Clustering during Scene Description from Memory

Karina Tachihara (karinat@illinois.edu)[1,2], Justin Horton[1], Adrian Zhou[3], Fernanda Ferreira[3]
[1]Department of Linguistics, University of Illinois, Urbana-Champaign
[2]Department of Psychology, University of Illinois, Urbana-Champaign
[3]Department of Psychology, University of California, Davis

Abstract

Language production requires speakers to resolve the problem of linearization, where they must choose what to talk about and in what order. Previous work has shown that scene descriptions while viewing the scene show evidence of clustering. Objects which are physically close together are mentioned close together and objects which are semantically similar are mentioned close together. Additionally, the transition time from one object to the next is longer when jumping across pre-defined physical and semantic clusters compared to when staying in the same cluster. We compared previous results with a new experiment where participants produced scene descriptions from memory. We find that descriptions from memory show evidence of clustering using physical and semantic relationships. Results also show that they rely more heavily on semantic relationships than descriptions generated while viewing the image. The work highlights the importance of task demands on linearization strategies in multiutterance production.

Keywords: language production; linearization; clustering; memory; chunking

Introduction

Describe the scene in front of you – this seemingly simple act is in fact a complex cognitive task that requires several interacting skills: visual perception, language production, planning, and more. The speaker needs to decide what to talk about, the order of the components, and whether to make a precise plan or not, all while keeping track of what has been said and what to talk about next. These choices are the central issues in the problem of linearization in language production, or the ordering of thoughts into a coherent, serial output (Levelt, 1981).

The problem of linearization is not unique to language production (Lashley, 1951). Goal-oriented, multistep processes often require a combination of planning and flexibility. Work in other cognitive domains has shown that the use of hierarchical structures can be helpful. For example, action sequences such as "driving into town" can be represented not by low-level actions such as the movement of every muscle through every inch of space but rather by high-level actions such as "start the car, turn left onto Orchard Street, then drive down main street" (Cohen, 2000). Spatial representation during navigation (Wiener & Mallot, 2003), event segmentation (Kurby & Zacks, 2008), foraging (Hills et al., 2020) and other types of multistep processes have shown evidence of hierarchical structures (Correa et al., 2023; Schapiro et al., 2013; Tomov et al., 2020). Clustering subcomponents is beneficial because a high-level plan can be made with the clusters without having to decide the order of every single item, drastically reducing the number of choices that need to be made before execution can begin. Language production may also benefit from a strategic organization that involve clustering of subcomponents. The difficulty and complexity of the task can impact the hierarchical structure in action sequences (Cohen, 2000). Similarly, the organization of language output may depend on the production procedure and the associated constraints. Here, we ask whether language production is organized differently when the contents are highly accessible during production compared to when the contents must come from memory.

Previous work investigating the ordering of components in language production has shown that people produce easy, previously mentioned, animate, and generally more cognitively accessible components earlier in the utterance (Christianson & Ferreira, 2005; Griffin & Bock, 2000; MacDonald, 2013; Morgan & Levy, 2016; Tachihara & Goldberg, 2020). This shows a tendency for language production to rely on incrementality, producing whatever comes to mind first and allowing other components to follow within the constraints of language. However, many of these studies were concerned with ordering of short utterances with few components, such as nouns in binomial expressions or active vs. passive structures in single sentences. Linearization may require a different strategy when the utterances in question are longer. When the number of items to be produced is larger, it may require more planning.

Work on multiutterance organization is more limited. (although see Roberts and Kirsner (2000) for work on fluency cycles). Descriptions of naturalistic scenes offer several advantages when investigating the nature of language production. When the task is open-ended, such as "describe the scene", scene images offer infinite possibilities, often with many items to talk about. These items are also interconnected and have meaningful relationships with one another. They do not exist independently in a random array. For example, in a kitchen scene, we expect a fork and a knife to be in closer physical proximity compared to a fork and a window. The viewer is also likely to see a fork and a knife as a meaningful set compared to a fork and a window. Because most multiutterance production in the wild deals with ideas that are interconnected rather than a random list, the task is likely to reveal beneficial strategies that are high in ecological

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

validity. Additionally, because we know which scenes the participants are describing, we are able to match their verbal descriptions to items in the scene. This allows us to take a data-driven approach to analyze the relationship between the descriptions and image properties.

Previous work on scene descriptions found that participants attend to and choose to talk about semantically meaningful areas of the scene (Hayes & Henderson, 2021; Ferreira & Rehrig, 2019; Henderson & Hayes, 2018; Henderson et al., 2018; Rehrig et al., 2020). Participants also often talk about the most interactable items earlier in the utterance (Barker et al., 2023). Tachihara et al., 2026 examined the relationships between all items mentioned within a description to investigate the overall organization of items. They found that the linearization problem in scene description is aided by grouping objects into meaningful clusters. Specifically, they found that 1) people talk about objects in close physical proximity in close temporal proximity; 2) people talk about objects which are semantically similar in close temporal proximity; and 3) people took longer to transition from one object to the next when jumping from one cluster to another compared to staying in the same cluster and produced more filler words during cluster jumps. Authors suggested that linearization in multiutterance production involves incremental planning, where speakers interleave planning and speech using meaningful clusters as boundaries.

In this paper, we asked whether task goals and constraints change clustering strategy during multiutterance production. Specifically, we were interested to see if scene descriptions from memory differ in how items are clustered from scene descriptions generated while viewing the scenes

It is well-established that chunking or clustering based on item characteristics, such as semantic categories, is an effective strategy in memory tasks (Miller, 1956; Thalmann et al., 2019). The idea is that clustering offloads some of the burden from episodic memory onto long-term memory. Therefore, on one hand, we could expect that descriptions from memory may show even stronger effects of clustering compared to descriptions that rely less on memory. On the other hand, clustering requires keeping a mental representation of the items and how they relate to each other. Memory of object-to-object relationships can be impacted when there is a change in the scene (Hollingworth, 2007). Therefore, if participants cannot keep a veridical representation of object-to-object relationships when the scene is not in front of them, then the resulting description may show less effects from these object-to-object relationships and thus show less evidence of clustering.

There are many possible factors along which speakers may cluster objects, such as physical location, semantic category, size, color, etc. We will concentrate on physical distance and semantic distance, given their relevance in guiding attention and production strategies in past studies (Hayes & Henderson, 2021; Tachihara et al., 2026). How much clustering relies on one type of information over the other may also change when the task shifts from production during viewing to production from memory. We hypothesized that scene description from memory would also rely on clustering as a strategy. However, the nature of the task may alter how much clustering relies on physical distance and/or semantic distance. While physical distance between objects is easily accessible during production if the scene is in front of them, it is inaccessible if the participant's memory of a scene does not include the location of objects. If participants choose to use a semantic clustering strategy to help them memorize the items in a scene, the output is likely to reflect greater semantic distances effects. We examined whether the organization of descriptions from memory reveal the same reliance on physical and semantic relationships between objects. The current project also shed light on the flexibility of clustering strategies. Is clustering an advantageous strategy in production under different task demands? Can it flexibly adapt given task constraints?

This paper will compare data and results from the study reported in Tachihara et al., 2026 available on OSF (https://osf.io/nymbh/overview?view_only=5df98bff1a4045a688685f6b1065db54) with a newly collected data set. In Tachihara et al., 2026, henceforth the viewing experiment, participants described scenes while they viewed the scenes in front of them. In the new study, henceforth the memory experiment, participants first viewed scenes, and then after a visual mask, described the scenes from memory while looking at a blank screen.

Methods

Participants

30 participants took part in the viewing experiment. Similarly, 30 new participants were recruited from University of California, Davis and participated for course credit in the memory experiment.

Stimuli

The same 30 naturalistic, indoor and outdoor scenes were used for both experiments. The scenes did not contain any animate entities except for one, which included a cat.

Procedure

The procedure of the study was closely matched to that of the viewing experiment except for the timing of the viewing phase and the description phase. In both studies, participants were asked to describe the scene freely for 30 seconds. In the viewing experiment, participants were asked to describe the scene while they viewed the scene in front of them. The 30 seconds of description phase started as soon as the image was presented. In the memory experiment, participants first viewed the scene silently for 30 seconds. The scene then disappeared from the screen. They were then presented with a visual mask for 1 second. After the visual mask disappeared, the 30 seconds of the description phase started, where participants described the scene verbally while viewing a blank screen.

Data processing

Figure 1. (Left) Sample scene image. A subset of items has been selected for visualization purposes. These items show object boundaries and labels. (Middle & Right) Sample scene image with object clustering. Object with the same colors belong to the same cluster. Middle image shows clustering based on physical distances between object centroids. Right image shows clustering based on semantic distances between object labels.

The data processing for the memory experiment followed the same steps described in Tachihara et al., 2026, which we summarize below.

Scenes Each scene was exhaustively segmented into objects with unique labels and boundaries using AnyLabeling (Nguyen, 2023) (Fig. 1 Left). Relationships between all objects in the scene were calculated. Physical distance used the segmented object boundaries to calculate the Euclidean distance between centroid points of each object pair. Semantic distance used the label given to each object to calculate inverse semantic similarity score (1 minus semantic similarity score) between object pairs. Semantic similarity was calculated using Concept Numberbatch (Version 17.06; Speer et al., 2017) by computing the cosine similarity of object vectors. We then created a cluster map for physical distance (Fig 1. Middle) and a separate cluster map for semantic distance (Fig. 1 Right). We used the k-medoid clustering algorithm within the scikit-learn-extra package to cluster the objects using their pairwise distances (Pedregosa et al., 2011). We found the optimal number of clusters by calculating the silhouette score for every cluster size between two and the number of objects in a scene minus one and then choosing the mutual cluster size with the lowest silhouette score. Note that we used a data-driven approach to object relations and clustering, where they were predefined using properties of the image without influence from the verbal descriptions of participants.

Verbal description Audio recordings from the description phase were transcribed using a speech-to-text algorithm, Whisper (OpenAI), and then manually corrected. The onset and offset of each word in the descriptions were identified using the Penn Phonetics Forced Aligner (p2fa) (Yuan & Liberman, 2008). Undergraduate research assistants identified the objects mentioned in the description using the transcript and the segmented scenes. We then looked at the pairwise object relationships between all objects that were mentioned in a single description of a scene by a participant and again calculated the physical and semantic distances. We also calculated the temporal distance, which was the amount of time that elapsed between the onset times of the mentioned objects. For objects which were mentioned directly one after the other or immediate transitions, we classified them as jumps if they crossed a cluster boundary or no jump if the two objects were in the same cluster. We classified jumps for the physical and semantic clusters separately. Since some words corresponded to many objects (ex: chairs), each relationship was given a weight according to how many objects were identified by a word (1/number of objects referred to by the first word * 1/number of objects referred to by the second word). This ensured that each mention of an object or objects were weighted equally informative in the model.

Results

We preregistered the comparative analyses between the viewing experiment and the memory experiment (https://aspredicted.org/nf8gj7.pdf). All models reported are maximally converging models.

Distance analysis

First, we examined if descriptions generated from the memory experiment are organized by physical and semantic distances of the objects. We found that objects that were mentioned close together were close together in space and meaning (Fig. 2). We ran a linear mixed effects model with temporal distance (ms) as our outcome measure, scaled physical and semantic distances as interacting fixed effects, and scene, participant, item 1, and item 2 as random effects with random intercepts, and included weights. We found significant main effects of physical distance (β =0.007, t = 2.76, p =0.006) and semantic distance (β =0.19, t =102.94, p<0.001) as well as a significant interaction effect (β =0.09, t =78.14, p<0.001). The interaction suggests that the effect of physical distance on temporal distance is stronger when

objects are semantically distant. The pattern of results was similar to that of the viewing experiment.

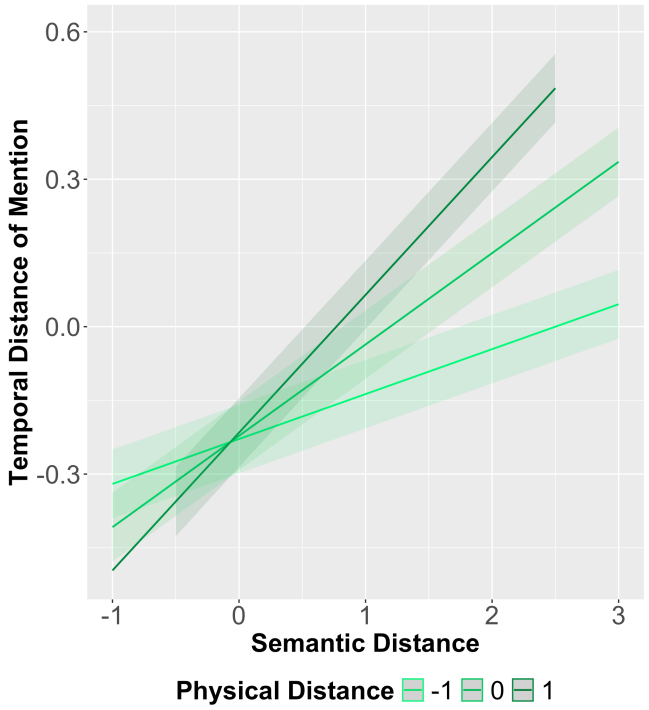

Figure 2. Model plots for the 2-way interaction between physical distance and semantic distance and the estimated temporal distance of mention (y-axis). Each line shows a different level of physical distance: 0 refers to the mean, −1 refers to −1 z score, and +1 refers to +1 z score. Shaded regions indicate 95 % confidence intervals.

Next, we wanted to know whether the organizations of descriptions differed in how much they relied on physical or semantic distances depending on the task – does describing a scene from memory differ from describing a scene while viewing the scene? Do they rely on physical and semantic distances equally? We first ran a linear mixed effects model with temporal distance (ms) as our outcome measure, scaled physical distance, scaled semantic distance, and experiment as interacting fixed effects, and scene, participant, item 1, and item 2 as random effects with random intercepts, and included weights. We found significant main effects of physical distance (β =0.19, t =8.59, p<0.001) and semantic distance (β =0.19, t =109.43, p<0.001), but not experiment (β =-0.02, t =-0.39, p=0.70). We also found significant interacting effects between all of the fixed effects. However, this combined model showed moderate collinearity (VIF scores of 6.2-7.3), making model interpretation difficult. Therefore, we examined physical and semantic distances in separate models. For physical distance, we found a main effect of physical distance (β =0.19, t =159.77, p<0.001) but no main effect of experiment (β =-0.06, t =-1.83, p = 0.07) or interacting effects of physical distance x experiment (β =-0.002, t =-1.43, p=0.15). This means that the descriptions from the viewing experiment and the memory experiment relied on physical distance to organize their utterance in equal amounts. For semantic distance, we found that organization of the descriptions from the memory experiment relied on semantic relations more heavily than from the viewing experiment. The model revealed a main effect of semantic distance (β =0.21, t =127.51, p<0.001), no significant main effect of experiment (β =-0.03, t =-1.05, p=0.30), and an interacting effect of semantic distance x experiment (β =-0.02, t =-7.57, p<0.001). This shows that semantic distance had a stronger effect on the temporal distance in the memory experiment compared to the viewing experiment.

Transition time analysis

To see if the predefined, image-driven physical and semantic clusters had an effect on the organization of the descriptions, we looked at how long it took to mention one object and then the next, which we will call immediate transitions. All immediate transitions were classified as either a jump, if the two objects belonged to different clusters , or no jump, if the two objects were in the same cluster. First, we examined if descriptions generated from the memory experiment were impacted by physical and semantic clusters, and found that they are. We ran a linear mixed effects model with temporal distance as our outcome measure, physical and semantic jumps (binary coded as -1=no jump or 1=jump) as interacting fixed effects, and scene, participant, item 1, and item 2 as random effects with random intercepts, and included weights.

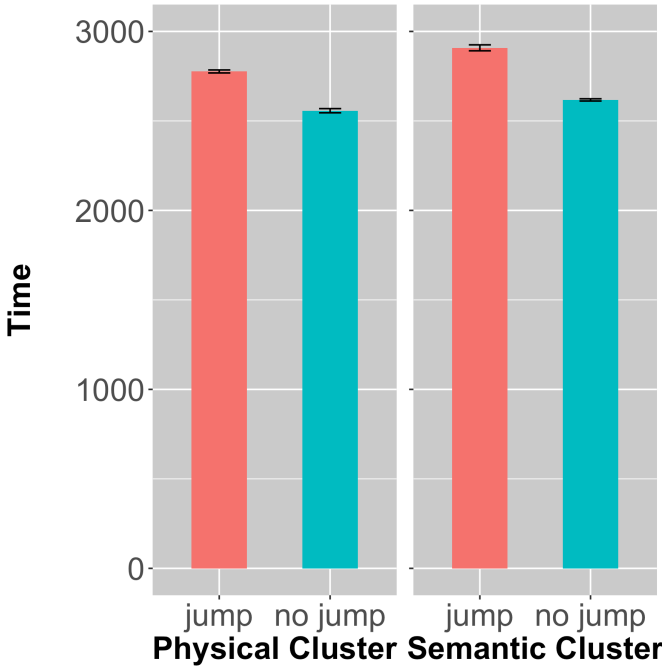

Figure 3. Time elapsed in ms between the onset of an object mention and the onset of the next object mentioned. The pink bar represents object transitions when the mentioned objects were in different clusters (jumps) and the blue bar when they were in the same cluster (no jump). Error bars represent standard error.

We found significant main effects of physical jump (β =0.17, t =25.82, p<0.001) and semantic jump (β =0.09, t =13.55, p<0.001), and a significant interaction effect (β =0.03, t =4.46, p<0.001). This means that it took longer to move onto the next object when that object belonged to a different physical cluster or semantic cluster compared to when the next object was in the same physical cluster or semantic cluster (Fig. 3). Cluster boundary effects were additive, meaning that when the transition involved jumping both a physical and a semantic cluster, it took the longest time.

Next, we wanted to see whether physical and semantic clusters had different effects on descriptions from the viewing experiment compared to descriptions from the memory experiment. We ran a linear mixed effects model with temporal distance as our outcome measure, physical jump, semantic jump, and experiment as interacting fixed effects, and scene, participant, item 1, and item 2 as random effects with random intercepts, and included weights. We found a significant main effect of physical jump (β =0.18, t =33.02, p<0.001), semantic jump (β =0.10, t =17.71, p<0.001), and experiment (β =-0.14, t =-2.17, p=0.03). This means that immediate transitions took longer for the memory experiment compared to the viewing experiment, which is to be expected given that participants have to find the next object to talk about from memory instead of choosing the next item from the scene in front of them. We also found significant 2-way interactions between semantic jump x physical jump (β =0.02, t =3.15, p=0.002), physical jump x experiment (β =-0.02, t =-2.44, p=0.01), and semantic jump x experiment (β =0.17, t =2.49, p=0.01). Interestingly, the direction of the interaction effects went in opposite directions, such that the difference between jump and no jump for physical clusters was larger for the memory experiment but the difference between jump and no jump for semantic clusters was smaller for the memory experiment compared to the viewing experiment. We interpret both of these interactions to be an effect of the longer immediate transition times. For the physical jumps, the longer transition times of the memory experiment led to a larger difference between jumps (viewing: M= 2274ms; memory: M= 2776ms,) and no jumps (viewing: M=2122ms; memory: M=2557ms). For semantic jumps, we observe a flooring effect. The transition times for the jumps are very close for both experiments (viewing: M=2909ms; memory: M=2902ms). The transition times for no jumps, however, is significantly lower for the viewing experiment compared to the memory experiment (viewing: M=1934ms; memory: M=2618ms). While no jump transitions take significantly less time than jump transition even in the memory experiment, it did not go below a certain threshold, presumably because of the effort it takes to recall the next object. This led to the flooring effect, where the difference between jump and no jump is smaller in the memory experiment than the viewing experiment, where the subsequent object in the same cluster is recalled at a much faster rate.

Discussion

In summary, we find that when describing a scene from memory, speakers continue to use a clustering strategy to organize their utterances. Because participants in the memory experiment did not have the scene that they are describing in front of them, the organization had to be based on the mental representation of the scene. The effects of physical distance and semantic distance on the description show that speakers not only maintained object-to-object relationships in their memory but that they used these relationships to guide the organization of multiutterance description from memory.

Compared to a closely matched experiment where the scene was visible during production, the memory experiment showed that the effect of semantic distance on the temporal distance between object mentions was larger. This is consistent with the chunking literature in memory, where memory for items is improved when they are grouped by associations in long-term memory, such as semantic relationships (Miller, 1956; Thalmann et al., 2019). Participants likely adopted a semantic clustering strategy precisely because they know that they will need to recall items in the scene once the scene is removed from the screen. While physical distance still had an effect on the descriptions from memory, its effect did not differ from the descriptions during viewing.

When looking at immediate transition times, we see that the average time to transition from one object to the next was longer for the memory experiment, leading to an increased clustering effect in the physical dimension but a flooring effect in the semantic dimension. While transitioning to objects in the same clusters is still quicker than jumping clusters, there may be limitations on how quickly one can transition from one object to the next when recalling objects from memory.

This study and its findings highlight the task-dependent nature of clustering during multiutterance production. We started the paper by describing how the ordering of components in production (i.e., linearization) may differ based on the length of what is being produced, with longer utterances requiring more planning. We add another layer of complexity to the problem of linearization by showing how task demands, such as reliance on memory, shift the organization of multiutterance production, even when the length of the production is the same (30 seconds).

The difference in clustering strategies could reflect changes in when clustering takes place. While the viewing experiment shows that clustering can start as soon as a scene is viewed, this does not preclude the possibility that clustering can take place later during the other phases. In the memory experiment, clustering may take place after viewing is complete. Additionally, there may also be a re-clustering process where the clusters and the objects that belong to them are rearranged based on the memory for the items. Clustering and/or re-clustering can occur during the masked period or during the production period. Clustering based on physical distance and clustering based on semantic distances can also occur at different times. For example, physical clustering

may take place during viewing because physical relationships are easier to calculate when the image is in front of them. On the other hand, semantic clustering may take place during production when items are being retrieved from memory. Due to semantic priming and spreading activation, retrieving one item will increase the likelihood that a semantically related item is retrieved next (Cutting & Ferreira, 1999). Future studies should investigate the timing of clustering and its relation to memory encoding, storage, and retrieval processes.

In conclusion, we find that clustering is an advantageous strategy in linearization during language production from memory and that the nature of clustering of can adapt to the constraints of the task to aid the organization of language production.

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. https://doi.org /10.1080/23273798.2023.2190136

Christianson, K., & Ferreira, F. (2005). Conceptual accessibility and sentence production in a free word order language (Odawa). *Cognition*, *98*(2), 105–135. https://doi.org/10.1016 /j.cognition.2004.10.006

Cohen, G. (2000). Hierarchical models in cognition: Do they have psychological reality? *European Journal of Cognitive Psychology*, *12*(1), 1–36. https://doi.org/10.1080/095414400382181

Correa, C. G., Ho, M. K., Callaway, F., Daw, N. D., & Griffiths, T. L. (2023). Humans decompose tasks by trading off utility and computational cost. *PLOS Computational Biology*, *19*(6), e1011087. https://doi.org/10.1371/journal.pcbi.1011087

Ferreira, F., & Rehrig, G. (2019). Linearisation during language production: Evidence from scene meaning and saliency maps. *Language, Cognition and Neuroscience*, *34*(9), 1129–1139. https://doi.org/10.1080/23273798.2019.1566562

Griffin, Z. M., & Bock, K. (2000). What the Eyes Say About Speaking. *Psychological Science*, *11*(4), 274–279. https://doi.org/10.1111/1467-9280.00255

Hayes, T. R., & Henderson, J. M. (2021). Looking for semantic similarity: what a vector-space model of semantics can tell us about attention in real-world scenes. *Psychological Science*, *32*(8), 1262-1270.

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6), 10. https://doi.org /10.1167/18.6.10

Henderson, J. M., Hayes, T. R., Rehrig, G., & Ferreira, F. (2018). Meaning Guides Attention during Real-World Scene Description. *Scientific Reports*, *8*(1), 13504. https://doi.org/10.1038 /s41598-018-31894-5

Hills, T. T., Todd, P. M., & Jones, M. N. (2015). Foraging in semantic fields: How we search through memory. *Topics in Cognitive Science*, 7(3), 513–534.

Hollingworth, A. (2007). Object-position binding in visual memory for natural scenes and object arrays. *Journal of Experimental Psychology: Human Perception and Performance, 33*(1), 31–47. https://doi-org.proxy2.library.illinois.edu/10.1037/0096-1523.33.1.31

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, *12*(2), 72–79. https://doi.org/10.1016/j.tics.2007.11.004

Lampe, L. F., Hameau, S., & Nickels, L. (2022). Semantic variables both help and hinder word production: Behavioral evidence from picture naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(1), 72–97. https://doi.org/10.1037/xlm0001050

Lashley, K. S. (1951). *The Problem of Serial Order in Behavior*.

Levelt, W. J. (1989). Speaking: From intention to articulation. MIT Press.

Miller, G. (1956). Human memory and the storage of information. *IEEE Transactions on Information Theory*, *2*(3), 129–137. https://doi.org/10.1109/TIT.1956.1056815

Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, *157*, 384–402. https://doi.org/10.1016/j.cognition.2016.09.011

Nguyen, V.A. (2023). *AnyLabeling - Effortless data labeling with AI support.* Version 1.2.0. https://github.com/vietanhdev/anylabeling

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (n.d.). Scikit- learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.

Rehrig, G., Hayes, T. R., Henderson, J. M., & Ferreira, F. (2020). When scenes speak louder than words: Verbal encoding does not mediate the relationship between scene meaning and visual attention. *Memory & Cognition*, *48*(7), 1181–1195. https://doi.org/10.3758/s13421-020-01050-4

Roberts, B., & Kirsner, K. (2000). Temporal cycles in speech production. *Language & Cognitive Processes*, 15(2), 129–157. https://doi.org/10.1080/016909600386075

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*(4), 486–492. https://doi.org/10.1038/nn.3331

Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. In S. Singh & S. Markovitch (Chairs), Proceedings of the ThirtyFirst AAAI Conference on Artificial Intelligence (pp. 44444451). AAAI Press.

Tachihara, K., Barker, M., Cotter, B., Hayes, T., Henderson, J., & Ferreira, F. (2026). Planning to Be Incremental: Scene Descriptions Reveal Meaningful Clustering in Language Production. *Cognition*, *266*(106330). https://doi.org/10.1016/j.cognition.2025.106330

Tachihara, K., & Goldberg, A. E. (2020). Cognitive accessibility predicts word order of couples' names in English and Japanese. *Cognitive Linguistics*, *31*(2), 231–249. https://doi.org/10.1515 /cog-2019-0031

Thalmann, M., Souza, A. S., & Oberauer, K. (2019). How does chunking help working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 37–55. https://doi.org/10.1037/xlm0000578

Tomov, M. S., Yagati, S., Kumar, A., Yang, W., & Gershman, S. J. (2020). Discovery of hierarchical representations for efficient planning. *PLOS Computational Biology*, *16*(4), e1007594. https://doi.org/10.1371/journal.pcbi.1007594

Wiener, J. M., & Mallot, H. A. (2003). "Fine-to-Coarse" Route Planning and Navigation in Regionalized Environments. *Spatial Cognition & Computation*, *3*(4), 331–358. https://doi.org/10.1207 /s15427633scc0304_5

Yuan, Jiahong, and Mark Liberman. "Speaker Identification on the SCOTUS Corpus." *The Journal of the Acoustical Society of America* 123, no. 5_Supplement (May 1, 2008): 3878–3878. https://doi.org/10.1121/1.2935783.