Adults and preschoolers seek visual information to support language comprehension in noisy environments

Kyle MacDonald, **Virginia Marchman**, **Anne Fernald**, and **Michael C. Frank**
{kylem4, marchman, afernald, mcfrank} @stanford.edu
Department of Psychology, Stanford University

Abstract

Language comprehension in grounded contexts is a multisensory process in whcih listeners rapidly integrate information from both the visual and the linguistic signals. But how should we prioritize these different information sources? Here, we test the hypothesis that even young listeners will flexibly adapt the dynamics of gaze to seek higher value visual information by increasing visual fixations to a speaker when the auditory signal is less reliable. We measured the timing and accuracy of adults (n=31) and children's (n=40, 3-5 y.o.) eye movements during a real-time language comprehension task. We found that both age groups delayed the timing of gaze shifts away from a speaker's face when processing speech in the presence of background noise. This delay resulted in listeners gathering more information from the visual signal, which results in more accurate gaze shifts, and fewer random eye movements to the rest of the visual world. These results suggest that even young listeners respond to the demands of different processing contexts by adapting gaze patterns to support language comprehension.

Keywords: eye movements; language processing; information-seeking; speech in background noise; development

Introduction

As skilled listeners, we continually integrate information from the visual and the linguistic signals to understand what others are saying. A classic demonstration of this integration process is the "McGurk effect" where a speaker's mouth movements suggest one sound while their acoustic output suggests another. This conflict results in the listener perceiving a third, intermediate sound (J. MacDonald & McGurk, 1978). Findings such as these have inspired prominent theories of speech perception (McClelland, Mirman, & Holt, 2006) and lexical processing (M. C. MacDonald & Seidenberg, 2006; Smith, Monaghan, & Huettig, 2017) that argue for the importance of *interactive* processes – where listeners integrate information from multiple sources in parallel. Moreover, empirical work on speech perception shows that adults are better able to recover linguistic information in noisy contexts when they have visual access to a speaker's face (Erber, 1969)

However, the usefulness of integrating visual information varies depending on features of the listener and features of the processing context. Consider the familiar example of a friend who asks you to "Pass the salt" in a noisy restaurant. Here, comprehension could be facilitated by gathering visual information by allocating visual attention to the speaker to read her lips or the direction of her gaze. A second case study is the understanding a visual-manual language like American Sign Language (ASL). Here, the value of allocating visual fixations to the language source (i.e., the signer) is high since all of the language-relevant information is available in that location.

In prior work, we showed that, compared to spoken language learners, ASL-learners will delay shifting gaze away from a language source until they have accumulated sufficient information to generate a highly-accurate eye movement (K. MacDonald, Blonder, Marchman, Fernald, & Frank, 2017). In contrast, spoken language learners were more likely to generate early, exploratory gaze shifts. We explained these differences using an information-seeking account: that listeners flexibly adapted the dynamics of gaze in response to the value of gathering visual information within a particular language processing context.

Our account represents a synthesis of ideas from several research programs. First, work on language-mediated visual attention shows that adults and children rapidly shift gaze upon hearing the name of an object in the visual scene (Allopenna, Magnuson, & Tanenhaus, 1998; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). The speed and consistency of this visual response has led to debates about whether language-mediated gaze shifts are automatic as opposed to under the control of the listener. Second, empirical work on vision during natural tasks shows that people overwhelmingly prefer to look at *goal-relevant* locations – e.g., an upcoming obstacle while walking (Hayhoe & Ballard, 2005). These accounts predict that gaze dynamics during language comprehension should adapt to the value of different fixation behaviors with respect to the listener's goal of language comprehension. Finally, work on "effortful listening" shows that listeners generate compensatory responses (e.g., increases in attention and working memory) within "challenging" comprehension contexts such as processing noisy or accented speech (Van Engen & Peelle, 2014). These accounts predict that young listeners might compensate for the reduced quality of the auditory signal by allocating gathering additional visual information.

Here, we test the generality of our information-seeking account of eye movements during grounded language comprehension. We ask whether listeners will adapt the timing of gaze shifts away from a speaker when the auditory signal becomes less reliable – as is the case when processing speech in noisy environments. Recent evidence suggests that gaze during lexical access can adapt to the demands of different processing contexts. For example, recent work by McMurray, Farris-Trimble, & Rigler (2017) shows that individuals with Cochlear Implants, who are consistenly processing degraded auditory input, are more likely to delay the process

of lexical access as measured by slower gaze shifts to named referents and fewer incorrect gaze shifts to phonological onset competitors, as compared to listeners with typical hearing. McMurray et al. (2017) also found that they could replicate these changes in lexcial access in adults with typical hearing by degrading the auditory stimuli in a way that shares features with the output of a cochlear implant (noise-vocoded speech).

A second goal of this work is to test whether children would show a similar pattern of flexibly adapting fixation behaviors in response to changes in the utility of gathering certain kinds of visual information. Recent developmental work shows that, like adults, preschoolers will flexibly adjust how they interpret ambiguous sentences (e.g., "I had carrots and *bees* for dinner.") by integrating information about the relibility of the incoming perceptual information with their expectations about the speaker (Yurovsky, Case, & Frank, 2017). While children's behavior showed impressive parallels to adults, they relied more on top-down expectations about the speaker perhaps because there was more perceptual noise compared to adults. These developmental findings provide insight into how children succeed in efficient language comprehension despite having partial knowledge of word-object links and a fully-developed internal language model.

Here, we hypothesized that a noisy auditory environment creates a scenario where the auditory signal becomes less reliable, and in turn increases the value of fixating on a speaker for the task of language understanding. Our key behavioral prediction is that listeners in noisy contexts will delay generating an eye movement away from a speaker until they have accumulated additional visual information about the identity of the named referent. We also predicted that preschoolers would show a parallel pattern of adaptation to noisy contexts and allocate more fixations to a speaker's face when it became more useful for maintaining accurate language comprehension. To quantify evidence for our predictions, we analyze the Accuracy and Reaction Times (RTs) of listeners' first gaze shifts after hearing the name of an object in the visual scene. We focus on first shifts because they provide a window onto changes in the underlying dynamics of decision processes that generate eye movements. However, it is important to point out that when we analyze differences in accuracy, we are not making claims about the overall amount of time spent looking at the target vs. the distractor image – a measure typically used in analyses of the Visual World Paradigm.

Experiment

In this experiment, we recorded adults and children's eye movements during a real-time language comprehension task where participants processed familiar sentences (e.g., "Where's the ball?") while looking at a simplified visual world with three fixation targets (see Fig 1). Using a within-participants design, we manipulated the signal-to-noise ratio of the auditory signal by convolving the auditory input with brownian noise (i.e., random noise patterns).

First, we present standard behavioral analyses of Reaction Time (RT) and accuracy of listeners' first gaze shifts after target noun onset. Then, we present two model-based analyses that link observable behavior to underlying psychological constructs. (1) We use an exponentially weighted moving average (EWMA) method (Vandekerckhove & Tuerlinckx, 2007) to classify participants' gaze shifts as language-driven or random. In contrast to the standard RT/Accuracy analysis, the EMWA approach allows us to quantify participants' willingness to generate gaze shifts after noun onset but before collecting sufficient information to seek the named referent. Higher values indicate that participants were shifting early and equally likely to land on the target or distractor image. (2) We use drift-diffusion models (DDMs) (Ratcliff & Childers, 2015) to ask whether behavioral differences in Accuracy and RT are driven by a more cautious responding strategy or by more efficient information processing – an important distinction for our theoretical account.

We predicted that processing speech in noisy contexts would make participants less likely to shift before collecting sufficient information, which in turn would lead to a lower proportion of shifts flagged as random in the EWMA analysis, and a pattern of DDM results that indicates a prioritization of accuracy over and above speed (see the Analysis Plan section below for more details on the models). We also predicted both (a) developmental differences: that children would produce a higher proportion of random shifts and accumulate information less efficiently compared to adults, and (b) developmental parallels: that children would show an adult-like pattern of behavioral/model-based results in the noisy vs. clear processing contexts.

Method

Participants Participants were native, monolingual English-learning children ($n = 39$; 22 F, 17 M) and adults ($n = 31$; 22 F, 9 M). All participants had no reported history of developmental or language delay and normal vision. 14 participants (11 children, 3 adults) were run but not included in the analysis because either the eye tracker falied to calibrate or the participant did not complete the task.

Stimuli *Linguistic stimuli.* The video/audio stimuli were recorded in a sound-proof room and featured two female speakers who used natural child-directed speech and said one of two phrases: "Hey! Can you find the (target word)" or "Look! Where's the (target word) – see panel A of Fig 1. The target words were: ball, bunny, boat, bottle, cookie, juice, chicken, and shoe. The target words varied in length (shortest = 411.68 ms, longest = 779.62 ms) with an average length of 586.71 ms.

Noise manipulation. To create the stimuli in the noise condition, we convolved each recording with Brown noise using the Audacity audio editor. The average signal-to-noise ratio[1]

[1]The ratio of signal power to the noise power, with values greater than 0 dB indicating more signal than noise.

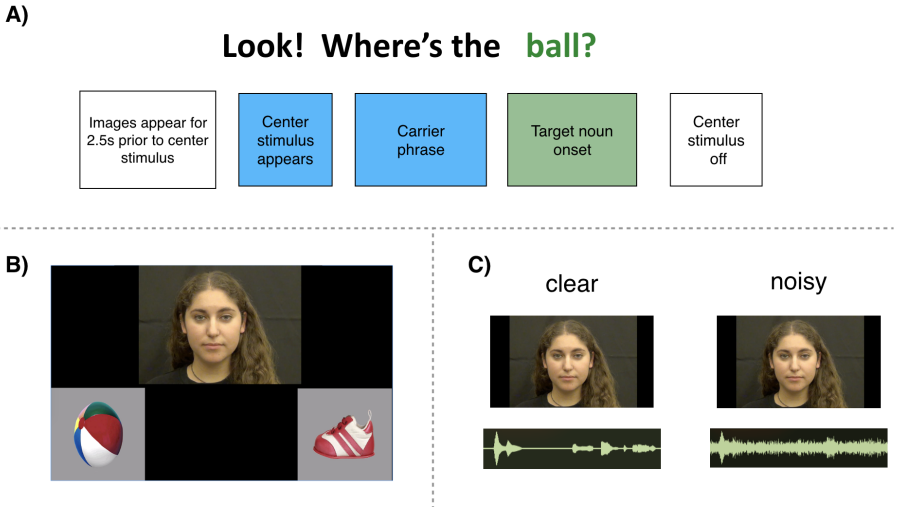

Figure 1: Stimuli information. Panel A shows the timecourse of the linguistic stimuli for a single trial. Panel B shows the layout of the three fixation locations (speaker, target, and distracter). And panel C shows a visual representation of the clear and noisy waveforms used in E1.

in the noise condition was 2.87 dB compared to the clear condition, which was 35.05 dB.

Visual stimuli. The image set consisted of colorful digitized pictures of objects presented in fixed pairs with no phonological overlap between the target and the distractor image (cookie-bottle, boat-juice, bunny-chicken, shoe-ball). The side of the target picture was counterbalanced across trials.

Design and procedure Participants viewed the task on a screen while their gaze was tracked using an SMI RED corneal-reflection eye-tracker mounted on an LCD monitor, sampling at 60 Hz. The eye-tracker was first calibrated for each participant using a 6-point calibration. On each trial, participants saw two images of familiar objects on the screen for two seconds before the center stimulus appeared (see Fig 1). Next, they processed the target sentence – which consisted of a carrier phrase, a target noun, and a question – followed by two seconds without language to allow for a response. Child participants saw 32 trials (16 noise trials; 16 clear trials) with several filler trials interspersed to maintain interest. Adult participants saw 64 trials (32 noise; 32 clear). The noise manipulation was presented in a blocked design with the order of block counterbalanced across participants.

Results and Discussion

Analysis plan First, we present behavioral analyses of First Shift Accuracy and Reaction Time (RT).[2] RT corresponds to the latency to shift away from the central stimulus to either picture measured from the onset of the target noun. (all RTs were analyzed in log space). Accuracy corresponds to whether participants' first gaze shift landed on the target or the distracter picture. We used the `rstanarm` (Gabry & Goodrich, 2016) package to fit Bayesian mixed-effects regression models. The mixed-effects approach allowed us to model the nested structure of our data – multiple trials for each participant and item, and a within-participants manipulation – by including random intercepts for each participant and item, and a random slope for each item and noise condition. We used Bayesian estimation to quantify uncertainty in our point estimates, which we communicate using a 95% Highest Density Interval (HDI). The HDI provides a range of credible values given the data and model. All analysis code can be found in the online repository for this paper: `https://github.com/kemacdonald/speed-acc/blob/master/paper/cogsci2018/README_cogsci2018.md`.

Next, we present the two model-based analyses – the EWMA and DDM. The goal of these models is to move beyond a description of the data and map behavioral differences in eye movements to underlying psychological variables. The EWMA method models changes in random shifting behavior as a function of RT. For each participant, the model classifies the proportion of shifts that were likely to be language-driven as opposed to random responding, which we call the *guessing* parameter.

After we fit the EWMA, we took shifts categorized as language-driven and fit a hierarchical Bayesian drift-diffusion model (HDDM). This model quantifies differences in separable parameters of the underlying decision process that lead to different patterns of behavior. The model assumes that people accumulate noisy evidence in favor of one alternative with a response generated when the evidence crosses a pre-defined decision threshold. Here, we focus on two parameters of interest: **boundary separation**, which indexes the amount of evidence gathered before generating a response (higher val-

[2] See `https://osf.io/g8h9r/` for a pre-registration of the analysis plan.

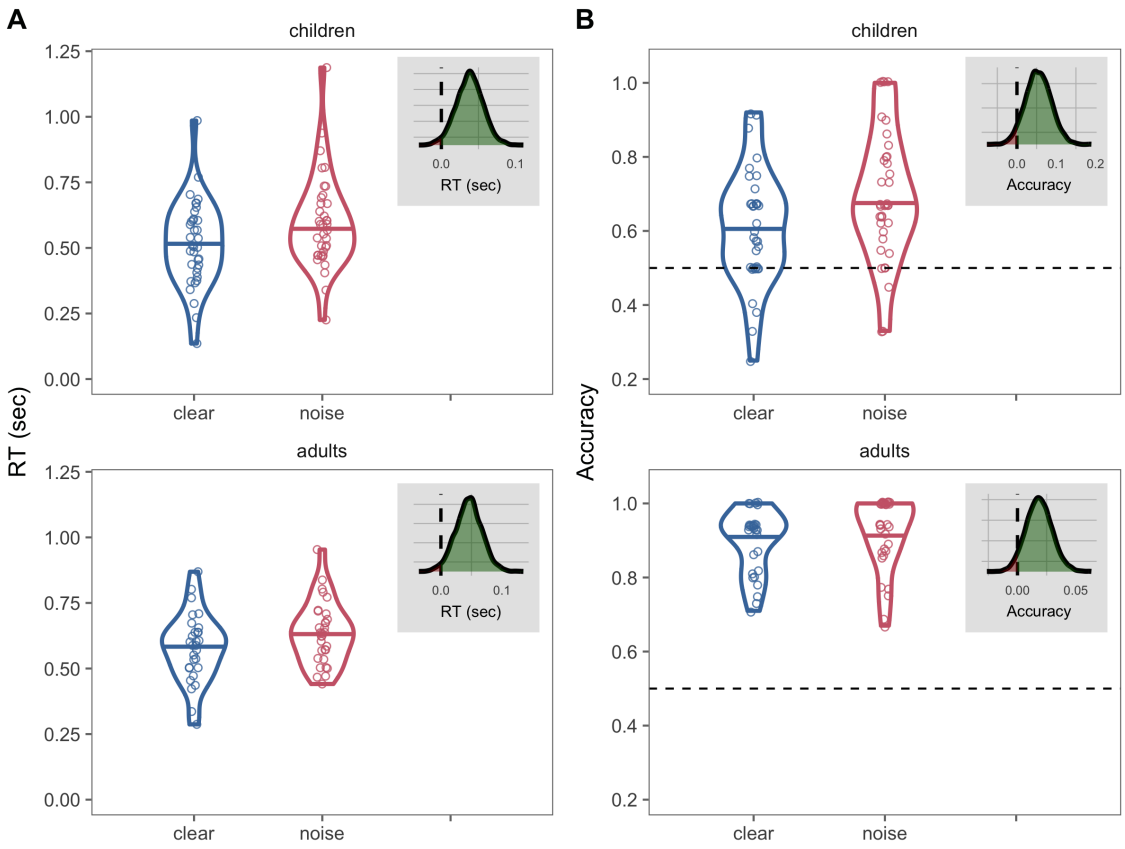

Figure 2: Behavioral results for first shift Reaction Time (RT) and Accuracy. Panel A shows violin plots representing the distribution of RTs for each participant in each condition. Each point represents a participant's average RT. Color represents processing condition. The grey insets show the full posterior distribution of the plausible RT differences across conditions with the vertical dashed line representing the null value of zero condition difference. The green shading represents estimates in the predicted direction and above the null value while the red shading represents estimates below the null. Panel B shows the same information but for first shift accuracy.

ues suggest more cautious responding) and **drift rate**, which indexes the amount of evidence accumulated per unit time (higher values suggest more efficient processing).

Behavioral analyses: RT. To make RTs more suitable for modeling on a linear scale, we analyzed responses in log space with the final model specified as: $log(RT) \sim noise\_condition + age\_group + (sub\_id + noise\_condition \mid item)$. Panel A of Figure 2 shows the full RT data distribution, the estimates of condition means, and the full posterior distribution of the estimated difference between the noise and clear conditions. Both children and adults were slower to identify the target in the noise condition (Children M_{noise} = ms; Adult M_{noise} = ms), as compared to the clear condition (Children M_{clear} = ms; Adult M_{clear} = ms). RTs in the noise condition were 41.73 ms slower on average, with a 95% HDI from 1.19 ms to 82.26 ms that did not include the null value of zero condition difference.

Accuracy. Next, we modeled adults and children's first shift accuracy using a mixed-effects logistic regression with the same specifications (see Panel B of Fig 2). Both groups were more accurate than a model of random responding (null value of 0.5 falling well outside the lower bound of the 95% HDI for all group means). Adults were more accurate ($M_{adults} = 90\%$) than children ($M_{children} = 62\%$). The key result is that both groups showed evidence of higher accuracy in the noise condition: children (M_{noise} = 67%; M_{clear} = 62%) and adults (M_{noise} = 92%; M_{clear} = 90%). Accuracy in the noise condition was on average 4% higher, with a 95% HDI from 0% to 11%. Note that the null value of zero difference falls at the very edge of the HDI such that 95% of the credible values fall below the null, providing evidence for higher accuracy in the noise condition.

Model-based analyses: EWMA. Figure 3 shows the proportion of shifts that the model classified as random vs. language-driven for each age group and processing contxt. Critically, processing speech in noise caused both adults and children to produce a higher proportion of language-driven shifts with the 95% HDI excluding the null value (see Table 1). This pattern suggests that the noise condition led participants to increase visual fixations to the language source, lead-

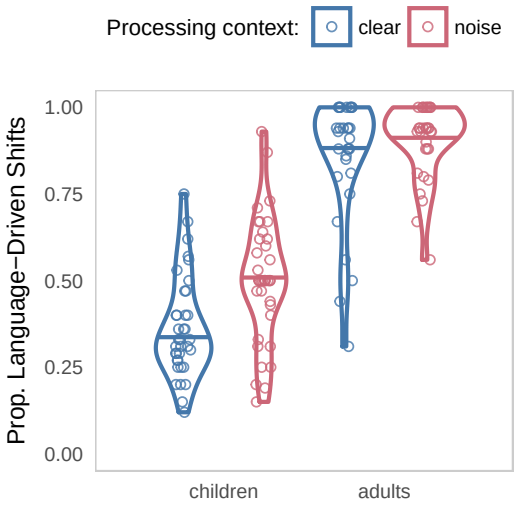

Figure 3: EWMA results for children and adults. Each point represents the proportion of shifts categorized as language-driven for a single participant. Color of represents the processing context: noise vs. clear.

Parameter	Contrast	Estimate (95% HDI)
Cut point	age group	0.21 [0.18, 0.24]
	noise	0.04 [0.01, 0.06]
Guessing	adults-children	0.46 [0.43, 0.49]
	noise-clear	0.12 [0.07, 0.17]

Table 1: EWMA results for E1 and E2. The guessing parameter refers to the proportion of gaze shifts classified as random vs. language-driven with higher values indicating more random responding. Estimate refers to the difference between condition or age group.

ing them to generate fewer exploratory, random shifts before accumulating sufficient information to respond accurately.

HDDM. Figure 4 shows the full posterior distributions for the HDDM output. Children had lower drift rates and boundary separation estimates as compared to adults, suggesting that children were less efficient and less cautious in their responding (see also Table 2). The noise manipulation only affected the boundary separation parameter, with higher estimates in the noise condition for both age groups. This result suggests that participants' in the noise condition prioritized information accumulation over speed when generating an eye movement in response to the incoming language. This increased decision threshold led to higher accuracy. Moreover, the high overlap in estimates of drift rate suggests that participants were able to integrate the visual and auditory signals such that they could achieve a level of processing efficiency comparable to the clear processing context.

Together, the behavioral and EWMA/HDDM results provide converging support for the predictions of our information-seeking account. Processing speech in noise caused listeners to seek additional visual information to support language comprehension. Moreover, we observed a strikingly similar pattern of behavior in children and adults, with both groups producing more language-driven shifts and prioritizing accuracy over speed in the more challenging, noisy context.

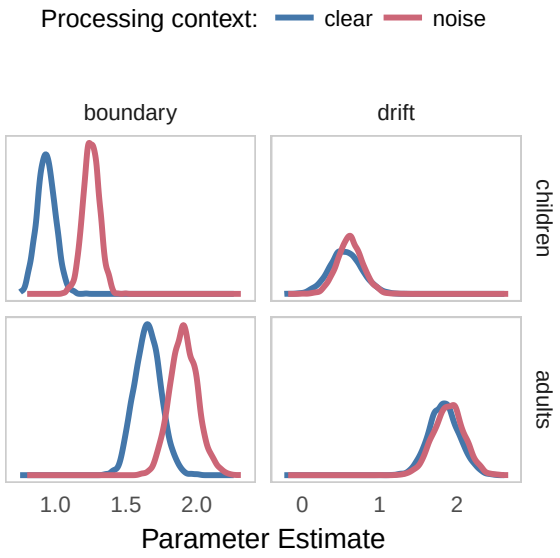

Figure 4: HDDM results. Each panel shows the posterior distribution for either the boundary separation or drift rate parameters for children (top panels) and adults (bottom panels).

Parameter	Condition	Age Group	Estimate [95% HDI]
boundary	clear	adults	1.68 [1.5, 1.87]
		children	0.96 [0.84, 1.09]
	noise	adults	1.65 [1.48, 1.83]
		children	1.23 [1.12, 1.34]
drift	clear	adults	1.86 [1.48, 2.25]
		children	0.58 [0.17, 0.99]
	noise	adults	1.94 [1.56, 2.33]
		children	0.59 [0.27, 0.93]

Table 2: HDDM parameter estimates for each age group and noise condition. The drift rate parameter indexes processing efficiency and the boundary separation parameter indexes participants' information accumulation threshold.

General Discussion

Language comprehension in grounded contexts involves integrating the visual and linguistic signals. But the value of gathering visual information can vary depending on features of the processing context. Here, we presented a test of an information-seeking account of eye movements during language processing – an account that we first proposed in K. MacDonald et al. (2017) to explain population-level differences in the dynamics of gaze between children learning ASL and children learning spoken English. Here, we showed that children and adults adapt to processing speech in noise by producing slower but more accurate gaze shifts away from a

speaker. Both groups also showed evidence of prioritizing information accumulation over speed (HDDM) while producing more language driven shifts (EWMA). It is interesting that listeners were able to achieve higher accuracy in the more challenging, noisy context. Together, the behavioral and modeling results suggest that when the linguistic signal is degraded, listeners adapt their eye movements to seek language-relevant information in the visual world.

These results bring together ideas from several research programs. First, work on language-mediated visual attention shows that adults and children rapidly shift gaze upon hearing the name of an object in the visual scene (Allopenna et al., 1998; Tanenhaus et al., 1995). The speed and consistency of this response has led to debates about whether language-mediated gaze shifts are automatic as opposed to under the control of the listener. While we do claim that listeners in our task have explicit access to the underlying decision process, our findings show that the dynamics of gaze during lexical access adapt to the information features of the context. This finding parallels recent work by McMurray et al. (2017), showing that adults with Cochlear Implants, who consistently process degraded auditory input, will delay the process of lexical access, waiting to begin until substantial information has accumulated.

Second, empirical work on vision during natural tasks shows that people overwhelmingly prefer to look at *goal-relevant* locations – e.g., an upcoming obstacle while walking (Hayhoe & Ballard, 2005). These accounts inspired our prediction that gaze dynamics during language comprehension should adapt to the value of different fixation behaviors with respect to the listener's goal of rapid language processing. And third, work on effortful listening shows that listeners generate compensatory responses (e.g., increases in attention and working memory) within "challenging" comprehension contexts such as processing noisy or accented speech (Van Engen & Peelle, 2014). These accounts predict that our young listeners might compensate for the reduced quality of the auditory signal by allocating gathering additional visual information.

This work has several important limitations that we hope will pave the way for future work. Here, we chose to focus on a single decision about visual fixation to provide a window onto the underlying dynamics of decision-making across different processing contexts. However, the decision to shift away from a language is just one of the many decisions that listeners make while processing language in real-time. Moreover, our analysis does not consider the rich information present in the gaze patterns that occur leading up to this decision. In our future work, we aim to quantify changes in the dynamics of gaze across the full sentence processing context. Finally, we used a simple visual world, with only three places to look, and very simple linguistic stimuli, especially for the adults. Thus it remains an open question how these results might scale up to more realistic language environments.

We designed this experiment to test the generalizability of our information-maximization proposal within the domain of familiar language comprehension. However, we think that the account is more general. And we are interested in applying this framework – in-depth analysis of decisions about visual fixation – to the language acquisition context. Consider that early in language learning children are acquiring novel word-object links while also learning about visual object categories. Both of these tasks produce goals that should in turn modulate children's decisions about where to allocate visual attention, e.g., seeking nonlinguistic cues to reference such as eye gaze and pointing become critical when you are unfamiliar with the information in the linguistic signal. More generally, we think that these results contribute to a recent theoretical emphasis on including goal-based accounts of eye movements during language comprehension (Salverda, Brown, & Tanenhaus, 2011). We hope that our approach presents a way forward for explaining fixation behaviors across a wider variety of populations, processing contexts, and during different stages of language learning.

Acknowledgements

We are grateful to the families who participated in this research. Thanks to Tami Alade and Hannah Slater for help with data collection. This work was supported by an NSF GRFP to KM.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, *12*(2), 423–425.

Gabry, J., & Goodrich, B. (2016). Rstanarm: Bayesian applied regression modeling via stan. r package version 2.10. 0.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Attention, Perception, & Psychophysics*, *24*(3), 253–257.

MacDonald, K., Blonder, A., Marchman, V. and, Fernald, A., & Frank, M. C. (2017). An information-seeking account of eye movements during spoken and signed language comprehension. In *Proceedings of the 39th annual conference of the cognitive science society*.

MacDonald, M. C., & Seidenberg, M. S. (2006). Constraint satisfaction accounts of lexical and sentence comprehension. *Handbook of Psycholinguistics*, *2*, 581–611.

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences*, *10*(8), 363–369.

McMurray, B., Farris-Trimble, A., & Rigler, H. (2017). Waiting for lexical access: Cochlear implants or severely

degraded input lead listeners to process speech less incrementally. *Cognition*, *169*, 147–164.

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision*, *2*(4), 237–279.

Salverda, A. P., Brown, M., & Tanenhaus, M. K. (2011). A goal-based perspective on eye movements in visual world studies. *Acta Psychologica*, *137*(2), 172–180.

Smith, A. C., Monaghan, P., & Huettig, F. (2017). The multimodal nature of spoken word processing in the visual world: Testing the predictions of alternative models of multimodal integration. *Journal of Memory and Language*, *93*, 276–303.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632.

Van Engen, K. J., & Peelle, J. E. (2014). Listening effort and accented speech. *Frontiers in Human Neuroscience*, *8*.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*(6), 1011–1026.

Yurovsky, D., Case, S., & Frank, M. C. (2017). Preschoolers flexibly adapt to linguistic input in a noisy channel. *Psychological Science*, *28*(1), 132–140.