But what can I do with it?: Speakers name interactable objects earlier in scene descriptions

Madison Barker (msbarker@ucdavis.edu), Gwen Rehrig, Fernanda Ferreira (UC Davis)

Introduction: Spoken language requires speakers to decide what to say and when; deciding on a linear order is the linearization problem of language production (Levelt, 1981). previous research has suggested that image salience influences word order (Gleitman et al., 2007). More recent work found that image salience and meaning are correlated (Henderson & Hayes, 2017; Henderson et al., 2018), but neither image saliency nor scene meaning predicted the order in which objects are mentioned (Rehrig et al., 2020). Perhaps linearization decisions are based on another type of information that is more relevant to a human agent, such as object affordances. One type of object affordance, graspability, has been shown to predict visual attention (operationalized as fixation density) as well as meaning (Rehrig et al., 2020a). This study investigates whether object affordances more generally, which we term "interactability", predicts the order in which objects are mentioned in speakers' verbal descriptions. We hypothesized that objects that received higher ratings of interactability would be more task-relevant and would occur earlier in speakers' descriptions of the scenes.

Methods: Thirty native English speakers verbally described 30 real-world scenes, each for 30s, while eye-movements and speech were recorded (Henderson et al., 2018; Rehrig et al., 2020; see Fig.1a). To measure interactability, a separate group of participants was shown a black and white version of the scene with a single object shown in color (Figure 1b). Participants were asked to indicate on a scale from 1 (Very Unlikely) to 7 (Very Likely) the degree to which a human would interact with the highlighted object (Figure 1c). To obtain meaning and saliency values, the same objects that were rated for interactability were parsed into polygons using CVAT and LabelMe (Figure 1c). Object name referents were identified using a window of time and fixation data based on eye-voice span estimates (see Rehrig et al., 2020b).

Results: To assess word order, object mentions were identified with respect to their temporal onset in the verbal description. Meaning map (*M* = 0.43, *SD* = 0.13), saliency map (*M* = 0.37, *SD* = 0.12), and object interactablity values (*M* = 4.52, *SD* = 0.95) were used as predictors of word onset (*M* = 13623.61 ms, *SD* = 8253.49 ms; Figure 2). Object map values were averaged over the entire polygon (parsed in CVAT/LabelMe). The correlations revealed that neither meaning map values (r = -0.021, *p* = 0.54, Fig.1a) nor saliency map values (r = -0.034, *p* = 0.34, Fig.1b) were correlated with the order in which objects were mentioned. Consistent with our hypothesis, whole object interactability values did predict the order in which objects were mentioned (r = -0.14, *p* < 0.001, Fig.1c).

Discussion: Consistent with previous results, we observed that neither meaning nor saliency values predicted the order in which objects were mentioned. In contrast, object interactability did predict sequencing: Objects rated as more interactable were mentioned earlier in participants' verbal descriptions. These results add to our growing understanding of how complex verbal descriptions are planned and sequenced, suggesting that the specific aspect of meaning that influences utterance sequencing decisions is object interactability. When speakers plan multi-utterance sequences such as scene descriptions, they begin by identifying objects with which they would be inclined to interact. Overall, this work provides compelling evidence for the role of object affordance information in language processing.

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.

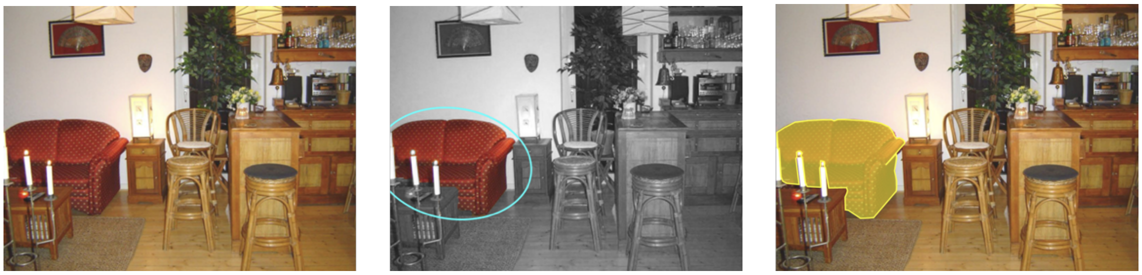
Levelt, W. J. (1981). The speaker's linearization problem. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *295*(1077), 305-315.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour, 1*(10), 743.

Henderson, J. M., Hayes, T. R., Rehrig, G., & Ferreira, F. (2018). Meaning guides attention during real-world scene description. *Scientific Reports, 8*, 13504.

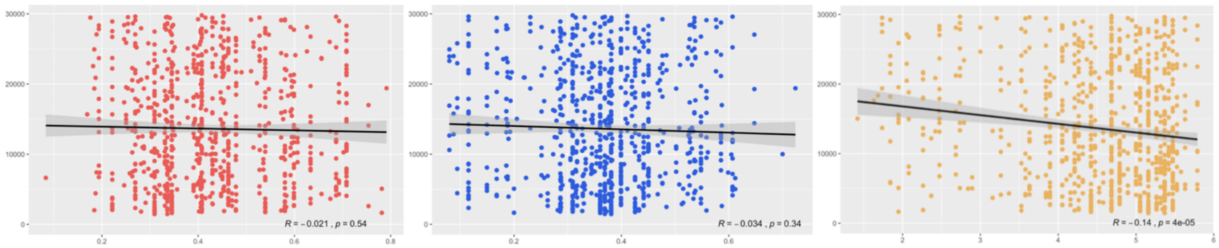
Rehrig, G., Peacock, C. E., Hayes, T. R., Henderson, J. M., & Ferreira, F. (2020a). Where the action could be: Speakers look at graspable objects and meaningful scene regions when describing potential actions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Rehrig, G., Hayes, T. R., Henderson, J. M. & Ferreira, F. (2020b). Setting the scene: Saliency and meaning in linearization during scene description. Poster presented on March 20th, 2020 at the 33rd Annual CUNY Human Sentence Processing Conference, University of Massachusetts, Amherst.



A. Scene B. Parsed for rating C. Parsed in LabelMe

Figure 1. A) Real-world scene presented to subjects in the description task. B) Object and scene context presented in the interactability rating task. C) Parsed object polygon overlaid on the scene. The average of the map values for pixels within the polygon served as meaning and saliency measures in the correlations.



A. Meaning B. Saliency C. Interactability

Figure 2. Scatterplots showing object name onset in the description on the x-axis (in ms) plotted against A) object meaning map values, B) object saliency map values, or C) whole object interactability ratings on the y-axis. Black regression lines indicate correlations.