Eye movements and multi-utterance event descriptions

Introduction. The way speakers choose to describe a visual environment is often dictated by properties of the environment itself (e.g., entities or events). Previous work has investigated utterance form at the sentence level through visual manipulations of simple transitive events (e.g., Gleitman et al., 2007) as well as a relationship between patterns of attention and production decisions (Gleitman et al., 2007; Griffin & Bock, 2000). However, little to no work has examined how speakers order multi-utterance descriptions of real-world events. We investigated the relationship between scenes, online event apprehension, and extemporaneous descriptions to investigate multi-utterance linearization decisions. We hypothesized that speakers would first describe the event to which they allocated their attention prior to beginning their descriptions, a pattern consistent with previous results from sentence-level production paradigms.

Method. *Scene norming.* Photorealistic scenes containing no text were normed using two separate tasks (Fig 1) to ensure that the experimental images depicted either a single or multiple event (30 each). Subjects either (1) mouse clicked on the event(s) they perceived and provided the corresponding event label (e.g., loading truck) or (2) they indicated the number of events they perceived (1-6+) and provided the corresponding event label. Both tasks used the same images to ensure cross-task agreement. *Verbal Data.* Sixty subjects were shown 30 digitized and luminance-matched photographs (1024 x 768 pixel) of indoor and outdoor real-world scenes that depicted both single (15 scenes) and multiple (15) events. Subjects were asked to describe aloud the events being depicted while their eye movements were recorded. Speech was transcribed using Whisper's speech-to-text algorithm (Open AI) and then was manually corrected by a research assistant. Word onset and offset were determined by using Whisper Timestamped. *Event segmentation.* Event boundaries were determined using coordinate data collected in the first norming task (1). Coordinates for each mouse click of an event were considered; a clustering algorithm (DBSCAN) was used to determine the grouping of clicks. A silhouette score was used to determine the optimal number of clusters. The event boundary was determined by generating a convex hull for each cluster of clicks (Fig 1c). Event size and distance from the center of the scene were also considered. *Fixation identification.* We determined whether each fixation fell within an event boundary using the 'shapely' package in Python. The event that received the highest proportion of looks for each description's pre-speech interval was considered to be the fixated event. *Event Mentions.* A research assistant naïve to the hypotheses manually identified the first event mentioned by reading each verbal description (Fig 1d).

Results. For both single and multiple events, speakers took about 1750 ms (*SD* = 726.8 ms) to begin their description and made about four fixations (*SD* = 2.75) during this interval. Speakers took the same amount of time when preparing to describe a single (1769 ms) and multiple (1696 ms) event scenes, as revealed with a Bayesian t-test. Speakers started with the event that they had spent the most time fixating on prior to beginning their description ($\beta = 0.25$, $z = 2.48$, $p = 0.013$). Larger events were described earlier ($\beta = 0.95$, $z = 10.86$, $p < 0.001$) as well as those closer to the center ($\beta = -0.31$, $z = -5.42$, $p < 0.001$).

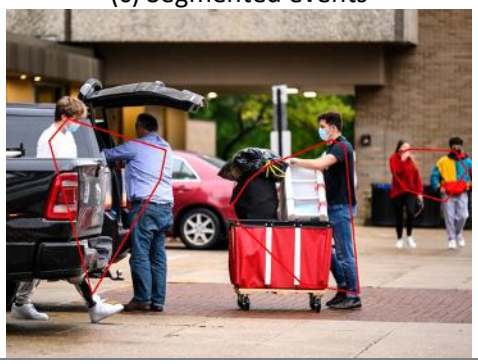
Discussion. Regardless of the number of events, speakers made about 4 fixations before beginning their event description. Additionally, we found that the event fixated on during the pre-speech interval was more likely to be mentioned first. Also consistent with previous work, size and center distance influenced linearization decisions. Overall, these results suggest considerable linguistic planning prior to articulation as well a relationship between initial looking patterns and linearization decisions, extending similar findings from previous work to include complex real-world scenes.

Figure 1. Single and multiple events

(a) Single event

(b) Multiple event

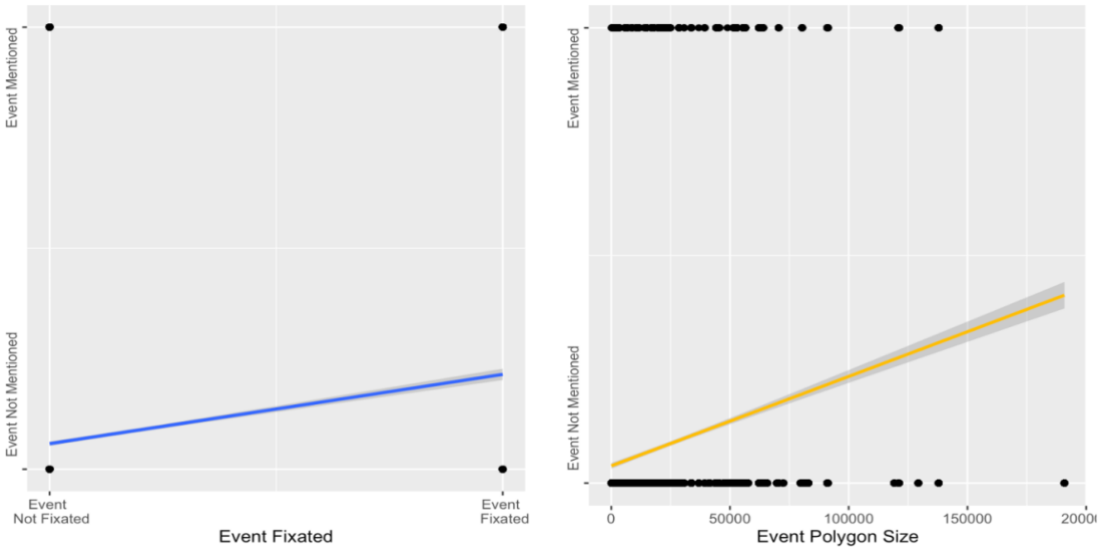
(c) Segmented events



(d) Description of scene (b)

"Looks like a guy unloading a car using a luggage cart. There is one guy in the car itself with the back hatch open. It looks like an SUV and he is taking stuff out, putting it into the cart that the other guy is holding for him. He might be checking into (NA) a hotel or an apartment or maybe a dorm."

Figure 3. Regression lines for predictors



1. Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. Journal of memory and language, 57(4), 544-569.
2. Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. Psychological science, 11(4), 274-279.