INDIAN INSTITUTE OF TECHNOLOGY, MADRAS

UNDERGRADUATE RESEARCH PROJECT - ID4900

Analysis of Indian Languages using Eye-Tracking

[Shobhith Vadlamudi - ED21B069]

[Guide: Prof. Anindita Sahoo, *Dept. of Humanities and Social Sciences*]

Contents

1 Objective

In this work, we explore the principle of agent-first in language. It states that the human brain has a tendency to interpret the first argument in a clause as the agent. This default assumption makes passive constructions, which violate this expectation and require syntactic reanalysis, cognitively more demanding. We propose an eye-tracking experiment that is capable of capturing gaze and saccade patterns when a participant is shown an image on the laptop screen. The agent-first bias is treated as an universal principle of sentence processing, by testing it experimentally we can validate a foundational claim about how the human brain processes language. In languages like Punjabi we have explicit grammatical cues like the word "dwara" which typically marks the agent in passive clauses. Unlike English, where passivization is signaled primarily through word order and auxiliary morphology, the overt marking in languages like Punjabi may give us some more insight. By using eye-tracking we are able to capture incremental, moment-to-moment processing, revealing mechanisms that are invisible in accuracy or reaction-time measures making it a suitable medium for our experiment. We wish to test the following hypotheses:

- Whether the agent will be given maximum attention and fixation time irrespective of the sentence construction especially in constructions with dropped agents across English and Punjabi.
- Whether passive constructions show higher cognitive load
- Whether there is a significant difference between the gaze patterns of Punjabi passives and English passives.

1.1 Literature Review

A thorough literature review was conducted on the agent-first bias in language as well as the different methods by which this phenomenon was studied previously. The Agent-First principle is rooted in the cognitive preference for the entity that performs an action. This bias is not merely a linguistic phenomenon but is deeply integrated into event apprehension which is the rapid cognitive process of segmenting and categorizing actions in the real world. (Cohn & Paczynski, 2013)
Research using visual narratives and event descriptions consistently demonstrated an "Agent Advantage",where agents are viewed longer than patients independent of their serial order in a visual sequence,suggesting their central role in event structure and prediction (Cohn & Paczynski, 2013). Furthermore, the earliest stages of visual information processing are susceptible to top-down influences from linguistic preferences. By analyzing landing positions and onset latencies of first fixations in a brief exposure paradigm, studies have shown that the Agent's visual representation is prioritized when comprehending dynamic events (Gerwien & Flecken, 2016). It is also seen that the syntactic structure used during language production or comprehension can influence the visual apprehension of subsequently viewed, unrelated events (Sauppe & Flecken, 2021), this work establishes a direct measurable link between grammatical structures (like passives and actives) and the way visual attention is allocated, providing the necessary rationale for using eye-tracking to test the cognitive salience of the Agent role. In existing literature the Agent Advantage was established but its persistence when the linguistic agent is dropped or unspecified remains unexplored in a cross-linguistic

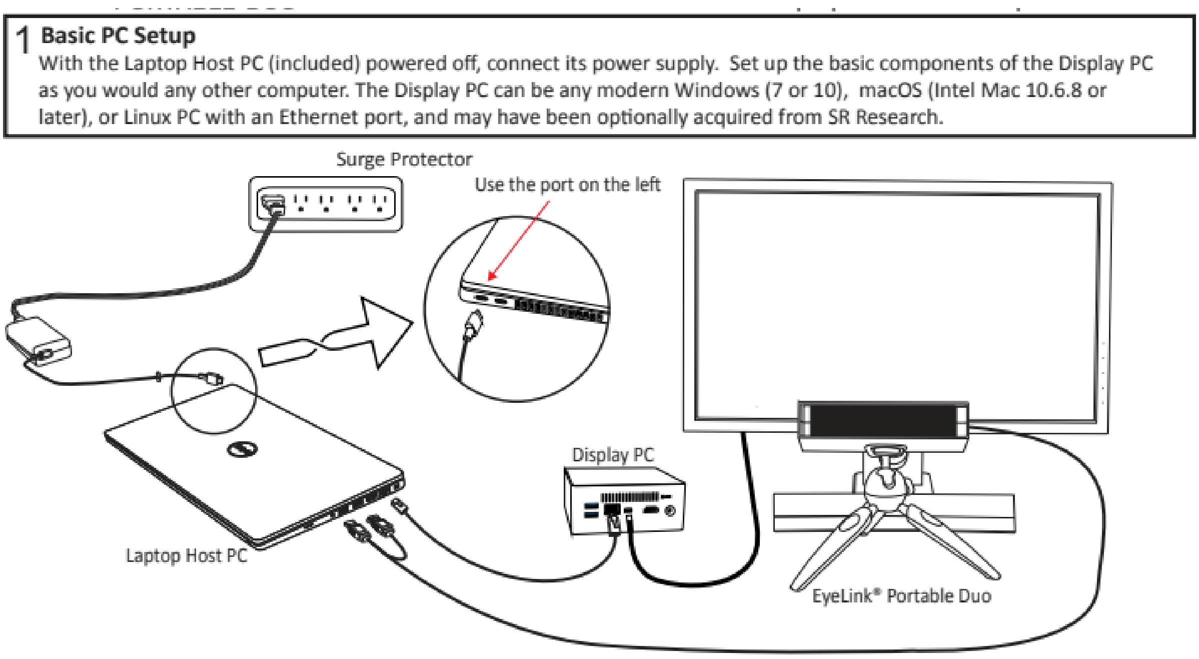
visual-world setting. Most of the foundational work focuses on European languages (English, German, Dutch, Spanish, Basque). The cross-linguistic differences in how agency is coded are known to affect memory for causal agents ((Fausey & Boroditsky, 2008), (Fausey et al., 2010)). Introducing Punjabi, a language that differs starkly in its passive construction, allows for a crucial test of whether the Agent-First bias is truly universal across languages with different canonical word orders and morphological marking systems.

Due to the nature of eye-tracking data, which is sequential, hierarchical and non-normally distributed standard approaches involve powerful statistical models. According to literature, in the visual world paradigm, the most revealing analysis involves plotting the fixation proportion (the probability of looking at the Agent vs. the Patient Area of Interest (AOI)) over time, often in small time bins. The statistical analysis of this time-course data is typically performed using Mixed-Effects Models (specifically, Generalized Linear Mixed Models with a logistic link function) to model the binomial outcome (Fixation on Agent AOI vs. Patient AOI). For aggregate measures (Gaze Duration, Total Fixation Time, Saccade Latency), Linear Mixed-Effects Models are the gold standard. These models account for the nested structure of the data and allow for simultaneous modeling of fixed effects (e.g., Sentence Type: Active vs. Passive, Language: English vs. Punjabi) and random effects (individual variability of participants and items/stimuli) ((Silva et al., 2016), (Hohenstein et al., 2017)). Recently analysis using Bayesian Regression models as an alternative to LMMs, which provide a full probability distribution over the parameter estimates is also seen as in (Isasi-Isasmendi et al., 2023).

2 Eye-Tracking Fundamentals

2.1 Device Specifications and Setup

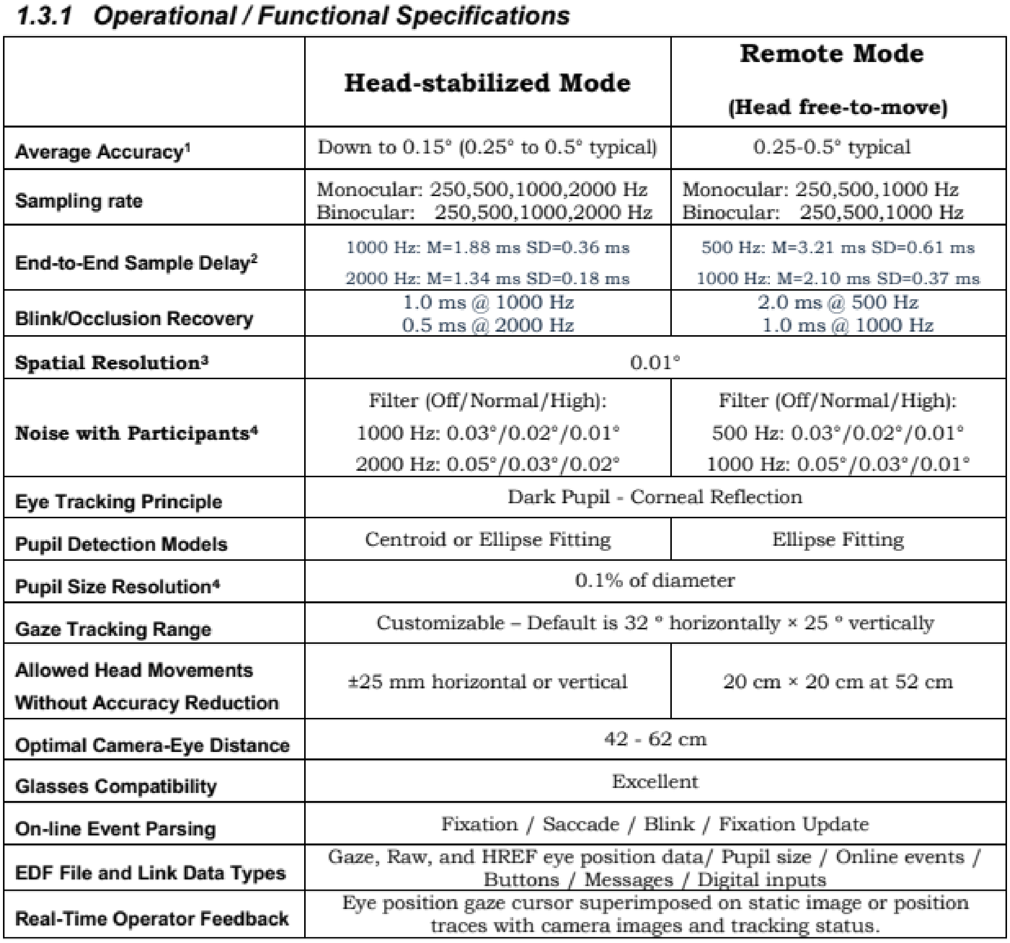
For our experiment we are using an Eye-Link Portable Duo eye tracker. The eye tracker illuminates the participant's face and eyes with invisible near-infrared light. A camera then acquires upto 2000 images per second which are processed by the attached host PC to identify landmarks on the eyes and face, allowing measurement of the participant's gaze location. It has a built in capability to be used as a high accuracy head stabilized eye tracker or as a head free to move remote eye tracker. For our experiment we used a chin rest attached to the table to stabilize the participant's head but allow for small motions. The remote mode of the eye-tracker is then used. A typical system consists of two computers- one is the host PC which connects to the eye-tracker and is dedicated to data collection. The second computer is called the display PC and it is used for presentation of stimuli to the participant. The two computers are connected via an ethernet connection. This connection allows the display PC to have access to the eye data such as ongoing gaze position and saccade and fixation events. It also allows the transfer of camera images from the host PC to the display PC. The eye tracker device will be connected to the host PC via USB. It is very important to ensure a power supply is connected to both systems before beginning the experiment. The eye tracker unit consists of a camera and an infrared illuminator.



1 **Basic PC Setup**

With the Laptop Host PC (included) powered off, connect its power supply. Set up the basic components of the Display PC as you would any other computer. The Display PC can be any modern Windows (7 or 10), macOS (Intel Mac 10.6.8 or later), or Linux PC with an Ethernet port, and may have been optionally acquired from SR Research.

The specifications of the eye-tracker are given below:



1.3.1 Operational / Functional Specifications

	Head-stabilized Mode	Remote Mode (Head free-to-move)
Average Accuracy[1]	Down to 0.15° (0.25° to 0.5° typical)	0.25-0.5° typical
Sampling rate	Monocular: 250,500,1000,2000 Hz Binocular: 250,500,1000,2000 Hz	Monocular: 250,500,1000 Hz Binocular: 250,500,1000 Hz
End-to-End Sample Delay[2]	1000 Hz: M=1.88 ms SD=0.36 ms 2000 Hz: M=1.34 ms SD=0.18 ms	500 Hz: M=3.21 ms SD=0.61 ms 1000 Hz: M=2.10 ms SD=0.37 ms
Blink/Occlusion Recovery	1.0 ms @ 1000 Hz 0.5 ms @ 2000 Hz	2.0 ms @ 500 Hz 1.0 ms @ 1000 Hz
Spatial Resolution[3]	0.01°	
Noise with Participants[4]	Filter (Off/Normal/High): 1000 Hz: 0.03°/0.02°/0.01° 2000 Hz: 0.05°/0.03°/0.02°	Filter (Off/Normal/High): 500 Hz: 0.03°/0.02°/0.01° 1000 Hz: 0.05°/0.03°/0.01°
Eye Tracking Principle	Dark Pupil - Corneal Reflection	
Pupil Detection Models	Centroid or Ellipse Fitting	Ellipse Fitting
Pupil Size Resolution[4]	0.1% of diameter	
Gaze Tracking Range	Customizable – Default is 32 ° horizontally × 25 ° vertically	
Allowed Head Movements Without Accuracy Reduction	±25 mm horizontal or vertical	20 cm × 20 cm at 52 cm
Optimal Camera-Eye Distance	42 - 62 cm	
Glasses Compatibility	Excellent	
On-line Event Parsing	Fixation / Saccade / Blink / Fixation Update	
EDF File and Link Data Types	Gaze, Raw, and HREF eye position data/ Pupil size / Online events / Buttons / Messages / Digital inputs	
Real-Time Operator Feedback	Eye position gaze cursor superimposed on static image or position traces with camera images and tracking status.	

We are using binocular mode with 500 Hz sampling frequency and remote operating mode for our experiment. The calibration mode has been set to a default 3x3 grid and is on auto mode.

2.2 How does it work?

Modern eye-trackers like this one work using a video-based infrared tracking system. The tracker illuminates the participant's eyes with invisible near-infrared (IR) light and records high-speed images of the eyes using an internal camera. When infrared light is directed at the eye, two stable visual features are produced:

- The pupil, which appears dark under IR illumination and
- The corneal reflection (CR), a small bright reflection created when IR light reflects off the curved surface of the cornea.

These features are reliably detectable across lighting conditions and do not interfere with vision. For each video frame, the system computes the relative position of the pupil center with respect to the corneal reflection. Because the corneal reflection remains relatively stable while the pupil moves with eye rotation, the vector between these two points provides a robust estimate of eye orientation.Using a prior calibration procedure, the tracker maps this eye orientation onto screen coordinates, allowing it to infer where on the screen the participant is looking at any given moment. The tracker samples eye position at rates ranging from 500-2000 Hz, meaning it records eye-position every 1-2 milliseconds. This allows for the precise detection of

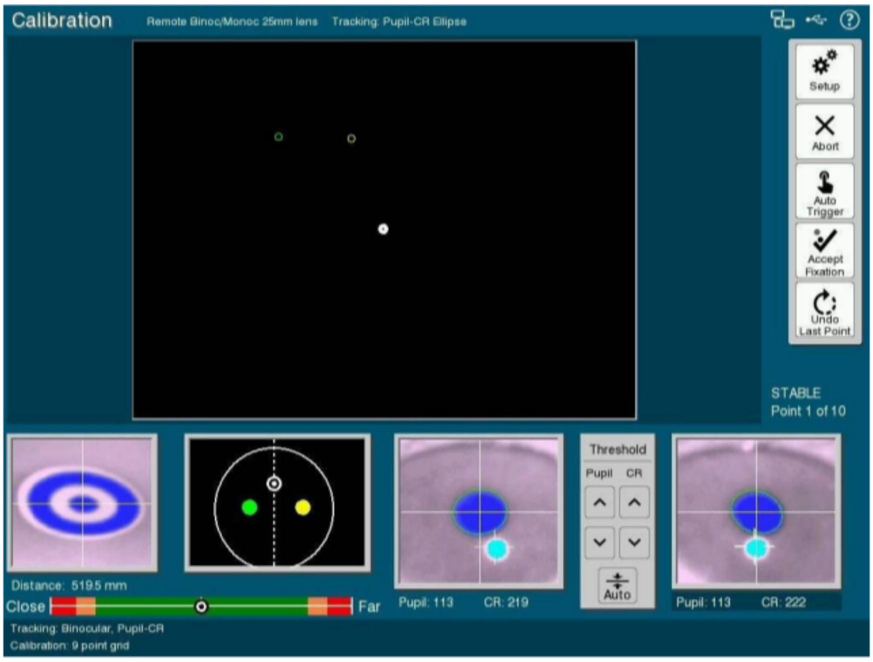
- Fixations (periods of relatively stable gaze)
- Saccades (rapid eye movements between fixations)
- Blinks

All of these events along with continuous gaze position and pupil size are stored in dedicated data files EDF format for later analysis. The eye-tracker system implements two main pupil tracking algorithms - Centroid fitting and Ellipse fitting. The centroid mode tracks the center of the thresholded pupil using a center-of-mass algorithm, whereas the ellipse mode determines the center of the pupil by fitting an ellipse based on the thresholded pupil mass. In head-stabilized mode the centroid algorithm is used by default, this has very mow noise. The remote tracking uses the ellipse-fitting pupil tracking method.

2.3 Calibration

calibration is a crucial step that establishes the mapping between the participant's eye position and the location on the screen. It is necessary to do this for each participant before the start of the experiment. During calibration the participants will be asked to fixate on a sequence of predefined targets presented at known locations on the screen. The calibration pattern can be chosen by us and we have gone with the basic 3x3 grid. The system records the corresponding pupil and corneal reflection positions for each target and uses this data to compute a transformation function that maps eye orientation to screen coordinates. If automatic sequencing has been enabled, targets will be presented and fixations collected without further intervention. Each time a new target is displayed, the participant should quickly make a saccade to it. The EyeLink system detects these saccades and the fixation following, producing an automated sequencing system.

2.4.2 Calibration Screen


Calibration Remote Binoc/Monoc 25mm lens Tracking: Pupil-CR Ellipse

STABLE
Point 1 of 10

Distance: 519.5 mm

Pupil: 113 CR: 219

Pupil: 113 CR: 222

Tracking: Binocular, Pupil-CR
Calibration: 9 point grid

Figure 2-4: Example Calibration Screen

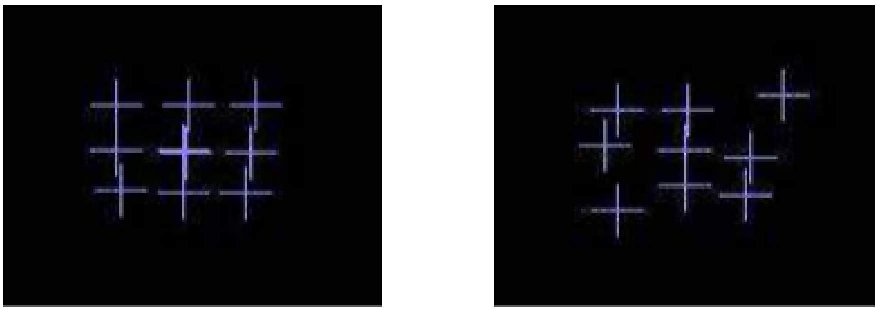

Good Calibration Poor Calibration

Figure 3-10: Calibration Grid

It is important that the result of the calibration should be a good grid structure, any other structures would point to lapses in the participant's attention, and most camera set-up problems. For some participants short fixations or lapses of attention make the automated procedure unusable, it is important to switch to the manual calibration process in this case. After calibration a message will be shown about the nature of the calibration either GOOD or FAILED.

After successful calibration we do a validation process where in targets are again presented to the participant in a random order, similar to the calibration procedure. When the participant fixates on these targets, the calibration model is used to estimate the gaze position of the participant, and the error (difference between actual target position and computed gaze position) is calculated. One of the following results is displayed depending on the outcome: GOOD (green background): Errors are generally acceptable. FAIR (grey background): Errors are moderate, calibration should be improved. POOR: (red background): Errors are too high for useful eye tracking.

3 Experiment Design

We had forty-eight English-Punjabi bilingual participants between the ages of eighteen and forty who participated in the study. All participants reported normal or corrected-to-normal vision and no history of neurological or language-related disorders. Informed consent was obtained prior to participation.
Stimuli consisted of simple, black-and-white drawings depicting transitive events involving two entities, one as the subject and the other as the object. Each image was paired with a sentence describing the depicted event. Images were designed to be visually balanced (equal area for the subject and the object) and as simple as possible to remove all distractions. It was mandatory that neither the subject or the object was visually salient prior to linguistic input. An example image is given below.

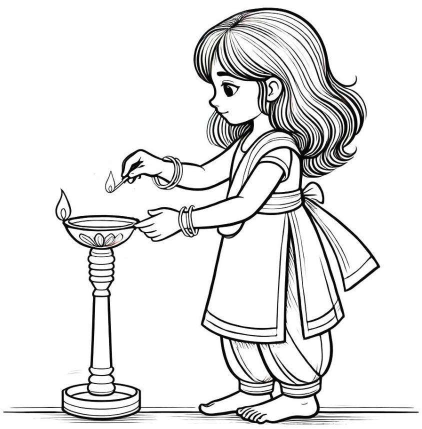


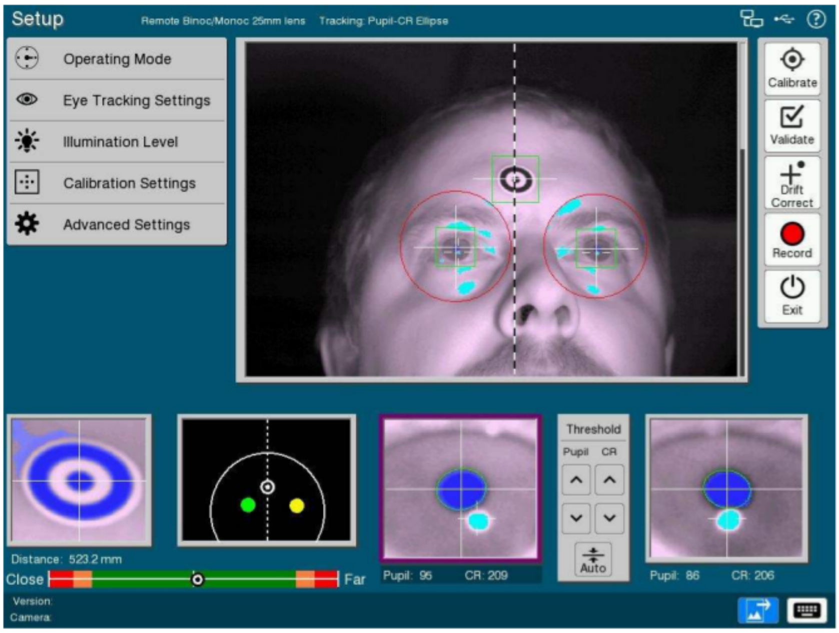
Figure 1: Girl lighting a lamp

The images were generated with careful prompting using Gemini AI. The Punjabi sentences were transcribed using a native speaker and a text-to-voice app was used to generate the audio both in Punjabi and English. This was done to ensure uniformity and as little bias as possible during the experiment design. The sentences were split into six different categories - Punjabi active, Punjabi passive, English active, English passive, Punjabi passive with dropped agent, English passive with dropped agent. The experiment was split into two sessions, session one was free viewing where the participants were shown plain images without any sentence context. This was done to establish a benchmark of their viewing pattern. In session two they were shown eighteen images with sentences and audio being played prior to the image being shown. The stimuli were randomized to ensure that the same images weren't shown in both the sessions.

4 Procedure

A major part of my project was conducting the experiment for 48 participants. Here I will mention the procedure followed for each participant to ensure uniformity.

- First the chin rest and the experiment space should be cleaned in the presence of the participant to make sure that they are not worried about the hygiene.
- I give the participants the consent form and ask them to sign it after reading it.
- Next I show them a presentation explaining about what they would need to do during the experiment. It is very important not to disclose any details about the purpose of the experiment as that would bias their viewing patterns. It is necessary to just tell them enough information so that they can complete the experiment. It is good to mention how long the experiment will take and to remove all sources of distractions such as a phone for the duration of the experiment. It is important to communicate to them that they are allowed to blink normally but shouldn't move their head too much. They are given time to adjust their chair height.
- One of the parameters recorded by the eye-tracker is eye dominance (left or right), to measure this I do a dominant eye test for the participant. I found this video to be helpful Eye dominance. The participant is also required to put a sticker on their forehead for remote-tracking mode.



Remote Binoc/Monoc 25mm lens Tracking: Pupil-CR Ellipse

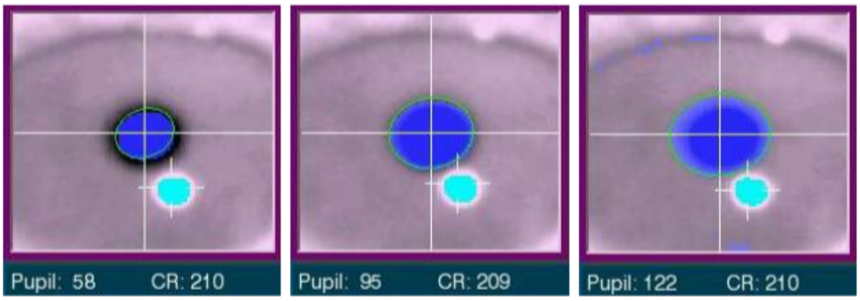
Distance: 523.2 mm

Pupil: 96 CR: 209

Pupil: 86 CR: 206

Figure 3-4: Setup Screen in the Remote Mode

- The eye tracker is then set up and the thresholds are adjusted for the participant. It is important to check the distance of the target from the tracker and ensure the angle is correct as well. The participant's sitting posture should be comfortable and should be able to see the whole screen. It is important to adjust the focus of the eye tracker before going for calibration. There is a small wheel at the bottom of the tracker that is used to adjust this.



Threshold too low Properly thresholded Threshold too high

- Next calibration and validation procedures are carried out as explained before. It is important to save the calibration results by pressing ENTER before starting validation. For participants with glasses it was helpful to manually reduce the size of the measuring area to prevent glare. This can be done using ALT and cursor keys on the keyboard.
- Now the experiment is started and they are asked to view the stimulus being presented. A brief gap is given between session one and session two.
- After successful completion of the experiment I briefly check the data and ensure its saved and then have the participant fill out a post-experiment survey. Now I explain the motivation behind the experiment and ask them what they think about it.

5 Data Analysis

The data analysis was done with the help of the Eyelink DataViewer application. The files were loaded and cleaned. Next the areas of interest (subject and object) for all the images were marked. Using the DataViewer gaze, saccade and fixation reports were generated and then further analysis was done using Python and R.

5.1 GAMM Models

Generalized Additive Mixed Models (GAMMs) are an extension of linear mixed-effects models that allow non-linear relationships between predictors and response variables while simultaneously accounting for random effects due to subjects and items. Formally, GAMMs model the response variable as the sum of:

- Parametric terms (e.g., experimental conditions such as language or voice)
- Smooth functions of continuous predictors (e.g., time, word position, trial progression)
- Random effects capturing subject- and item-level variability

This flexibility makes GAMMs particularly well-suited for time-resolved behavioural data such as eye movements. We cannot use traditional linear models here as

- Eye movements are inherently non-linear. We see rapid early effects, delayed divergences and plateau or recovery phases in gaze behaviour.
- Eye-tracking data points are not independent as consecutive samples within a trial are correlated, and fixation patterns depend on earlier processing stages.

Let y_i denote the response variable for observation i (e.g., fixation probability or dwell time). GAMMs model the expected value of y_i using a link function $g(\cdot)$ as follows:

$$g(\mathbb{E}[y_i]) = \beta_0 + \sum_p \beta_p x_{ip} + \sum_j f_j(z_{ij}) + \mathbf{Z}_i\mathbf{b}, \tag{1}$$

where β_0 is the intercept, β_p are coefficients associated with parametric predictors (such as Language and Voice), $f_j(\cdot)$ are smooth functions of continuous predictors (such as time), and $\mathbf{Z}_i\mathbf{b}$ represents random effects associated with participants and items.
Each smooth function $f(x)$ is represented using a basis expansion:

$$f(x) = \sum_{k=1}^{K} \alpha_k \phi_k(x), \tag{2}$$

where $\phi_k(x)$ are basis functions (typically thin-plate regression splines), α_k are coefficients to be estimated from the data, and K determines the maximum flexibility of the smooth. This formulation allows the model to capture complex, non-linear relationships without imposing a fixed functional form.
To prevent overfitting, GAMMs include a smoothness penalty on the curvature of the function:

$$\lambda \int (f''(x))^2 \, dx, \tag{3}$$

where λ is a smoothing parameter controlling the trade-off between model fit and smoothness. Larger values of λ enforce smoother functions, while smaller values allow greater flexibility. The optimal value of λ is estimated from the data using restricted maximum likelihood (REML).
Random effects are incorporated to account for systematic variability across participants and items. Random intercepts are assumed to follow a normal distribution:

$$b_s \sim \mathcal{N}(0, \sigma_s^2). \tag{4}$$

In addition to random intercepts, GAMMs allow the inclusion of random smooths, which permit individual-specific temporal trajectories:

$$f_s(t) = \sum_k \alpha_{sk} \phi_k(t), \quad \alpha_{sk} \sim \mathcal{N}(0, \sigma_k^2). \tag{5}$$

6 Results

First, the stimuli images were analyzed and bounding boxes were drawn for the subject and the object marking the AOIs. Then reports were generated using the EyeLink Data Viewer software for fixation, saccade and interest areas. Across all 48 participants, the experimental design included a consistent total of 36 trials per participant, comprising 18 free-viewing trials and 18 voice trials. The voice condition was fully balanced, with each participant completing 6 Active (ACT), 6 Passive No Agent (PNA), and 6 Passive With Agent (PWA) trials (33.3% each). Language distribution within voice trials was also uniform, with 9 English and 9 Punjabi trials per participant. Language was not coded for free-viewing trials.

On analyzing the first fixations of each image for all the participants, it was found that about 90% of them were in the center of the screen. This showed a clear bias of the participants to look at the center of the screen first. These fixations were removed and the first non-center fixation was considered. The results of the first fixations organized by voice type are given below.

Table 1: First Non-Center Fixation Location by Voice Type

Voice Type	Total	Subject	Object	Subject %	Object %
ACT	257	114	143	44.4%	55.6%
PNA	261	107	154	41.0%	59.0%
PWA	264	111	153	42.0%	58.0%
Overall	**782**	**332**	**450**	**42.5%**	**57.5%**

Across all three voice types, the distribution of referential roles was highly consistent, with approximately 58% of responses referring to the object and 42% referring to the subject. A chi-square test confirmed that these proportions did not differ significantly across conditions ($\chi^2(2) = 0.62$, $p = 0.731$), and the effect size was negligible (Cramér's $V = 0.028$). Follow-up pairwise comparisons (ACT vs. PNA, ACT vs. PWA, and PNA vs. PWA) also revealed no significant differences, indicating that voice type did not meaningfully influence whether the first fixation landed on the subject or the object.

6.1 Total Fixations and Viewing Time

Although the first fixation showed a slight preference for objects, the overall distribution of fixations across trials revealed a different pattern. When examining total fixations throughout the viewing period, subjects received more attention than objects. Table 2 presents the total fixation counts by AOI type and voice condition.

Table 2: Total Fixations by AOI Type and Voice Condition

Voice Type	Subject	Object	Subject %	Object %
ACT	2,286	1,906	54.5%	45.5%
PNA	2,074	1,909	52.1%	47.9%
PWA	2,084	1,891	52.4%	47.6%
Overall	**6,444**	**5,706**	**53.0%**	**47.0%**

The overall pattern showed that subjects received 53.0% of total fixations compared to 47.0% for objects, with an average of 7.42 fixations per trial for subjects versus 6.64 for objects. This reversal from the first fixation pattern suggests a two-stage attention process: initial attention capture by objects, followed by sustained attention to subjects.

6.2 Fixation Duration and Viewing Time

To examine the temporal dynamics of attention allocation, we analyzed total fixation time (dwell time) on each AOI type. Table 3 presents the mean total fixation time by AOI type and voice condition.

A Kruskal-Wallis test revealed no significant differences in total fixation time across voice types for subject AOIs ($H(2) = 1.35$, $p = 0.509$) or object AOIs ($H(2) = 1.24$,

Table 3: Total Fixation Time (ms) by AOI Type and Voice Condition

Voice Type	Subject (ms)	Object (ms)
ACT	$M = 1,845.2$	$M = 2,089.3$
PNA	$M = 1,912.1$	$M = 2,178.4$
PWA	$M = 1,900.1$	$M = 2,191.5$
Overall	$M = 1,885.8$	$M = 2,153.1$

$p = 0.538$). When examining the proportion of viewing time allocated to each AOI, subjects received 46.9% of total viewing time compared to 53.1% for objects, which was statistically significant (Mann-Whitney U test, $p < 0.001$). This pattern suggests that while subjects receive more fixations, objects receive longer individual fixations and greater total viewing time, and this pattern is consistent across all voice types.

6.3 AOI Switching Patterns

To understand the dynamics of attention shifts, we analyzed the number of switches between subject and object AOIs. Table 4 presents the switching patterns by voice type.

Table 4: AOI Switching Patterns by Voice Type

Voice Type	n Trials	Total Switches	M Switches/Trial	Subj→Obj	Obj→Subj
ACT	286	1,022	3.57	507	515
PNA	285	946	3.32	465	481
PWA	283	955	3.37	472	483
Overall	**854**	**2,923**	**3.42**	**1,444**	**1,479**

The switching patterns were highly balanced, with nearly equal numbers of subject-to-object and object-to-subject transitions across all voice types. A Kruskal-Wallis test revealed no significant differences in switch counts across voice types ($H(2) = 2.18$, $p = 0.336$), indicating that voice type did not affect the frequency or direction of attention switches.

6.4 Free Viewing vs. Voice Condition Comparison

To examine the effect of voice presentation on attention patterns, we compared the free viewing (baseline) condition with the voice condition. Table 5 presents fixation duration comparisons between conditions.

Table 5: Fixation Duration (ms) Comparison: Free Viewing vs. Voice Condition

AOI Type	Free Viewing	Voice	Difference	p-value	Cohen's d
Subject	$M = 349.9$ $SD = 351.5$	$M = 360.9$ $SD = 345.1$	+11.0 ms (3.1%)	< 0.01**	0.032
Object	$M = 333.4$ $SD = 323.5$	$M = 356.0$ $SD = 371.9$	+22.6 ms (6.8%)	< 0.01**	0.065

The voice condition showed significantly longer fixations on both subject and object AOIs compared to free viewing, though effect sizes were small (Cohen's $d < 0.1$). The effect was larger for objects (6.8% increase) than for subjects (3.1% increase), suggesting that voice guidance particularly enhances attention to object regions.

6.5 Language Effects

To examine whether language (English vs. Punjabi) affected attention patterns, we compared eye-tracking measures across languages. Table 6 presents the first fixation location by language.

Table 6: First Fixation Location by Language

Language	Subject	Object	Subject %	Object %
English	156	158	49.7%	50.3%
Punjabi	159	158	50.2%	49.8%

A chi-square test revealed no significant association between language and first fixation location ($\chi^2(1) = 0.0016$, $p = 0.968$, Cramér's $V = 0.0016$). Similarly, no significant differences were found for total fixation distribution ($\chi^2(1) = 2.45$, $p = 0.117$), fixation duration (Subject: $p = 0.129$, Object: $p = 0.052$), or proportion of viewing time ($p = 0.131$). These findings indicate that attention patterns were consistent across languages, suggesting that the observed effects are driven by voice type and condition rather than language.

6.6 Generalized Additive Mixed Model (GAMM) Analysis

To account for temporal dynamics and participant-level variation, we fitted a Generalized Additive Mixed Model (GAMM) with polynomial time terms and random participant effects. The model included fixed effects for AOI type (subject vs. object), voice type (ACT, PNA, PWA), and time from trial start (linear, quadratic, and cubic terms), with random intercepts for participants.

The polynomial mixed model (Model 3) provided the best fit, accounting for non-linear temporal effects. Key findings from the model are presented in Table 7.

Table 7: GAMM Results: Fixed Effects for Fixation Duration

Predictor	Coefficient	SE	z	p-value
Intercept	276.44	12.09	22.87	< 0.001***
AOI (Subject)	−7.20	4.79	−1.50	0.133
Voice (ACT)	−14.19	5.79	−2.45	0.014*
Voice (PNA)	−6.84	5.86	−1.17	0.243
Time (linear)	71.39	13.27	5.38	< 0.001***
Time (quadratic)	1.29	6.80	0.19	0.849
Time (cubic)	−3.67	0.95	−3.88	< 0.001***

Note: Reference categories are Object AOI and PWA voice type. $n = 11,086$ fixations, 49 participants. Random participant variance = 4,245.74.

The model revealed several key findings: (1) Fixation duration increased significantly over time (linear term: $\beta = 71.39$ ms/second, $p < 0.001$), with a significant non-linear component (cubic term: $\beta = -3.67$, $p < 0.001$), indicating that fixations became longer as trials progressed but with a decelerating trend; (2) ACT voice type showed significantly shorter fixations than PWA ($\beta = -14.19$ ms, $p = 0.014$), even after controlling for temporal effects; (3) The AOI effect (subject vs. object) was not significant when controlling for time, suggesting that the observed differences in first fixation patterns do not persist throughout the trial; (4) Participant-level random effects accounted for substantial variance (4,245.74), indicating meaningful individual differences in fixation patterns.

6.7 Summary of Key Findings

The results reveal several important patterns in attention allocation during image viewing:

1. **First fixation bias**: Objects are slightly preferred as the first non-center fixation target (57.5% vs. 42.5%), but this preference does not vary by voice type.
2. **Sustained attention to subjects**: Despite initial object preference, subjects receive more total fixations (53.0% vs. 47.0%) and more fixations per trial (7.42 vs. 6.64), suggesting a shift in attention allocation over time.
3. **Objects receive more viewing time**: Objects receive a greater proportion of total viewing time (53.1% vs. 46.9%), indicating longer individual fixations on objects despite fewer total fixations.
4. **Voice type effects**: ACT voice type is associated with shorter fixation durations compared to PWA, but voice type does not significantly affect first fixation location, total fixations, or switching patterns.
5. **Temporal dynamics**: Fixation duration increases over the course of trials, with a non-linear decelerating trend, suggesting that attention becomes more focused and sustained as viewing progresses.
6. **No language effects**: Attention patterns are consistent across English and Punjabi, indicating that the observed effects are driven by visual and voice characteristics rather than linguistic factors.
7. **Balanced switching**: Attention switches between subject and object AOIs are highly balanced, with nearly equal transitions in both directions, suggesting active comparison of referential roles throughout viewing.

These findings support a two-stage attention model: initial attention capture by objects, followed by sustained attention allocation to subjects, with voice presentation enhancing overall fixation duration but not fundamentally altering the distribution of attention between referential roles.

References

Cohn, N., & Paczynski, M. (2013). Prediction, events, and the advantage of agents: The processing of semantic roles in visual narrative. **Cognitive Psychology**, **67**(3), 73–101. https://doi.org/10.1016/j.cogpsych.2013.07.002

Fausey, C. M., & Boroditsky, L. (2008). English and spanish speakers remember causal agents differently. **Proceedings of the Annual Meeting of the Cognitive Science Society**, **30**.

Fausey, C. M., Long, B. L., Inamori, A., & Boroditsky, L. (2010). Constructing agency: The role of language. **Frontiers in Psychology**, **1**. https://doi.org/10.3389/fpsyg.2010.00162

Gerwien, J., & Flecken, M. (2016). First things first? top-down influences on event apprehension. **Proceedings of the Annual Meeting of the Cognitive Science Society**, **38**(0). https://escholarship.org/uc/item/2871b1gc

Hohenstein, S., Matuschek, H., & Kliegl, R. (2017). Linked linear mixed models: A joint analysis of fixation locations and fixation durations in natural reading. **Psychonomic Bulletin & Review**, **24**(3), 637–651. https://doi.org/10.3758/s13423-016-1138-y

Isasi-Isasmendi, A., Andrews, C., Flecken, M., Laka, I., Daum, M. M., Meyer, M., Bickel, B., & Sauppe, S. (2023). The agent preference in visual event apprehension. **Open Mind: Discoveries in Cognitive Science**, **7**, 240–282. https://doi.org/10.1162/opmi_a_00083

Sauppe, S., & Flecken, M. (2021). Speaking for seeing: Sentence structure guides visual event apprehension. **Cognition**, **206**, 104516. https://doi.org/https://doi.org/10.1016/j.cognition.2020.104516

Silva, B. B., Orrego-Carmona, D., & Szarkowska, A. (2016). Using linear mixed models to analyze data from eye-tracking research on subtitling. **Journal of Eye Movement Research**, **9**(5), 1–15.