Scene descriptions reveal meaningful clustering in language production.

Language production requires us to linearize our thoughts into an ordered form. If we want to talk about five objects, for example, we need to decide which object will be mentioned first, second, etc. Past research has shown that production is incremental, where planning and production are interleaved into small chunks (Griffin & Bock, 2000; Ferreira & Swets, 2002). We do not need to wait until the whole utterance is planned out before starting to speak. However, most research on linearization has focused on the production of phrases or sentences. How one plans and executes a multi-sentence utterance is largely unknown. The present study begins to answer this question by looking at verbal descriptions of naturalistic scenes. Participants had a range of choices regarding what to talk about, how to talk about it, and in what order.

Thirty real-world images of indoor and outdoor scenes with no animate entities were used as stimuli (Figure 1). Thirty English-speaking participants viewed each image for 30 seconds and were asked to freely describe the image out loud in English during those 30 seconds. We used transcriptions of the audio recordings to generate a list of the objects in the scene that were mentioned with the onset time of each mention.

Past work using the same scene description task has found that individual objects' affordance-based features, such as how interactable an object appears to be, predict the order of mention (Barker et. al., 2023). Here, we investigate how relationships between objects play a role in how we organize multi-utterance production. We hypothesized that participants would create meaningful clusters of objects and balance local exploitation (talking about objects within a cluster) and global exploration (talking about objects in a new cluster). In support of the idea that descriptions have clusters, we found that objects that are mentioned close together are close together in space and meaning (Figure 2). Specifically, we ran a mixed effects model with differences in onset times of mentioned objects as the outcome (ms). We included physical distance (Euclidean distance between mentioned objects) and semantic distance (semantic similarity of mentioned objects calculated using ConceptNet multiplied by negative 1) as interacting fixed effects with random effects for subject, scene, and objects. We find a significant main effect of physical distance ($\beta = 838$, $t = 32.53$, $p < 0.0001$) and semantic distance ($\beta = 326$, $t = 9.43$, $p < 0.0001$), but no interaction ($\beta = -10$, $t = 0.20$, $p = 0.85$).

Next, we divided all objects into pre-specified physical and semantic clusters based on their physical location and object labels using a k-medoids clustering algorithm. We found that people's production of the next word takes substantially longer when jumping to an object that belongs to a new cluster compared to when they talk about another object within a cluster. We ran a mixed effects model with differences in onset times of mentioned objects as the outcome. Whether or not the next mentioned object was in a new cluster or the same cluster was coded as jumps (1 for a new cluster and -1 for the same cluster). We included jumps in physical clusters and jumps in semantic clusters as interacting fixed effects with random effects for subject, scene, and objects. We find a significant main effect of jumps in physical clusters ($\beta = 813.60$, $t = 38.40$, $p < 0.0001$) and jumps in semantic clusters ($\beta = 109.83$, $t = 4.92$, $p < 0.0001$). We also find a significant interaction ($\beta = 88.66$, $t = 4.17$, $p < 0.0001$), indicating a larger effect of jumps in physical clusters over semantic clusters (Figure 3).

Clustering in scene description may facilitate linearization by allowing for incremental production planning between clusters. The clustering we observed also suggests that planning multi-utterance production can be analyzed through the same lens as foraging behavior in animals, where animals try to find the optimal strategy for finding food by deciding between exploiting a patch (i.e., staying on the same tree to find more bananas) and exploring a new patch (i.e., traveling to a new tree to find more bananas). Ongoing work looks at how eye-tracking data compares to verbal descriptions and how people's clustering preferences differ when they describe scenes from memory. We explore the guiding principles of multi-utterance language production by examining how descriptions are influenced by task-oriented goals, semantic memory, and the balance between exploitation and exploration.

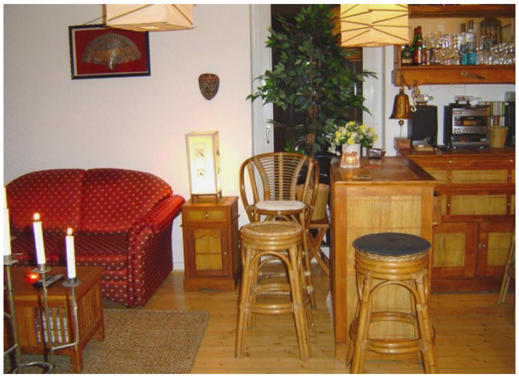

Figure 1. Example image of a scene participants were asked to describe.

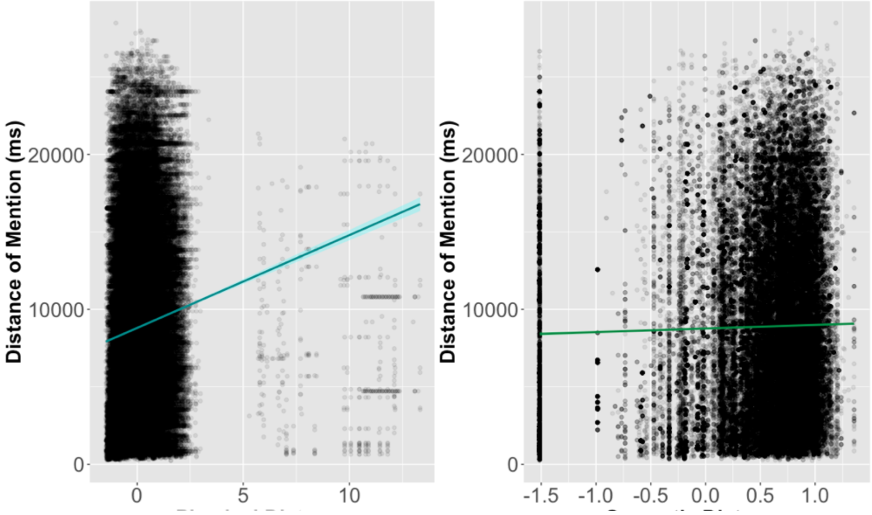

Figure 2. Scatter plots show the relationship between distance of mention and physical distance (left) and semantic distance (right). Each point represents a relationship between two mentioned objects. The blue line and green line represent best-fitting linear regressions and the shadings represent confidence intervals.

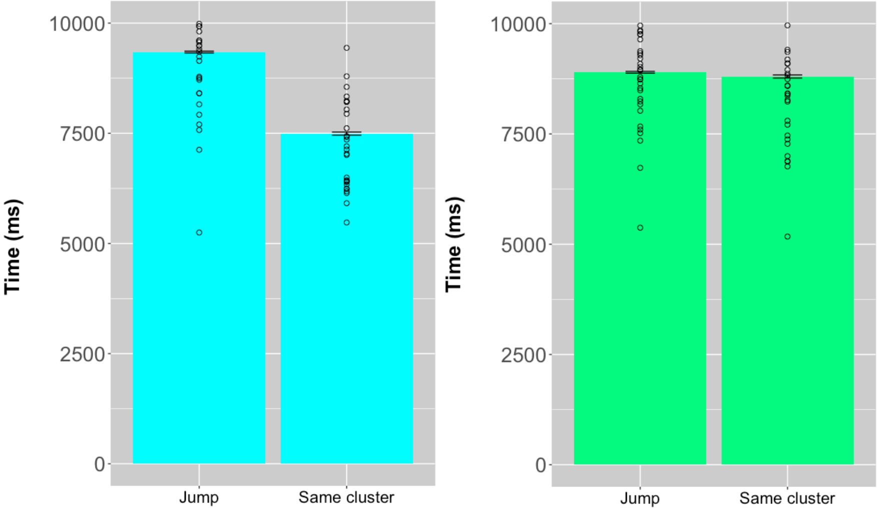

Figure 3. Bar plots show the time it took to talk about the next object in a new cluster (Jump) or the same cluster. Error bars represent standard error. Circles represent mean time for each participant.

References

Barker, M., Rehrig, G., & Ferreira, F. (2023) Speakers prioritise affordance- based object semantics in scene descriptions. *Language, Cognition and Neuroscience*, 38:8, 1045-1067

Ferreira, F., & Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, 46(1), 57–84.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, 11(4), 274–279.