Effects of Eye Movement Patterns and Scene-Object Relations on Description Production

Pelin Çelikkol (acelikkol@uni-potsdam.de)
Cognitive Sciences
University of Potsdam

David Schlangen (david.schlangen@uni-potsdam.de)
Department of Linguistics
University of Potsdam

Jochen Laubrock (jochen.laubrock@uni-potsdam.de)
Department of Psychology
University of Potsdam

Abstract

This study investigates whether fixation behaviour during scene viewing can offer insights into sequentialisation in verbal scene description production. We explored the correlation between visual and linguistic attention on naturalistic scenes using scene descriptions and eye movement measures. Results demonstrate an overlap in object prioritization during scene viewing and describing. Our additional analysis of scene descriptions reveals a tendency towards selecting and prioritizing category-specific objects.

Keywords: hierarchical perception; scene perception; dynamics of language production

Introduction

Language production entails organizing the cognitive content we aim to express into incremental segments. When describing our surroundings, we arrange our thoughts into a sequence to effectively convey the elaborate information we encounter. Thus, producing scene descriptions involves encoding the visual stimuli and articulating certain parts of the environment while simultaneously planning what to mention next.

Previous research on the interaction between perceptual understanding and language production focused on eye movement patterns during scene description generation and investigated whether visual attention informs language formulation (Bock, Irwin, Davidson, & Levelt, 2003; Gleitman, January, Nappa, & Trueswell, 2007; Griffin & Bock, 2000). Griffin and Bock (2000) provided evidence for a temporal coupling between the attended and mentioned parts of a scene when people were asked to describe pictures of events. In a similar experimental setting, Gleitman et al. (2007) further demonstrated a consecutive interaction between eye movements and spoken utterances, in which a fixation on a character preceded its mention. Bock et al. (2003) investigated eye movements during time-telling across different languages and time displays and observed a consistent trend in which the fixation points preceded the uttered expressions. The alignment between the attentional shifts and spoken utterances argues in favor of a process where the incremental sentence formulation is anticipated by comprehension (Griffin & Bock, 2000) and points to the need to linearise language due to attentional constraints (Levelt, 1981).

Levelt (1981) introduced the linearization problem in language production, which refers to the necessity of arranging our thoughts into order, thus prioritizing certain expressions over others. Previous research on the guiding mechanisms of linearisation suggested that more accessible information might be prioritized to reduce the cognitive load during simultaneous speech planning (MacDonald, 2013). In this view, speakers produce language by constructing a hierarchically organized cognitive schema depending on the task at hand. Shanon (1984) illustrated an example of such a structure in room descriptions: When people were asked to describe indoor places, they often started by categorizing the room, followed by relatively larger objects, and mentioned more fine-grained details towards the end. Dobnik, Ilinykh, and Karimi (2022) provided further support for this formulization and found that when people were asked to produce multi-sentence descriptions of indoor images, they tended to give essential information about the room accompanied by larger objects early on and focused on smaller objects and fewer details later. Ferreira and Rehrig (2019) investigated the linearisation phenomenon in the context of visual attentional guidance and analyzed the viewing behavior of speakers while they were describing indoor scenes. They found that a topographical representation of the scene informativeness accounted for more of the unique variability in participants' eye movements than a solely feature-based measure, suggesting a link between the modulating mechanisms of scene perception and language planning. They argued that following a quick extraction of a scene's gist, people's attention was driven towards more informative regions when describing the scene content. Barker, Rehrig, and Ferreira (2023) investigated the influence of different object features on their mention order during a scene description task and found that informative and interactable objects tended to receive earlier mentions as opposed to graspable objects that were mentioned later. They argued in favor of a linearization strategy in which speakers deemed recognizable and interactable objects as easily

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

accessible, thus prioritizing them when describing a scene. In this view, our generic scene knowledge and the information structures we have about the typical compositions in our environment drive our attention during perceptual processing and regulate the scene-related expectations (Henderson, Hayes, Peacock, & Rehrig, 2019), consequently influencing the linearization strategies during description production.

In the present study, we aimed to address two questions: First, we investigated whether fixation behavior during scene viewing provides insight into how the objects are prioritized during the production of scene descriptions, even if the production is recorded independently from the scene viewing. To this aim, we analyzed scene descriptions and eye movement measures collected in two independent studies. We obtained scene descriptions from the Tell-me-more corpus, which provides multi-sentence descriptions for real-world indoor images (Ilinykh, Zarrieß, & Schlangen, 2019). We used eye tracking data separately collected for a subset of the Tell-me-more images (Çelikkol, Laubrock, & Schlangen, 2023). We analyzed the relationship between visual and linguistic attention to individual objects using verbal object labels provided by the ADE20k image database (Zhou et al., 2019), on which Tell-me-more is based. Second, we assessed the modulating mechanisms of object prioritization during language production and focused on the effect of scene-object relations on the produced sentences. We utilized the linguistic object labels and adopted a statistical document analysis method, namely, the term frequency-inverse document frequency (TF-IDF) (Jones, 1988) when scoring object labels. This approach allowed us to obtain object ratings based on their categorical specificity (Çelikkol et al., 2023). We then used the object scores to predict if and when an object was mentioned during description production.

Methods

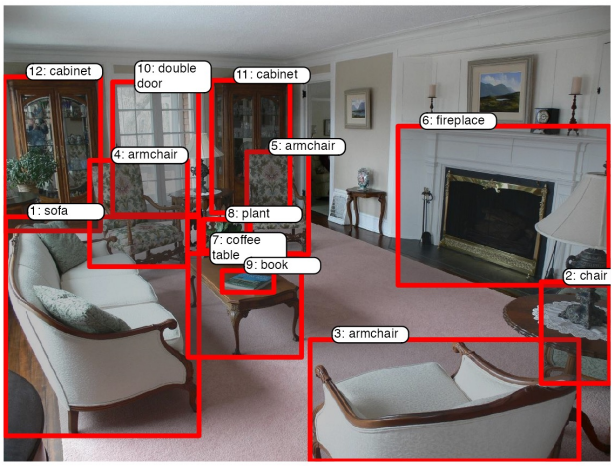
Datasets

Image Description Sequences We used verbal scene descriptions from the Tell-me-more dataset, which contains multi-sentence descriptions for 4410 real-world indoor images (Ilinykh et al., 2019). Descriptions were collected by asking participants to generate five successive sentences, imagining they were to describe an image to a partner who successively asked for more information in order to single out one image from a set of similar images. Participants were instructed to provide descriptions as quickly as possible and given separate subsequent text fields for each descriptive sentence, thus limiting their opportunity to make corrections. There were no restrictions put on the participants on how long or often they were to look at the images while describing. Images were taken from the ADE20k dataset, which additionally provides dense verbal object annotations (Zhou et al., 2019).

Eye Tracking Data We utilized eye movement data obtained using a subset of 145 images from the Tell-me-more dataset (Çelikkol et al., 2023). Data was collected from 23 participants who were asked to view each scene for 10 seconds and complete a caption-matching task. The task required participants to read a description after the image presentation and indicate with a key press whether the caption was correct or incorrect. The correct and incorrect conditions were balanced across the dataset, and incorrect captions were chosen from the same image category. The dataset provided eye movement metrics, including fixations and viewing times, on the corresponding object label obtained from the ADE20k dataset (Zhou et al., 2019).

Data Preprocessing

To compare the attentional priority during description production and scene viewing, we preprocessed the scene descriptions so that the noun phrases in descriptive sentences matched the corresponding ADE20k object annotation label. Human annotators inspected Tell-me-more descriptions and ADE20k object labels simultaneously and annotated the noun phrases in descriptions with the matching object labels. We followed the procedure described by Loáiciga, Dobnik, and Schlangen (2021) and annotated images to include those used in the eye tracking experiment. Figure 1 visualizes the results of the annotation process. This procedure allowed us to use identical object labels when comparing object mentions and eye tracking data.



1. This is a very formal looking living room with a sofa[1] and four chairs[2,3,4,5].
2. The sofa[1] and one chair[3] are white and the other chairs[2,4,5] have a print pattern on them.
3. There is a large pinkish color area rug on the floor and a fireplace[6] across from the sofa[1].
4. The coffee table[7] in front of the sofa[1] has a plant[8] on it and one book[9].
5. At the end of the room are french doors[10] with curio cabinets[11,12] on either side.

Figure 1: An example of an annotated image and its description sequence is shown. Bounding boxes represent the object labels obtained from the ADE20k dataset (Zhou et al., 2019), and the corresponding phrases in descriptions are shown in red with their numerical references. We only show bounding boxes of a few mentioned object labels for visualization purposes.

Data Analysis

Scene description data consisted of 1025 images that were described and later preprocessed, as detailed in the previous section. 320 unique object labels out of a total of 584 labels were identified as *mentioned* as a result of the preprocessing procedure. Eye tracking measures were available for a subgroup of 145 images containing 304 unique object labels. We used generalized linear mixed-effect models (GLMM) to assess the influence of eye movement metrics and object TF-IDF scores on the mentioned objects. The dependent variable of interest was the object's mentioned/not-mentioned status or the ordinal number of the object's mention within the description.

We first tested the predictors of interest using the subgroup of 145 images to jointly assess the effect of eye movement measures and object TF-IDF scores on the mentioned objects. We conducted an additional analysis including all images to test whether the effect of object TF-IDF scores holds when analyzed using a larger database. We constructed the models such that the baseline structure remained intact and the remaining covariates were adjusted based on the research question of interest. We fit the models using *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) in the R statistical environment (R Core Team, 2022). In the following, we describe the model structures and variables of interest at each analysis step.

Baseline Predictors Each model included the baseline predictors of object size and center distance. Object size and center bias are well-established influences on visual attention (Biederman, Mezzanotte, & Rabinowitz, 1982; Henderson & Ferreira, 2013). We included these variables as baseline predictors to control for any effects that might be introduced by visual attentional bias. We added images as random effects to account for any variability due to different scenes.

Eye Movement Metrics As eye movement measures, we used fixation probability and first arrival time, which refer to whether or not an object was fixated and the time the gaze first arrived at an object relative to the beginning of the trial, respectively. We averaged the measures over subjects to proxy average viewing behavior. To predict the likelihood of an object being mentioned, we fit a binomial GLMM with a logit link function and added the object's mentioned/non-mentioned status as the dependent variable. In addition to the baseline model structure, we added the object's fixation probability as the predictor of interest. We fit a second GLMM with a Poisson link function to predict the object's mention order and added the first arrival time as the predictor of interest.

Object TF-IDF Scores We scored object labels using the TF-IDF algorithm to investigate the influence of an object's contextual informativeness on description production. TF-IDF is a method used to assess the importance of a term in a document among a collection of documents (Jones, 1988). The term statistic correlates with the term frequency in a given document and is weighted by the inverse of the document frequency in which the term appears, consequently highlighting the document-specific terms. We used the following formula:

$$TF(t,d) = \frac{f_d(t)}{\max\limits_{w \in d} f_d(w)} \tag{1}$$

$$IDF(t,D) = log(\frac{|D|}{|d \in D : t \in d|}) \tag{2}$$

$$TFIDF(t,d,D) = TF(t,d) \times IDF(t,D) \tag{3}$$

The frequency of the term t in a document d is denoted by $TF(t,d)$, and $IDF(t,D)$ is the total number of documents $|D|$ divided by the number of documents in which the term t appears. We applied the TF-IDF method to the object and category labels obtained from the ADE20k data set, following the procedure described by Çelikkol et al. (2023). The resulting scores emphasized category-specific objects (e.g., *a toilet* in *a bathroom*) while outweighed object labels that frequently appear regardless of the image category (e.g., *floor*).

When analyzing the eye tracking subgroup, we added object TF-IDF scores along with the eye movement measures and conducted likelihood ratio tests (LRT) to assess the goodness of fit among the models containing baseline covariates, eye movement measures, and TF-IDF scores as predictors. We constructed two additional models and tested the effect of object TF-IDF on the mention probability and order using the extended database of 1025 images.

Permutation Tests To deal with the fact that we have only a small number of data points available when testing the mention order, we conducted permutation tests with 1000 simulations. We obtained p-values to infer statistical significance using the following formula:

$$p = P(|T| \geq t_{\text{obs}} \mid H_0) \tag{4}$$

where T denotes the test statistic and t_{obs} is its observed value under the null distribution.

Results

Table 1 summarizes the results of each full GLMM analyzing the effect of eye movement metrics and object TF-IDF scores on the mention probability and mention order. Figure 2 visualizes model predictions for each predictor of interest concerning the initial analysis of the eye tracking subgroup. LRT results comparing the baseline and the models of interest are shown in Table 2. This section first describes the results concerning the baseline covariates, followed by the effects pertaining to our research questions.

Baseline Predictors

All full model results showed that object size significantly affected the mention probability and order. Bigger objects were more likely to be mentioned and received earlier mentions

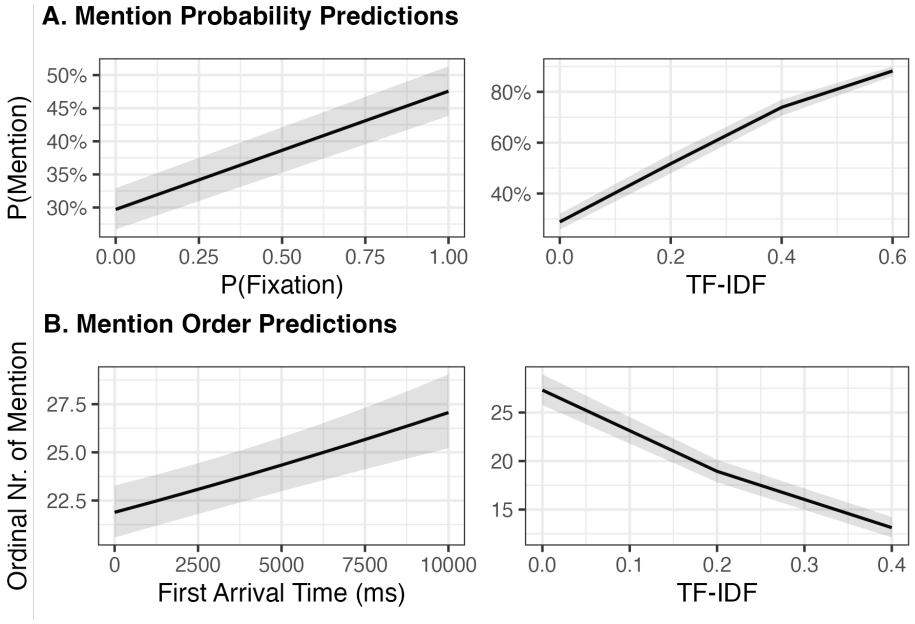


Figure 2: The full GLMM predictions are displayed for A) The probability of mention given the fixation probability and object TF-IDF scores B) The mention order given the first arrival time and object TF-IDF scores.

Table 1: Results of the full GLMMs

Variable Type	DV		Estimate	SE	Z	p
EM + TF-IDF	P(Mention)	Intercept	-0.593	0.076	-7.825	<0.001
		P(Fixation)	0.365	0.009	42.682	<0.001
		TF-IDF	0.387	0.008	45.616	<0.001
		Object Size	0.514	0.010	51.970	<0.001
		Center Distance	0.046	0.009	5.273	<0.001
	Mention Order	Intercept	3.163	0.029	109.643	<0.001
		First Arr.	0.040	0.006	6.273	<0.001
		TF-IDF	-0.145	0.007	-21.071	<0.001
		Object Size	-0.079	0.007	-11.231	<0.001
		Center Distance	0.027	0.006	4.271	<0.001
TF-IDF	P(Mention)	Intercept	-0.477	0.027	-17.932	<0.001
		TF-IDF	0.532	0.019	28.103	<0.001
		Object Size	0.325	0.020	16.539	<0.001
		Center Distance	-0.101	0.020	-4.978	<0.001
	Mention Order	Intercept	3.059	0.016	194.816	<0.001
		TF-IDF	-0.154	0.003	-47.409	<0.001
		Object Size	-0.118	0.003	-35.845	<0.001
		Center Distance	0.040	0.003	11.537	<0.001

Note. The GLMM coefficients of the variable type EM + TF-IDF represent the analysis carried out on the eye tracking subset of 145 images, followed by the TF-IDF analysis conducted with the extended database of 1025 images. EM: Eye movement measures.

than smaller objects. The effect of object distance on the mention probability and order was consistent in most cases: Peripheral objects were less likely to be mentioned and received later mentions than the central objects. The only exception was the positive correlation between the mention probability and center distance resulting from the analysis of the eye movement subset.

Table 2: LRT Comparison Results

DV	Model	BIC	Chisq	Df	p
P(Fixation)	Baseline	97982.22			
	Baseline + EM	96041.76	1951.84	1.0	<0.001
	Baseline + EM + TFIDF	93915.88	2137.26	1.0	<0.001
Mention Order	Baseline	16623.59			
	Baseline + EM	16582.99	47.73	1.0	<0.001
	Baseline + EM + TFIDF	16132.75	457.38	1.0	<0.001

Eye Movement Metrics

The analysis of the relationship between the fixation and mention probability revealed that an object's fixation probability significantly increased the likelihood of an object being mentioned. Among objects that were both fixated and mentioned, objects fixated earlier were significantly more likely to be mentioned earlier. The permutation test results revealed the significance of the observed effect of first arrival time on the mention order ($p = 0.037$).

Object TF-IDF Scores

We found a significant effect of TF-IDF scores on mention probability and order in the subgroup and extended analysis. There was a positive correlation between object TF-IDF scores and the object's mention probability, and earlier mentions were associated with higher TF-IDF scores. The effect of TF-IDF scores held when we added interaction effects with the baseline predictors in both the eye-tracking subgroup analysis ($\beta = 0.52$, $SE = 0.01$, $z = 56.09$, $p < .001$; $\beta = -0.18$, $SE = 0.01$, $z = -23.72$, $p < .001$, for mention probability and order, respectively) and in the extended analysis ($\beta = 0.63$, $SE = 0.02$, $z = 29.44$, $p < .001$; $\beta = -0.18$, $SE = 0.003$, $z = -50.11$, $p < .001$, for mention probability and order, respectively).

The permutation tests confirmed the significance of the object scores on the mention order ($p < 0.001$).

LRT Results

LRT comparisons carried out on the eye tracking subset revealed that the model including the eye movement measures better fit the data than the baseline model, both when predicting the mention probability and order. Adding the predictor of object TF-IDF scores further improved the model fit.

Discussion

We investigated the scene-object relations modulating prioritization during scene description production and whether eye movement control during scene viewing might inform the linguistic formulation process. Our results showed a correlation between the eye movement patterns and mentioned objects. Given that we obtained verbal and psychophysical data from studies conducted under different task instructions, our results can point to several directions: First, the more frequently mentioned objects were associated with a higher fixation probability, which may suggest that a common cognitive structure guides object selection during both scene viewing and describing. In the eye tracking study, participants were instructed to determine whether a description was correct for the preceding image, thus requiring them to memorize the scene content to a certain extent. Considering that fixation behavior often correlates with the recall of objects in memory tasks (Draschkow, Wolfe, & Vo, 2014; Tatler & Tatler, 2013), the distribution of fixations may offer insights into which objects were deemed crucial to memorize to succeed in the given task. The objects given attentional priority overlapped with those chosen for mention, which may imply a shared pattern in the information people extracted from the given environment. Similarly, we observed that the fixation sequence was predictive of the mention sequence, suggesting that object prioritization strategies were consistent across the two tasks. Given that both tasks required participants to inspect and differentiate a particular scene's content, the objects they prioritized may reflect the strategies used to optimize these processes.

In the context of top-down guidance of scene perception, attention is driven by the world knowledge accumulated through repeated exposure to typical environments. In this view, scene-based expectations are activated upon realizing a scene's gist, driving attention towards informative or meaningful regions (Henderson et al., 2019; Võ, 2021). Meaningfulness may refer to different object characteristics, including recognizability or interactability (Barker et al., 2023), and can be defined based on the environment type or task demands. We put emphasis on object distinctiveness reflected by the degree to which an object is diagnostic to a scene category and quantified it utilizing an easy-to-compute method, namely, object TF-IDF scores. We found that category-specific objects were more likely to be mentioned and given priority during the formulation of descriptions. These results are in line with a linearization process in which the high categorical relevancy of the objects made them more accessible after rapidly determining the scene type. At the same time, such objects may have been deemed more noteworthy in efficiently conveying the scene's content. The shared attributes

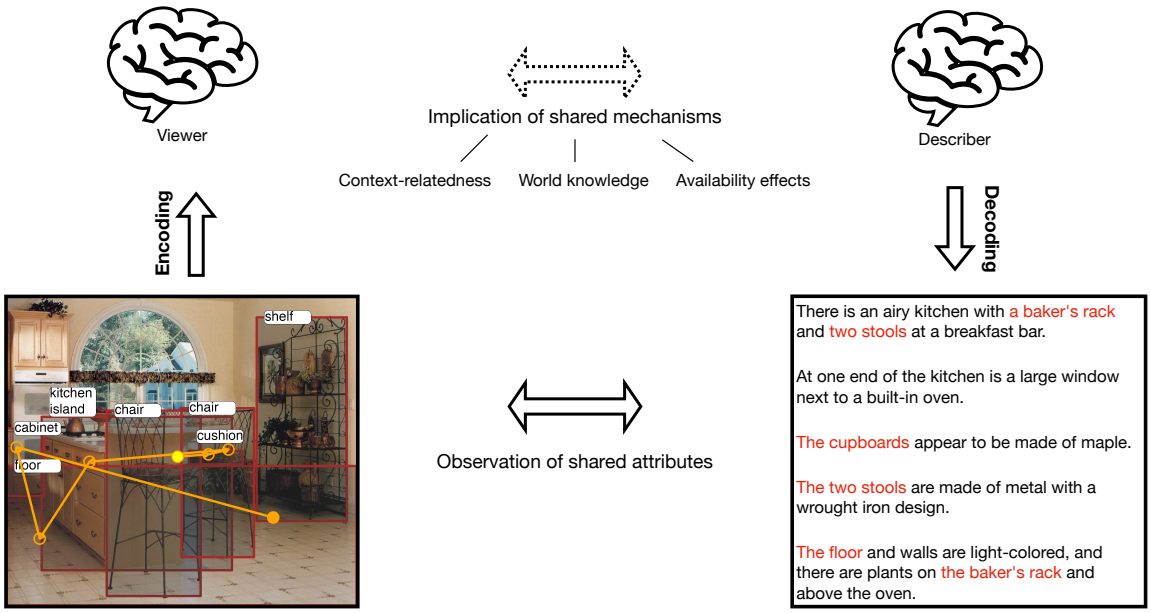


Figure 3: A schematic example of the visual encoding and linguistic decoding processes over the same image is presented. On the bottom left, a participant's scan path on an image is represented, and the starting point is shown with a yellow dot. Regressions to the same objects are not shown for simplicity. The descriptions obtained for the same image were presented at the bottom right. Both fixated and mentioned objects are shown in red. Given the shared attributes (the probability and order of the mentions/fixations) resulting from the two distinct tasks, potential shared mechanisms that may operate in optimizing these tasks are indicated.

observed during visual encoding and linguistic decoding of a scene may point towards underlying common mechanisms rooted in high-level world knowledge (Figure 3).

It should be noted that the information content of the objects in the current image database is limited due to the typicality of scenes. Given that contextual inconsistencies and surprisal have been found to influence the visual processing of scenes (Coco, Nuthmann, & Dimigen, 2020; Nuthmann, De Groot, Huettig, & Olivers, 2019), exploring this phenomenon during description production would provide further insights into the relationship between scene-object regularities and linearisation strategies.

The present study is subject to several limitations: The scene description data used in the study is limited to describers' written expressions, so we can not make inferences about how they visually processed the scenes at conceptualization or production stages. Thus, our interpretations of description production only concern the characteristics of the objects chosen to be mentioned and given priority, as well as potential similarities in how the same objects were visually attended. Given that the descriptions and eye movement data collected separately, our aim is not to establish a direct correspondence in the way visual and linguistic processes unfolded in the given studies but to point out that the regulating mechanisms of one modality may inform the other. Future studies concurrently investigating scene perception and language formulation will provide further insights into the underlying mechanisms of multimodal processing.

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. Retrieved from https://doi.org/10.1080/23273798.2023.2190136 (PMID: 37841974) doi: 10.1080/23273798.2023.2190136

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. Retrieved from https://doi.org/10.18637/jss.v067.i01 doi: 10.18637/jss.v067.i01

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, *14*(2), 143–177.

Bock, K., Irwin, D. E., Davidson, D. J., & Levelt, W. J. (2003). Minding the clock. *Journal of Memory and Language*, *48*(4), 653–685.

Çelikkol, P., Laubrock, J., & Schlangen, D. (2023). Tf-idf based scene-object relations correlate with visual attention. In *Proceedings of the 2023 symposium on eye tracking research and applications* (pp. 1–6).

Coco, M. I., Nuthmann, A., & Dimigen, O. (2020). Fixation-related brain potentials during semantic integration of object–scene information. *Journal of Cognitive Neuroscience*, *32*(4), 571–589.

Dobnik, S., Ilinykh, N., & Karimi, A. (2022). What to refer to and when? reference and re-reference in two language-and-vision tasks. *Proceedings of DubDial-Semdial*, 146–159.

Draschkow, D., Wolfe, J. M., & Vo, M. L.-H. (2014). Seek and you shall remember: Scene semantics interact with visual search to build better memories. *Journal of Vision*, *14*(8), 10–10.

Ferreira, F., & Rehrig, G. (2019). Linearisation during language production: evidence from scene meaning and saliency maps. *Language, Cognition and Neuroscience*, *34*(9), 1129–1139.

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.

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

Henderson, J. M., & Ferreira, F. (2013). Scene perception for psycholinguists. In *The interface of language, vision, and action.* Psychology Press.

Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2019). Meaning and attentional guidance in scenes: A review of the meaning map approach. *Vision*, *3*(2), 19. Retrieved from https://doi.org/10.3390/vision3020019 doi: 10.3390/vision3020019

Ilinykh, N., Zarrieß, S., & Schlangen, D. (2019, October–November). Tell me more: A dataset of visual scene description sequences. In *Proceedings of the 12th international conference on natural language generation* (pp. 152–157). Tokyo, Japan: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W19-8621 doi: 10.18653/v1/W19-8621

Jones, K. S. (1988). A statistical interpretation of term specificity and its application in retrieval. In *Document retrieval systems* (p. 132–142). GBR: Taylor Graham Publishing.

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

Loáiciga, S., Dobnik, S., & Schlangen, D. (2021, June). Annotating anaphoric phenomena in situated dialogue. In *Proceedings of the first workshop on multimodal semantic representations.* Online: Association for Computational Linguistics.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in psychology*, *4*, 226.

Nuthmann, A., De Groot, F., Huettig, F., & Olivers, C. N. (2019). Extrafoveal attentional capture by object semantics. *PLoS One*, *14*(5), e0217051.

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Shanon, B. (1984). Room descriptions. *Discourse Processes*, *7*(3), 225–255.

Tatler, B. W., & Tatler, S. L. (2013). The influence of instructions on object memory in a real-world setting. *Journal of vision*, *13*(2), 5–5.

Võ, M. L.-H. (2021). The meaning and structure of scenes. *Vision Research*, *181*, 10–20. Retrieved from https://doi.org/10.1016/j.visres.2020.11.003 doi: 10.1016/j.visres.2020.11.003

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. (2019). Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, *127*(3), 302–321. Retrieved from https://doi.org/10.1007/s11263-018-1140-0 doi: 10.1007/s11263-018-1140-0