A Novel Use of Eye-Tracking to Measure Real-Time Speech Recognition

Senior Thesis

Presented to

The Faculty of the School of Arts and Sciences
Brandeis University

Undergraduate Program in Neuroscience
Dr. Arthur Wingfield

In partial fulfillment of the requirements for the degree of Bachelor of Science

by

Boris Osipov

April 2013

Copyright by
Boris Osipov

Committee members (if applicable):

Name: ________________________  Signature: ________________________

Name: ________________________  Signature: ________________________

Name: ________________________  Signature: ________________________
ABSTRACT

With this study I wanted to see if eye-tracking could be used as an effective online measure of speech comprehension in the processing of sentences. Two experimental levels of syntactic complexity were used in this paradigm: subject relative embedded clauses and objective relative embedded clauses. English-speaking young adults listened to simple narrative sentences while looking at a visual display containing potential referents within the sentence and were asked to determine the agent of the action. The goal was to see if eye-gaze to the correct agent would signify the moment at which participants comprehended the sentences. This experiment was able to define a temporal window during which participants resolved syntactic ambiguity in real-time. The key behavioral finding showed that the probability of saccading to the target agent was considerably higher for subject-relative than object-relative sentences indicating a difference in sentence processing due to syntactic complexity.

INTRODUCTION

Spoken language is very rapid, consisting of 140-180 words per minute, and is often under-articulated. When listening to speech, we are actively engaged in decoding rapidly changing acoustic patterns to recognize words, parse out sentences, and process associations and semantics. Yet despite speech being grammatically complex and often not grammatical at all, human listeners can still understand speech efficiently. Understanding how spoken language comprehension proceeds over time has been central to the development of theories of human language use. In this study, we examine how comprehension for speech unfolds over time.
Online vs. Offline Tests of Comprehension

Most studies of speech comprehension have utilized off-line behavioral measures, requiring participants to recall or paraphrase information after hearing it. In such studies, a participant hears a sentence—for example, “The eye, that was observed by the keen fox, is watchful”—and is then asked, “Who is the agent of the action?” In this case, who did the observing? The identification of the agent of the action serves as a measure of comprehension. In these studies, the effects of syntactic structure on speech perception can be examined by varying syntactic complexity.

An example of a subject relative central embedded clause that was used in this experiment is, “The eye, who observed the keen fox, is watchful,” and an object relative central embedded clause would read, “The eye, who was observed by the keen fox, is watchful.” Numerous studies have demonstrated that object relative sentences have heavier processing loads than do subject relative sentences (Betancort, Carreiras, & Sturt, 2009). This seems to be the case because the thematic roles in object relative sentences are not canonical and require considerable thematic integration (Gibson, 1998; Warren & Gibson, 2002). Furthermore, one must maintain the subject of the sentence for a longer time in object-relative than in subject-relative sentences to accurately comprehend these thematic roles (Cooke et al., 2001). As a result, it's been inferred that processing object-relative sentences is more resource demanding than processing subject-relative sentences (Ferreira, Henderson et al., 1996; Cooke et al.). This has been supported by functional imaging studies showing increased patterns of neural activity for processing of object-relative sentences (Cooke et al.; Just et al., 1996; Peelle, McMillan, Moore, Grossman, & Wingfield, 2004), and that object-relative sentences produce more comprehension errors (Just & Carpenter, 1992; Wingfield et al., 2003). The choice of object relative sentences and subject relative sentences, was based on our desire to work with sentences that differed in known processing difficulty. Although these offline measures reveal how difficult or confusing a sentence is to understand, they do not tell us about the real-time nature of speech recognition. I therefore incorporated
these traditional methods with eye-tracking, a potential online measure of the time course of sentence processing.

**Testing Sentence Comprehension Using Eye-tracking**

Eye-tracking uses a camera to locate the gaze-point of a participant on the screen. The use of this methodology as a means of measuring comprehension has had some limited use in studying speech perception (Tanenhaus et al., 1995; Boaz et al., 2011). My question is whether it will be sensitive to differences in syntactic complexity arising from the embedding of subject relative and object relative clauses. The goal of this project is to serve as a proof of concept, to determine whether eye-tracking can be used as an effective online measure of speech comprehension in the processing of sentences. Addressing the time course of sentence processing requires using methodologies that can provide immediate information regarding how each word is interpreted as the sentence unfolds. Unlike a voluntary motor response which can be on the order of seconds, eye gaze is very rapid and can be directed on the order of milliseconds (Ferreira, Engelhardt, & Jones, 2009). In this thesis I use an eye-tracking monitoring task where the participant is allowed to process the input of language naturally. Potentially, eye-gaze on the picture of the correct agent will tell us the moment at which the actor of the sentence was recognized and when the sentence was fully comprehended.

**METHOD**

**Participants**

Sixteen younger adults who were all native English speakers participated in the study. The younger adults were undergraduate volunteers from Brandeis University and were paid $10/hour. Participants were provided with written informed consent and took part in a one-study visit.
Materials

Sentence stimuli.
Sixty grammatically simple sentences were constructed that contained two nouns, two adjectives, and one action verb. An example of such a sentence was, “The watchful eye observed the keen fox.” These sentences served as the basis for the design of sentence stimuli varying in syntactic complexity. We then modified each of these simple sentences to assemble two versions with subject relative central embedded clauses and two versions with object relative central embedded clauses. An example of a subject-relative sentence is, “The eye, that observed the keen fox, is watchful.” Here, the subject of the sentence is performing the action, the eye. An example of an object-relative sentence is, “The eye, that was observed by the keen fox, is watchful.” In addition, filler sentences were created which were neither subject relative nor object relative and used a different structure. Table 1 shows examples of all sentence subtypes. There was no recycling of nouns, adjectives, or verbs, amongst any of the 60 base sentences as to avoid redundancy of objects. This design resulted in 240 unique experimental sentences and 180 filler sentence that were counterbalanced in seven versions of the sentence sets.

<table>
<thead>
<tr>
<th>Sentence Type</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject Relative</td>
<td>The eye, who observed the keen fox, is watchful.</td>
</tr>
<tr>
<td>Subject Relative</td>
<td>The fox, who observed the keen eye, is watchful.</td>
</tr>
<tr>
<td>Object Relative</td>
<td>The eye, who was observed by the keen fox, is watchful.</td>
</tr>
<tr>
<td>Object Relative</td>
<td>The fox, who was observed by the keen eye, is watchful.</td>
</tr>
<tr>
<td>Filler</td>
<td>The watchful eye observed the keen fox.</td>
</tr>
<tr>
<td>Filler</td>
<td>The fox, who is watchful, observed the keen eye.</td>
</tr>
<tr>
<td>Filler</td>
<td>The watchful fox observed the eye, who is keen.</td>
</tr>
</tbody>
</table>

Table 1. Syntactic Manipulation of Sentence Structure.
Example of seven sentence subtypes including subject relative, object relative, and filler sentences.

Filler sentences were used to minimize the creation of a response strategy. These sentences were presented alternatingly with subject relative and objective relative versions so that experimental sentences would never be heard back-to-back. The auditory stimuli were recorded by a female speaker with a natural speech rate unto computer sound files.
Eye-tracking apparatus.
Eye-movements were recorded using a table-mounted head-free eye tracking apparatus (ASL Eye-Trac 6, D6 Desk Mounted Optics, Manual Version 3.00) and GazeTracker software. GazeTracker is a multipurpose tool that allows for the collection, analysis and resolution of recorded eye-gaze and computer data. This model records the movements and positions of participants' right eye with a camera and a non-invasive beam of infrared light. The camera takes pictures of the eye 60 times a second (approximately every 17 ms). Each picture is analyzed to find the pupil and corneal reflections and measurements are interpreted via a logical process called discrimination. By this eye-tracking system, visual fixations are defined as series of gazes in which an individual stays within 1° visual angle for 100 ms or longer.

Participants were seated 61-66 cm from a table-mounted computer monitor, which was placed directly above the eye-tracking apparatus. This model did not require participants to wear a head-mounted devise because it utilized head-tracking to subtract participant movement from eye-tracking measures. Due to the extreme sensitivity of eye-tracking, we still asked that participants kept their chin on a chin rest throughout the experiment. Image 1 shows participant and eye-tracker set-up.

Image 1. Eye-tracker set-up with participant.
**Procedure**

In this study, participants were seated in front of the eye-tracker and asked to listen to sentences while simultaneously looking at a monitor. Each sentence stimulus was presented with a visual display showing four objects which were presented in the four corners of a 2 x 2 grid displayed on a computer screen. On a given experimental trial, two images represented the two nouns heard in the sentences were presented with two unrelated clip art objects that would serve as distractors. Auditory stimuli was played through a headset and at a volume level (40 db) that was consistent with all participants. Participants were instructed to identify the noun performing the action (the agent) by pointing, to the correct agent with a computer mouse, which assumes necessary gazing at the image. The position of the correct object varied from sentence to sentence to prevent gaze tendencies and strategy formation. Also, to control for any potential effects of image frequency, visual stimuli were organized in such a way that no participant would see the same distractor image twice. A total of 240 unique images were used in the design of this experiment. These pictures were drawn from the Snodgrass and Vanderwart (1980) picture set for normed name agreement, image agreement, familiarity, and visual complexity. An example of an experimental display showing the 2 x 2 grid set-up is presented in Figure 1.

![Experimental Display of Response Objects](image)

*Figure 1. Experimental Display of Response Objects.*

Presentation of response objects was organized into a 2 x 2 grid. In every trial, two words represented in the sentence were presented with two distractor images.
Calibration was performed for each participant followed by several practice trials to ensure understanding of the task. To begin, participants pressed the spacebar on a keyboard to begin the presentation of a sentence. The display of the four clip art objects was shown immediately and synced with the start of the sentence. Visual displays persisted for 1,970 to 2,030 ms after the end of the audio to provide subjects with a reasonable amount of time to make their decision and then disappeared to signal the end of the trial. Each participant was presented with a total of 30 experimental trials, which were separated by a short pause screen. To re-center participant gaze trial to trial, a grey fixation cross was shown in the center of the monitor. Accuracy of using the computer mouse to point to the correct object was recorded, along with eye-gaze data.

RESULTS

In this study we compared comprehension for subject-relative sentences and object-relative sentences. Our first measure was accuracy. Figure 2 plots the proportion of trials in which participants correctly identified the agent in subject-relative and object-relative sentences. Participants correctly recognized the agent in 100% of subject-relative trials and 95% correctly in object-relative trials. Though there was a trend for the agent to be incorrectly identified in object-relative trials more often, this trend was insignificant (paired t-test, p=0.14)

Figure 2. Plot of proportion correct for each sentence type. Blue represents subject-relative sentences and green represents object relative sentences.
**Temporal Discourse of Sentence Events**

Figure 3 shows the average time of events in both versions of sentence stimuli: subject relative and object relative versions. In this study, an event was defined as the onset of noun 1, end of the verb, and onset of noun 2. These markers define the relevant components of the sentence, allowing for measures of response latency. The difference in object-relative times is due to the passive voice sentence structure. For example, the verb in a subject relative sentence will simply read “observed,” while in object relative sentences the same verb will exist as, “was observed by.”

Onset of noun was used as an event marker because we believed taking response latency measures from the end of the noun would produce more variability when comparing trials due to nouns being of differing lengths. Our logic was that by using onset of noun this could be more controlled because participant predictions of the unfolding word would not be affected by word length. Thus, participant gaze would be directed to an image of a shorter word like lobster upon hearing “lob-” to the same capacity as to an image of a longer, multi-syllable word like rhinoceros upon hearing “rhi-.”

![Figure 3. Graph of the average time of events for each sentence type. Three events are defined: noun one onset is the beginning of the first noun in the sentence, end of the verb, and noun two onset marks the beginning of the second noun. Blue represents subject-relative sentences and green represents object-relative sentences.](image-url)
Relation between sentence complexity and listener eye movements

We used several measures of gaze to determine how syntactic complexity affected sentence comprehension. Figure 4 shows the average latencies of the first saccade to target after each event (onset of noun 1, end of verb, and onset of noun 2), for both subject-relative and object-relative sentence types. The average latency of the first saccade to target after noun one onset in subject-relative trials was 1.27 sec versus 1.79 sec in object-relative trials. An ANOVA found this difference in latency to be significant ($F(1,30)=6.40, p=0.017$). We found that the average latency of the first saccade to target after end of verb in subject-relative trials was 1.11 sec versus 1.13 sec in object-relative trials. No significant difference was observed between the average response latency after end of the verb between subject relative sentences and object relative sentences. The average latency of the first saccade to target after onset of noun two in subject-relative trials was 1.02 sec versus 0.72 sec in object-relative trials. An ANOVA found this difference in latency to be marginally significant ($F(1,30)=4.11, p=0.051$).

![First Saccade to Target, After Event](image)

Figure 4. Graph showing the average latencies of the first saccade to the target after each event (noun one onset, end of verb, and noun two onset). Blue represents subject-relative sentences and green represents object-relative sentences.

Figure 5 illustrates the total number of saccades to the target, competitor, and unrelated images after each event. Sentence type did not exert a significant effect on the number of saccades to target. The average number of saccades to the target after noun one onset for subject-relative trials was 5.6 and for
object-relative it was 4.8. Number of saccades to the target after end of the verb was 4.0 for both subject-relative and object-relative sentence types. After the onset of noun two, the number of saccades to the target was marginally greater in object-relative sentences, 3.74, than in subject-relative sentence, 3.59. Sentence type was found to have significantly interacted with number of saccades to competitor in subject-relative versus object-relative sentences after noun one onset, but not after end of verb or onset of noun two. The number of saccades to competitor in subject-relative trials after noun one onset was 2.1 and in object-relative trials it was 3.1. ANOVA found this difference to be significant (F(1,30)=9.83, p=0.004). This can be explained by the fact that the first noun stated in object-relative sentences was always the competitor so participants naturally gazed at the first mentioned referent.

![Graph showing the total number of saccades to target, competitor, and unrelated images after each event (noun one onset, end of verb, and noun two onset). Blue represents subject-relative sentences and green represents object-relative sentences.](image)

Figure 5. Graph showing the total number of saccades to target, competitor, and unrelated images after each event (noun one onset, end of verb, and noun two onset). Blue represents subject-relative sentences and green represents object-relative sentences.

It was expected that the overall proportion of time spent fixating on the target word in subject relative trials would be higher than in object-relative trials because listeners would be more effective in identifying the agent due to the decreased syntactical ambiguity of subject-relative sentences. Figure 6
shows the proportion of time spent fixating on the target, competitor, and unrelated image for both subject relative and object relative sentences at the three different event times. Following onset of noun one, the proportion of time spent gazing at the target, competitor, and unrelated images was 0.55, 0.14, and 0.31, respectively in subject-relative sentences. In object-relative trials time spent looking at target, competitor, and unrelated images following noun one onset was 0.42, 0.24, and 0.34, respectively. ANOVA showed that the difference in proportion of time spent looking at the target following onset of noun one was marginally significant when comparing subject-relative and object-relative trials; (F(1,30)=3.481, p=0.072). Proportion of time spent fixating on the distractor after noun one onset was found to be much more significant running ANOVA when comparing subject-relative and object-relative trials; (F(1,30)=13.48, p=0.001). Sentence type did not exert a significant effect on proportion of time spent fixating on the three image groups at other event times. The greater amount of time spent fixating on the competitor in object-relative trials after noun one onset is also likely a result of object-relative sentences always beginning with the competitor noun.

Figure 6. Bar graph depicting the proportion of time spent fixating on target, competitor, and unrelated images after each event (noun one onset, end of verb, and noun two onset). Dark blue represents target, blue represents competitor, and light blue represents unrelated images.
The most interesting and robust representation of online sentence processing differences due to syntactic complexity that we found with this project is shown in Figure 7. This graph shows a plot of the probability of saccading to the target word as a function of time and clearly illustrates the evolution of sentence comprehension in real-time. The blue curve represents the probability of looking at the target for subject-relative sentences. We see that as the sentence begins, the curve rises rapidly with the onset of noun one because the first noun was always the agent in subject-relative trials. As the participant continues to listen to the sentence, probability of gaze at the target word decreases as they begin to look around the display in anticipation of other potential referents. This period where the blue curve slopes downward marks the listener's active search for syntactic resolution where they are attempting to integrate what is being heard and visually perceived. The probability of saccading to the target word in object relative sentences is represented by the green line. It begins low because the distractor noun in those sentences was always stated first. Incremental unfolding of the sentence led to a slow increase in the probability of gazing at the agent until the onset of the second noun. Onset of noun two in object relative trials is marked by a steep increase in the probability of saccading to the target noun. Comprehension of subject-relative and object-relative sentences is similarly progressive because in both cases, the visual search is defined by the need to establish and update reference in real-time.

Figure 7. Plot of the probability of saccading to the target word as a function of time. Event times (noun one onset, end of verb, and noun two onset) are marked in seconds per sentence time. Blue curve represents subject-relative sentences and green curve represents object-relative sentences.
DISCUSSION

The goal of this study was to show that an online measure of sentence comprehension could be achieved through the use of eye-tracking as an online methodology. Overall, results were not as explicit as I had anticipated, but differences in sentence processing due to syntactic complexity were clearly observed. Prior research has shown that object-relative sentences are more tasking and cognitively demanding than their subject-relative counterparts (Ferreira, Henderson et al., 1996; Cooke et al.). This was expected to be reflected behaviorally through eye-tracking in a number of ways: object relative trials were to be characterized by more delayed initial saccades to the target, a lower number of total saccades to the target, and less time spent fixating on the target. The current study did not reveal such trends because the task was likely too easy. This idea is supported by the accuracy results of this study. Although the accuracy measures I obtained were not found to be significant, it is worth noting that participants never incorrectly identified the agent in subject-relative trials and only a few incorrect judgments were made in object-relative trials across all participants. In concordance with previous studies, this also demonstrates that object-relative sentences are more likely to produce comprehension errors (Just & Carpenter, 1992; Wingfield et al., 2003). Furthermore, figure 7 shows a significant difference in the probability of saccading to the target between the subject-relative and object-relative curves as the end of the sentence unfolds. This indicates that on average, hearing a subject-relative sentence versus an object-relative results in better comprehension for spoken language.

With this study, I hoped to also provide a more definite link between listener fixation patterns and the unfolding of syntactically complex sentences. It’s intuitive that comprehension of spoken language is incremental as the establishment of reference revolves around interpreting and integrating incoming words. An attractive explanation as to why subject-relative sentences are easier to process and immediately comprehend than object-relative sentences is because the grammatical components are
organized in a successive left-to-right string. This idea is supported by the Syntactic Prediction Locality Theory which states that the greater the distance between an incoming word and the dependent to which it attaches, the greater the integration cost for comprehension (Gibson, 1998). For instance, in a subject-relative sentence like, “The dog, that chased the cat, is vicious” the incoming word would be the second noun and the dependent would be the verb that acts on it. In this case, the integration cost would be low because the two words are extremely close and follow one another in sequential order. For the object-relative version of that sentence, “The dog, that was chased by the cat, is vicious,” the concept of dog must be kept active until it is resolved who is acting on it. In this case, the integration cost is much higher because the incoming second noun and dependent are not clearly related due to the syntactic structure.

The results of this study raise an interesting point regarding the discourse of subject-relative and object-relative sentence processing that may not be completely explained by this theory. Participants were instructed to make a decision regarding the identity of the actor as soon as they could. As seen in the figure 7 plot, the probability of saccading to the target word in subject-relative trials continued to decrease after unfolding of the end of verb and even onset of noun two. In theory, hearing the end of the verb in in a subject-relative trial should have given listeners sufficient information to complete the task. Yet on average, they would continue to saccade to other referents until about 400 ms after onset of noun two. Only at that time do we see a rise in the probability of saccading to the target. Contrary to this, the onset of noun two in object relative trials led to an increase in the probability of saccading to the target noun. Considering that object-relative sentences are supposed to be more confusing and difficult to parse through overall, how come resolution of these sentences occurred so immediately with the onset of noun two? It may be that speech recognition is not solely about logically piecing together the components of sentence as efficiently as possible. Though participants clearly understood the task, as is evidenced by the accuracy measures, there appears to be some inclination towards understanding
the entirety of the sentence that overrides simply trying to complete the task. It would be interesting to see if these trends are replicable or would strengthen with more participants.

**Future Directions:**

*Effects of Aging and Background Noise on Speech Comprehension:*

Speech is naturally rapid and thus requires significant effort to maintain recently heard dialogue in working memory. It's been found that a decline in working memory capacity can contribute to substantial difficulties in the comprehension of spoken language (Wingfield et al., 2005). Declines in working memory associated with adult aging (Wingfield et al., 2005) are thought to contribute to age-related comprehension failures, especially in the case of syntactically complex speech (Carpenter, Miyaki, & Just, 1994; Kemper, 1992; Wingfield, Peelle, & Grossman, 2003). Furthermore, deficits in central auditory processing which result in reduced resolution of incoming speech are believed to contribute to older adults' greater difficulties with parsing between speech and background noise (Wingfield et al., 2005). It would be interesting to expand the current study to include older adults to see how the combination of hearing loss with age related cognitive decline manifests into challenges in the comprehension of speech. In addition to using syntactic structure as a means of regulating comprehension difficulty, various levels of acoustic masking, in the form of background noise, might also be employed in this paradigm. Implementation of such online behavioral measures with eye-tracking may further define the interaction of the effects of hearing acuity, age, working memory, and competing sound on sentence comprehension.

*Effects of Embedded Clauses in English and Hebrew:*

English and Hebrew differ markedly in grammatical and syntactic structure. Through collaboration with Boaz M. Ben-David and colleagues in Israel, this specific research paradigm will be tested using both English and Hebrew native speakers as participants. This will offer the potential for gaining insight into the universal structure of spoken language comprehension.
References


