The interaction of image salience with language: how much is language driving eye-movements in Visual World Paradigms?

All visual world paradigms (VWPs) use drawings, cartoons, or photographs as referential context to test language processing mechanisms. While most researchers control for features that might draw attention due to visual features (color, contrast, size, spacing) other factors are less easily controlled (e.g. line orientations within the image). Further, by removing each of these features, experiments become less ecologically valid, destroying the vast variability found in real life that may lead to unexpected interactions. The goal of this work was to quantify how much eye-movement data can be explained by visual salience, and how much can be explained by language in a VWP. We accomplished this using a variant of the VWP and a Receiver Operating Characteristics (ROC) analysis.

Thirty-six participants were given course credit to participate in this study. Three participants' data were discarded due to inaccurate tracks and data loss. After a 9-point calibration procedure, participants were given headphones and told to listen and view images. They were told there may or may not be an auditory story, and it might come prior to or concurrent with the image. A diverse collection of 24 photographs, computer graphics, or illustrations that were different along a variety of dimensions were chosen to ensure the procedure and analysis were robust to image variability.

Each audio clip was a short vignette of approximately the same length describing a third-person experience of a fictional character related to the image (see Figure 1). For the audio-first condition, participants viewed a grey screen while listening to the audio stimulus, after which the image was immediately presented. All images for every condition were shown for 20 seconds, with 5 seconds of a black screen between trials. Trial order was pseudorandomized and item x condition was counterbalanced across three lists. Data were raw eye-movement coordinates sampled at 60Hz with a Tobii X2 remote eye-tracker. While accuracy with this hardware cannot reliably capture the exact pixel at the center of a fixation, using various sizes for margin of error did not affect the pattern of results.

An ROC analysis begins with plotting the true positive rate against the false positive rate. The baseline/ground truth for the ROC analysis utilized algorithmically generated saliency maps (GBVS, Harel, Koch & Perona, 2007), where a continuous-value salience map is binarized at various thresholds from 0-1, creating maps with very few salient pixels to the maximum amount of salient pixels. If the eye gaze fell within range (20 pixels) of the pixel considered to be salient, it increased the true positive rate, while looks outside a salient region decreased the true positive rate. Thus, if all fixations were to salient regions ("perfect classification"), the area under the ROC curve would be 1, the absolute maximum. If fixations were randomly distributed, the ROC curve would be along the line of non-discrimination. For every trial, the ROC curve was computed, and then averaged by condition. The results showed a clear difference between no-audio, audio-first, and audio-concurrent conditions ($F(2,64)=5.5$, $p<.01$, linear trend $F(1,32)=9.2$, $p<.01$). The area under the curve was highest for the audio-concurrent condition, second highest for the audio-first, and lowest for no audio (see Figure 2).

Adaptations of this method and analysis could be utilized for any VWP study. This may allow for a much richer variety of images that would not need to be as constrained for color, contrast, and other salient features. Demonstrating that differences are not due to visual salience alone will be required for more visually complex VWPs. Instead of controlling for visual complexity and diminishing real-world variability, this method can disentangle the two and serve to corroborate and validate any differences found between various linguistic manipulations.

"There have been many different James bond movies over the years. Mary liked the Bond from 1974 the best, the man with the golden gun. She loved the 24k gun icon that represented the movie name. Her second favorite movie is gold finger from 1964."

"The baby chick was born near the ocean and was surrounded by seagulls. It wanted to fly away with the other birds as they flew south. This baby chick is a light yellow color when its born. It gradually changes color to a lighter white as it grows older."

Figure 1 *Example trial* – Each image was used in every condition across participants, i.e. each participant saw each image only one time.

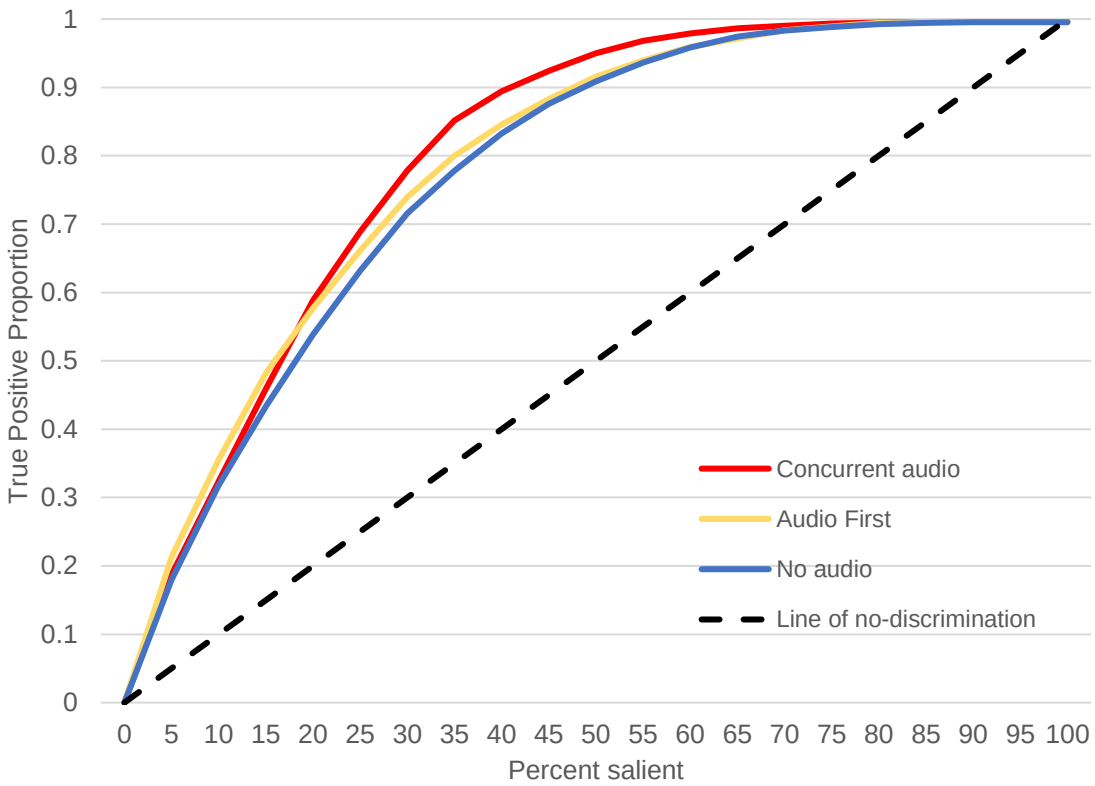

Figure 2 *ROC curves* – The x-axis is the percent salient pixels from the saliency map, and the y-axis is the proportion of eye-movements that hit a region close enough to the salient region.