis of social cue following, we were interested in both pure gaze following and gaze following supplemented by pointing. We identified papers using a Google Scholar search for "gaze following" and included those studies that A) included data from typically-developing children, B) used a standard face-to-face gaze-following task, and C) reported percentage accuracy (rather than a score or other composite measure). Although we coded all papers that fit these criteria, we focused on papers with a simple two-alternative forced choice (9 papers); integrating across different numbers of alternatives added additional complexity to our model.

In our first iteration of this analysis, we found that very few studies reported reaction times for gaze following, and those that did had no data from children older than 15 months and no data from gaze plus pointing. Estimating developmental curves for these data was difficult; to remedy this issue we include new analyses of data from Yurovsky, Wade, & Frank (2013) and Yurovsky & Frank (2015). Both of these paradigms were word learning experiments in which a social cue (either brief or extended) was used to indicate a referent. Data from these studies can thus be used to estimate social cue following time; conveniently, both studies included large numbers of participants at older age ranges, constraining our analysis. We also added 440ms as a floor adult reaction time, on the basis of measurements by Driver et al. (1999).

Figure 5 shows reaction time and accuracies for gaze following, both with and without pointing. Again we see relatively good qualitative fit by the developmental models. Details of these fits are identical to our analysis of word recognition, except that we use a standard logit function. We experimented with using a half-logit link, but found that many

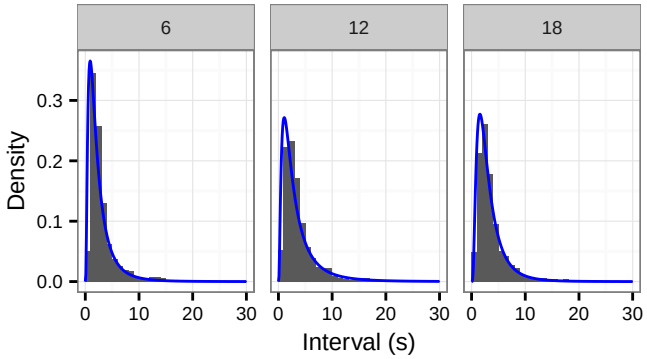

Figure 6: Histograms of inter-utterance intervals in child-directed speech. Panels show data from 6-, 12-, and 18-month-olds. Blue curves show the best-fitting log normal distribution for the full dataset.

studies reported accuracies below .5 due to children's failure to disengage from the experimenter's face.

Estimating a temporal threshold

We next estimate plausible values for θ and ι, which control the distribution of temporal thresholds at which referential utterances must be processed. To estimate these values, we turned to the Fernald and Morikawa corpus (Frank, Tenenbaum, & Fernald, 2013), which contains a set of transcribed interactions between caregivers and children as they play with pairs of objects. Critically, this corpus contains approximate timing information (Rohde & Frank, 2014) as well as annotations for social cues used by the caregivers (e.g., gaze, pointing, etc.) to indicate which toy is being talked about.

The primary variable of interest for our analysis was the distribution of time intervals between utterances using social cues to refer to objects. Preliminary analyses indicated that there were no differences in timing between referential utterances (those containing a concrete reference to an object in the current context) and non-referential utterances, so we examined the full distribution of inter-utterance intervals. Figure 6 shows the empirical distribution across ages, along with the best-fitting log-normal distribution. The mean time between utterances was 3.61 and the median was 2.5. Perhaps surprisingly, there were no major differences in the distribution between age groups (for example, $\text{median}_6 = 2$, $\text{median}_{12} = 2.5$, and $\text{median}_{18} = 2.5$), suggesting that parents were not substantially adjusting the pace of conversation to children, at least in this corpus. For our simulations below, we use the parameters of the best-fitting distribution across ages ($\theta = 0.78$, $\iota = 0.86$).

Simulations

We now run the same developmental simulations described above, but using the estimated numerical parameters for word processing, social cue following, and the timing of utterances in child-directed speech. Figure 7 shows the probability of successful ostensive learning by age, split by social cue.

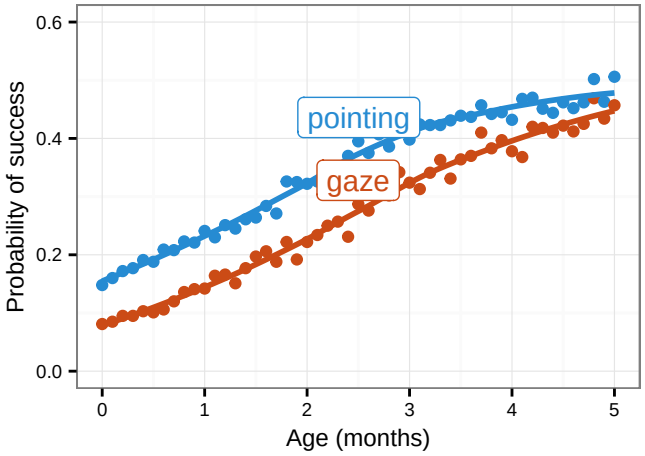

Figure 7: Simulation results using meta-analytic parameters estimated for speed and accuracy in social cue following and word recognition.

Pointing is substantially more effective than gaze following, but under this na"ive model, neither cue ever leads to success more than half of the time. While low accuracy constrains performance early in development, accuracy is above 80% by age 3 and reaction time is the bounding factor.[5]

A second way of viewing these same simulations is shown in Figure 8, which shows the average number of learning instances a learner would need to achieve a single successful mapping trial under this model. This number declines from more than 10 in early infancy to an asymptote of approximately two. Here the difference between gaze and pointing is more apparent, especially earlier in development. This difference is congruent with the difficulties in word learning from gaze observed by Yurovsky et al. (2013).

Is it reasonable to assume that not all word mapping opportunities succeed? After all, 3–4-year-old children have been show to learn words from a single exposure (Carey, 1978). We would argue that a success rate of approximately one half at age three is actually very congruent with accuracies in the 60-80% rate shown by Markson & Bloom (1997) and others (given a handful of exposures during training). And rates of learning for younger children also show some numerical congruence (e.g., Woodward, Markman, & Fitzsimmons (1994) showed some evidence of learning from nine exposures in 13- and 18-month-olds). In sum, even though average results suggest that some children may learn from a small set of exposures, they do not imply that *all* children have learned.

Discussion

We presented a model of children's developing processing capabilities in which the sole age-related changes are in the

[5]One consequential decision for the model is the assumed response function for social cues. Here we have used a half-logit link function for accuracy, assuming a two-alternative forced choice between referents. Without this assumption, learning probabilities for early infancy go down substantially, but the asymptote remains unchanged.

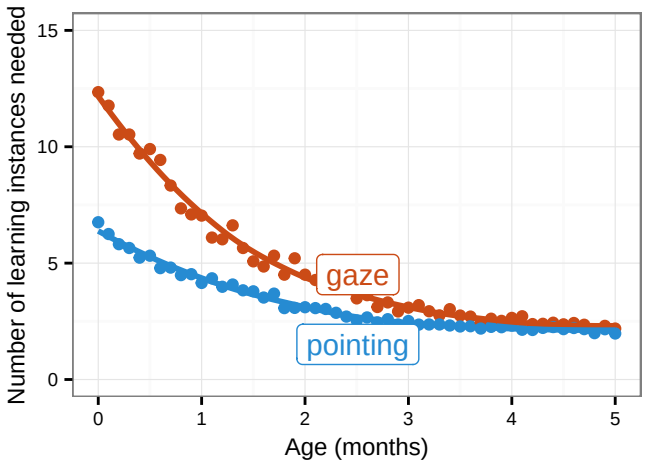

Figure 8: Average number of learning instances needed per effective learning instance, plotted by age.

speed and accuracy of mental operations. This strong null model is a first step towards a baseline model of cognitive development, attempting to answer the question of what changes we would see even if there were *no* substantive differences in children's internal representations across ages. Using measurements extracted from the literature, we used the model to mkae predictions about early word learning, a domain in which there has been uncertainty about the presence of early representational shifts. Without including any representational changes, the model still produced large developmental changes and showed *prima facie* congruence with some previous experimental work.

The simple model we described here has many limitations. First, we estimated parameters from the data that were available rather than the data we would have liked to have (e.g., reaction times from familiar word recognition rather than from novel word learning). Second, we assumed a purely serial model of responding in which accuracy and reaction time were independent from one another; more sophisticated models of decision-making might link processes in a race model or yoke accuracy and reaction-time in a speed-accuracy tradeoff. Third, although we did not find evidence of large developmental changes in the pace of utterances, parents likely still adapt to their children's speed of processing in some instances, leading to better outcomes for those instances.

Despite the many assumptions and limitations of our model, the results should still constrain and inform our theories. If we discard the parametric form of the model and simply examine the meta-analytic reaction times we estimated, we see that they are generally longer than the interval between new utterances, suggesting that word learning through gaze is likely to be difficult in the first year under any model. This qualitative observation suggests the basic intuitions derived from our model may be useful for analyzing other domains. More generally, the null model we articulated here should reinforce the point that—even with the most sensitive measurements available—we should not infer a lack of competence from a failure in performance.

Acknowledgements

This work supported by NSF BCS #1528526. Thanks to Ellen Markman, Rebecca Saxe, and members of the Language and Cognition Lab at Stanford for helpful discussion.

References

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355.

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*, *109*(9), 3253–3258.

Carey, S. (1978). The child as word learner. In *Linguistic theory and psychological reality* (pp. 264–293). Cambridge, MA: MIT Press.

Carey, S. (2009). *The origin of concepts*. Oxford, UK: OUP.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT press.

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures. *Child Development*, *85*(4), 1330–1345.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition*, *6*(5), 509–540.

Fenton, L. F. (1960). The sum of log-normal probability distributions in scatter transmission systems. *IRE Transactions on Communications Systems*, *8*(1), 57–67.

Fernald, A., Pinto, J. P., Swingley, D., Weinbergy, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*, *9*(3), 228–231.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, *9*(1), 1–24.

Hollich, G. J., Hirsh-Pasek, K., & Golinkoff, R. M. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, 1–135.

Kail, R. (1991). Developmental change in speed of processing during childhood and adolescence. *Psychological Bulletin*, *109*(3), 490.

Keen, R. (2003). Representation of objects and events why do infants look so smart and toddlers look so dumb? *Current Directions in Psychological Science*, *12*(3), 79–83.

Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, *385*(6619), 813–815.

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, *317*(5838), 631–631.

Piaget, J., & Inhelder, B. (1969). *The psychology of the child*. Basic Books.

Rohde, H., & Frank, M. C. (2014). Markers of topical discourse in child-directed speech. *Cognitive Science*, *38*(8), 1634–1661.

Scaife, M., & Bruner, J. S. (1975). The capacity for joint visual attention in the infant. *Nature*.

Tomasello, M. (1995). Joint attention as social cognition. *Joint Attention: Its Origins and Role in Development*, 103–130.

Vouloumanos, A., Onishi, K. H., & Pogue, A. (2012). Twelve-month-old infants recognize that speech can communicate unobservable intentions. *Proceedings of the National Academy of Sciences*, *109*(32), 12933–12937.

Woodward, A. L., Markman, E. M., & Fitzsimmons, C. M. (1994). Rapid word learning in 13-and 18-month-olds. *Developmental Psychology*, *30*(4), 553.

Yurovsky, D., & Frank, M. C. (2015). Beyond nave cue combination: Salience and social cues in early word learning. *Developmental Science*.

Yurovsky, D., Wade, A., & Frank, M. C. (2013). Online processing of speech and social information in early word learning. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*.