Where the Action Could Be: Speakers Look at Graspable Objects and Meaningful Scene Regions when Describing Potential Actions

Gwendolyn Rehrig, Candace E. Peacock, Taylor R. Hayes, John M. Henderson and Fernanda Ferreira

University of California, Davis

Corresponding author:

Gwendolyn Rehrig

University of California, Davis

Department of Psychology

Davis, CA 95616

glrehrig@ucdavis.edu

Abstract

The world is visually complex, yet we can efficiently describe it by extracting the information that is most relevant to convey. How do the properties of real-world scenes help us decide where to look and what to say? Image salience has been the dominant explanation for what drives visual attention and production as we describe displays, but new evidence shows scene meaning predicts attention better than image salience. Here we investigated the relevance of one aspect of meaning, graspability (the grasping interactions objects in the scene afford), given that affordances have been implicated in both visual and linguistic processing. We quantified image salience, meaning, and graspability for real-world scenes. In three eyetracking experiments, native English speakers described possible actions that could be carried out in a scene. We hypothesized that graspability would preferentially guide attention due to its task-relevance. In two experiments using stimuli from a previous study, meaning explained visual attention better than graspability or salience did, and graspability explained attention better than salience. In a third experiment we quantified image salience, meaning, graspability, and reach-weighted graspability for scenes that depicted reachable spaces containing graspable objects. Graspability and meaning explained attention equally well in the third experiment, and both explained attention better than salience. We conclude that speakers use object graspability to allocate attention to plan descriptions when scenes depict graspable objects within reach, and otherwise rely more on general meaning. The results shed light on what aspects of meaning guide attention during scene viewing in language production tasks.

Keywords: language production, scene processing, object affordances, graspability, actions

Introduction

The world around us is visually rich and complex, yet we are able to describe what we see with relative ease. To do so, we must determine what details are relevant to talk about and what to describe first. When exploring a new environment, we may pay special attention to what actions can be carried out in the space: Is there anything to eat? Is there a coffee pot and the usual accompaniments? Is there a place to sit down? If we were to describe the new environment, would we first mention objects that one can interact with (e.g., “there’s coffee here”)? In the current study, we investigate what properties of real-world scenes drive attention as speakers plan and execute spoken descriptions. Novel to the current study, we ask whether the possible actions an agent can perform on an object, known as object affordances, influence visual attention relative to other scene properties. The other scene properties we considered were image salience and scene semantics. We chose to operationalize object affordances in terms of grasping interactions specifically, which we call ‘graspability’ throughout.

Object affordances may be special because object-related cognitive processes evoke motor activation associated with an action the object affords, as evidenced by studies that tap affordances directly (see Martin, 2007 for review). Silently naming or viewing an object activates regions of the premotor cortex (Grafton, Fadiga, Arbib, & Rizzolatti, 1997), as do lexical decisions about abstract words that relate to human motion (Harpaintner, Sim, Trumpp, Ulrich, & Kiefer, 2020), which suggests that motor representations for possible interactions with the object are evoked during both visual and linguistic processing (but see Mahon & Caramazza, 2008). Studies have suggested that comprehenders activate object affordances when interpreting a sentence’s meaning, especially those affordances that are most relevant based on recent experience or the current episodic context, which suggests that we understand the meaning of sentences in the context of human action (Kaschak & Glenberg, 2000; Glenberg & Kaschak, 2002; Glenberg, Becker, Klötzer, Kolanko, Müller, & Rinck, 2009). In an ERP study, both semantic and object affordance primes facilitated performance on a go/no-go task, but the time course of activation differed across prime types in go vs. no-go trials, suggesting that object semantics are not necessarily prioritized over affordances (Feven-Parsons & Goslin, 2018). When viewing graspable objects, subjects were faster to decide whether the observed object and an object mentioned in a sentence matched if the size and orientation of the object afforded grasping, which indicates that visual and linguistic processes are sensitive not only to what actions the agent could perform on an object, but also to how readily the object affords those actions (Borghi & Riggio, 2009; Borghi, 2012). Taken together, these studies demonstrate that object affordances are intrinsically linked to cognitive processes that operate on objects.

However, it is unclear whether the findings generalize to everyday cognition, or if associated affordances exert greater influence when they are task-relevant (Ostarek & Huettig, 2019).

Using the visual world paradigm, language studies have implicated object affordances in language comprehension more broadly (see Chambers, 2016 for review). Psycholinguists use eye movements in the visual world paradigm to study language comprehension (following Allopenna, Magnuson, & Tanenhaus, 1998). In a typical visual world paradigm study, participants' eye movements are recorded while they view a visual stimulus (an array of images, or a simplified scene) and hear recorded speech. Fixations are assumed to reflect listeners' expectations about upcoming speech based on linking hypotheses that posit a relationship between lexical activation and subsequent eye movements (Tanenhaus, Magnuson, Dahan, & Chambers, 2000; Altmann & Kamide, 2007; Salverda, Brown, & Tanenhaus, 2011). Research with this paradigm has revealed that listeners use the meaning of a verb to constrain potential referents to objects that afford compatible actions (Altmann & Kamide, 1999; 2007; 2009; Kako & Trueswell, 2000; Kamide, Altmann, & Haywood, 2003) and they can use affordances of objects in the display to resolve syntactic ambiguities (Chambers, Tanenhaus, Eberhard, Filip, & Carlson, 2002; Chambers, Tanenhaus, & Magnuson, 2004). In light of these findings, Altmann and Kamide (2007) suggested that eye movements during spoken language comprehension are intrinsically tied to object affordances: Objects in the display that are compatible with the verb attract visual attention. Relevant to our study, the authors predicted that the relationship may also explain eye movements during language production. Salverda et al. (2011) proposed that listeners use task goals to constrain the potential referents in the display based on object affordances. By and large, these results suggest that object affordances constitute the relevant semantic information listeners and speakers extract from scenes. We explore this idea in the current study by investigating whether object graspability drives productions that describe possible actions.

There is evidence that the interactions that an agent can perform on an object influence how observers process scenes. For example, scene functions (e.g., what someone could do in the scene) successfully predict scene categorization (Greene, Baldassano, Esteva, Beck, & Fei-Fei, 2016). Viewers categorize scenes based on whether the scene depicts a reachable space in which graspable objects would be in reach of the viewer (Josephs & Konkle, 2019), and neuroimaging studies have found activation linked to navigating reachable spaces (Bonner & Epstein, 2017; 2018). Several studies have suggested that objects and their functions influence attention allocation in scenes. Scene viewing tasks have shown that attention is allocated to meaningful objects in scenes, rather than to regions that contrast in image salience (Nuthmann

& Henderson, 2010; Einhäuser, Spain, & Perona, 2008). In related work, the functions of objects have been shown to influence visual attention (Malcolm & Shomstein, 2015; Castelhano & Witherspoon, 2016). When observers in Castelhano and Witherspoon's (2016) study saw novel objects and learned either about the function or the features of the object, those who learned the objects' functions used their knowledge to facilitate search. Object affordances constrain attention during visual search (Gomez & Snow, 2017), and object affordances exert greater influence on attention when the objects are physically present and within reaching distance, as compared to physically present objects that are out of reach, or to 2D and 3D object representations (Gomez, Skiba, & Snow, 2018). These findings suggest that we constrain visual attention to objects when we view scenes based on their affordances. In the current study, we tested whether object graspability preferentially guides attention relative to other properties.

While there has been considerable research on object affordances in language comprehension and visual processing, few studies have investigated how scene properties affect vision-language interactions in the context of language production, let alone whether object affordances influence seeing for speaking. Griffin and Bock (2000) studied eye movements during single-sentence descriptions of simple two-participant events. Speakers waited approximately 1,600 ms on average after the scene appeared to start speaking, which the authors suggested was the time required to interpret the scene before describing it. Gleitman, January, Nappa, and Trueswell (2007) conducted a follow-up study that employed a similar paradigm. Speech onset began approximately 2,000 ms after the speaker saw the scene, and eyetracking measures indicated the utterance was being planned within 200 ms of the scene's appearance. The pattern of viewer fixations before speech began was consistent with the order in which elements of the scene were mentioned. Gleitman et al. (2007) proposed that speakers both extract the gist of the scene and plan at least some of the utterance before speech begins, after which the remainder of the description unfolds incrementally.

In our own work (Ferreira & Rehrig, 2019; Henderson, Hayes, Rehrig, & Ferreira, 2018), speakers described real-world scenes either by freely describing what they saw or by describing what actions could be carried out in the scene (but were not depicted). Speakers who described the scene freely began speaking earlier on average (1,678 ms) than those who instead described the actions that someone could perform in the scene (2,548 ms). Because speakers saw the same scenes in both tasks, the difference in speech onset cannot be attributed to processes of scene gist extraction, but rather to task-specific planning demands. Furthermore, in our task speakers produced complex multi-clausal utterances, yet the pre-speech interval was comparable to those of Griffin and Bock (2000) and Gleitman et al. (2007), whose speakers

produced single sentences, which suggests that speakers use the pre-speech interval to form a general strategy, not to plan an entire utterance.

Griffin and Bock (2000) and Gleitman et al. (2007) both suggested that image salience influenced speakers' productions. Models of vision-language interactions generally assume an influence of image salience on language processes (Cohn, Coderre, O'Donnell, Osterby, & Loschky, 2018; Vogels, Krahmer, & Maes, 2013). These assumptions follow from saliency-based theories of visual attention (e.g., Itti & Koch, 2001; Wolfe & Horowitz, 2017) in which visual attention is drawn to peaks in image salience, defined as regions that stand out from the rest on the basis of low-level features (e.g., luminance, orientation, color). An alternative explanation for what guides attention in scenes comes from cognitive guidance theory (Henderson, 2007): The cognitive system directs gaze to informative areas for detailed scrutiny (Henderson, Malcolm, & Schandl, 2009; Henderson, 2017; Henderson & Hayes, 2017; Henderson, Hayes, Rehrig, & Ferreira, 2018). Image salience plays a lesser role in cognitive guidance theory, initially helping to determine the lay of the land, but scene regions are selected for attention on the basis of meaning, modulated by the viewer's current goals (both meaning and task-relevance guide attention).

Consistent with cognitive guidance theory, during visual search, viewers fixate regions that are both meaningful and salient (Henderson et al., 2007) and use scene semantics to guide search even when the image salience of search targets is low, despite the presence of a salient distractor in the scene (Henderson et al., 2009). Henderson and Hayes (2017) quantified the spatial distribution of meaning across a scene using meaning maps for direct comparison with image salience. To create meaning maps, raters were shown small patches that were taken from the scene at fine and coarse scales and were asked to rate how meaningful or recognizable the contents of the patches were (see Henderson, Hayes, Peacock, & Rehrig, 2019 for review). When meaning maps were compared to saliency maps generated using graph-based visual salience (GBVS, Harel et al., 2007), meaning and saliency were highly correlated.

Because meaning and saliency are related, it is possible that some of the evidence favoring saliency-based theories in natural scenes may be driven by meaning. In subsequent scene memorization and aesthetic judgment tasks, the meaning and saliency maps were compared to attention maps empirically derived from viewer fixations. Meaning explained attention better than salience did, especially when the shared variance explained by both meaning and salience was partialled out. Subsequent work that used meaning maps has garnered further support for cognitive guidance theory. Henderson and Hayes (2018)

reanalyzed data from the original study (Henderson & Hayes, 2017) by incorporating fixation durations in empirically derived attention maps so that longer fixations, when subjects presumably processed the fixated region more carefully, contributed more to the attention map. Meaning also explained visual attention better than image salience did when speakers described rich real-world scenes, regardless of whether their task was to describe the scene freely or to describe what actions could take place in the scene (Henderson et al., 2018) and at all stages of description planning: prior to speaking, during filled and silent pauses, and during the entire speaking period (Ferreira & Rehrig, 2019). Peacock, Hayes, and Henderson (2019a) found that even when subjects rated the brightness of or counted bright patches in scenes, in which case the viewer's goal explicitly required attention to salient regions, scene meaning (not image salience) explained where they looked. Consistent with cognitive guidance theory, viewers attend not to parts of the scene that have high image salience, but instead to regions of the scene that are meaningful.

Meaning maps have provided compelling evidence that cognitive systems push visual attention to regions of the scene that are meaningful. What remains unclear is whether viewers tap into specific kinds of information, rather than meaning in a broadly defined sense, when doing so may facilitate the task at hand. In the current study, we showed real-world scenes to speakers and asked them to describe the actions that someone could perform in each scene. Because object affordances have been shown to influence attention (e.g., Gomez & Snow 2017, Gomez et al., 2018), it is possible that meaning accounts for variance in attention because it captures something about whether we can interact with what is shown in the scene. In the current study we dissociated grasping affordances from meaning by pitting meaning against graspability. While object affordances are not limited to grasping interactions, we chose to examine graspability because many commonplace objects that appear in real-world scenes afford grasping interactions. We expect graspability to be particularly relevant when speakers describe possible actions in a scene.

The type of scene (e.g., a kitchen) and the objects in the scene (e.g., a tea kettle) determine what actions are possible. It follows that speakers may preferentially attend to graspable objects in the scene and draw from a set of afforded actions to include in their descriptions. This hypothesis is consistent with work on object affordances and with cognitive guidance theory: When describing the actions that could be carried out in a scene, speakers may push visual attention to task-relevant objects in the scene based on their affordances. Because previous studies on the relationship between visual attention and object affordances in scenes did not also consider scene meaning and image salience, our hypotheses about how

graspability relates to the other scene properties we have quantified are more speculative. To the extent that objects constitute midlevel scene features (Malcolm & Shomstein, 2015), graspability may overlap with image salience, especially if salient object boundaries (edges) or other low-level features contribute to graspability. Insofar as graspability measures semantic information about the scene (as suggested about affordances more broadly by Altmann & Kamide, 2007), graspability may share features with meaning. Indeed, given how broadly defined meaning is in the rating task used to make meaning maps in Henderson and Hayes (2017), we may expect graspability to tap into some of the same information as general meaning, especially if raters interpret “informative” and “recognizable” (the definition of meaning used by raters for constructing meaning maps) as properties that are strongly tied to objects.

In the current study, we quantified three scene features for a series of rich, full color, real-world scenes: we used graph-based visual salience (GBVS; Harel et al., 2007) to generate saliency maps, and we crowdsourced rating judgements for scene patches from two spatial scales (coarse and fine) based on either meaning or graspability. We then separately constructed meaning and grasp maps by averaging their respective ratings. In a third experiment, we additionally probed reachability ratings and constructed reach maps. In three experiments, we presented scenes individually to subjects and asked them to describe what actions would be possible in each scene. No actions were depicted in the scenes, so speakers had to generate actions that were compatible with the objects shown in the scene, and with the scene’s category. During the viewing period, we recorded speakers’ eye movements and spoken descriptions. To investigate processes of language planning and production further, we identified the onset and offset of speech, and calculated description duration. Following our previous work, we constructed attention maps from viewer fixations and examined the relationship between attention and each of the feature maps to determine which best explained attentional guidance in scenes. Using the same approach, we further analyzed visual attention at individual fixations, with an emphasis on early fixations that occurred during the pre-speech interval, to determine which scene features drive language planning.

Based on our previous work (Henderson & Hayes, 2017; 2018; Henderson et al., 2018), and counter to some claims in the literature (e.g., Parkhurst et al., 2002; Gleitman et al., 2007), we anticipate that speakers will not attend to regions of the scene to describe on the basis of image salience. We expect to once again replicate the robust advantage of scene meaning over image salience that we have found in multiple studies and across different tasks (Henderson & Hayes, 2017; 2018; Henderson et al., 2018; Peacock et al., 2019a; 2019b; Hayes & Henderson 2019a; 2019b). If the semantic information viewers extract from scenes is reducible to object

affordances, then we expect meaning and grasp maps to share considerable overlap, and to account for variance in attention equally well. If object affordances differ from meaning, and if they are (presumably) more task-relevant than meaning, we expect graspability to guide visual attention in our task better than scene meaning, especially during pre-speech fixations when speakers engage in macroplanning. We do not have a specific *a priori* prediction about how graspability will fare against image salience, but it follows from the aforementioned predictions that we should find an advantage of graspability over image salience.

Experiment 1

For this experiment we conducted a new analysis on a subset of the data previously reported in Henderson et al. (2018). In the current analysis, our goal was to determine whether graspability preferentially guided visual attention more than meaning when speakers described actions that could be performed (but were not shown) in the scene.

Subjects

Thirty-two undergraduate students enrolled at the University of California, Davis participated in exchange for course credit. All subjects were native speakers of English, were at least 18 years old, and had normal or corrected-to-normal vision. They were naive to the purpose of the experiment and provided informed consent as approved by the University of California, Davis Institutional Review Board. Two subjects were excluded from analysis because their eyes could not be accurately tracked; data from the remaining 30 subjects were analyzed.

Stimuli

Thirty digitized (1024x768) and luminance-matched photographs of real-world scenes depicting indoor and outdoor environments were presented (see Henderson et al., 2018 for details on all images tested). A subset of 15 scenes that were mapped for graspability were analyzed. Of these, 3 scenes depicted outdoor environments (2 street views), and 12 depicted indoor environments (2 kitchens, 3 living rooms, 2 desk areas). People were not present in any of the scenes.

Meaning, Graspability, and Saliency Map Generation

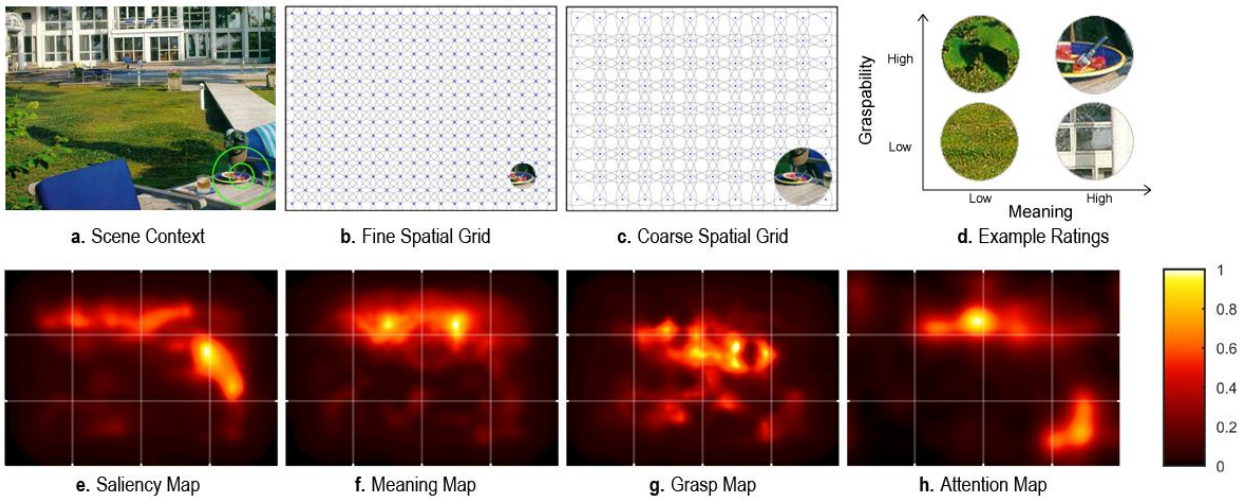
Meaning Maps. We generated meaning maps using a contextualized rating procedure adapted by Peacock, Hayes, and Henderson (2019b) from the context-free rating method introduced in Henderson & Hayes (2017). Each of the 20 scenes (1024 x 768 pixel) was decomposed into a series of partially overlapping circular patches at fine and coarse spatial scales (Figure 1b&c). The decomposition resulted in 6,000 unique fine-scale patches (93 pixel diameter) and 2,160 unique coarse-scale patches (217 pixel diameter), totaling 8,160 patches

across scales. Subjects viewed scene patches presented alongside an image of the scene. A superimposed green circle indicated what region the patch came from to provide context (Figure 1a).

Raters were 84 subjects recruited from the Amazon Mechanical Turk crowdsourcing platform. All subjects were located in the United States, had a HIT approval rating of 99% or more, and only participated once. Subjects provided informed consent and were paid $0.50 upon completing the task.

All but one subject rated 300 random patches extracted from the 20 scenes, presented alongside a small (256 x 192 pixel) image of the scene for context. Subjects were instructed to rate how informative or recognizable each patch was using a 6-point Likert scale (‘very low’, ‘low’, ‘somewhat low’, ‘somewhat high’, ‘high’, ‘very high’). Prior to rating patches, subjects were given two examples of low-meaning and two examples of high-meaning scene patches in the instructions to ensure that they understood the task. Scene-patch pairs were presented in random order. Each unique patch was rated 3 times by 3 independent raters for a total of 24,480 ratings. Because there was a high degree of overlap across patches, each fine patch contained data from 27 independent raters and each coarse patch from 63 independent raters (see Figure 1d for example of patches rated low and high in meaning).

Meaning maps were generated from the ratings by averaging, smoothing, and combining the fine and coarse scale maps from the corresponding patch ratings. The ratings for each pixel at each scale in each scene were averaged, producing an average fine and coarse rating map for each scene. The fine and coarse maps were then averaged [(fine map + coarse map)/2]. This procedure was used for each scene. Because subjects in the eyetracking task tended to look closer to the center of the image than the periphery, a phenomenon called center bias (Hayes & Henderson, 2019a), center bias was applied to the maps by downweighting the periphery of the maps. The final map was blurred using a Gaussian filter via the Matlab function ‘imgaussfilt’ with a sigma of 10 (see Figure 1f for an example meaning map).



a. Scene Context **b.** Fine Spatial Grid **c.** Coarse Spatial Grid **d.** Example Ratings

e. Saliency Map **f.** Meaning Map **g.** Grasp Map **h.** Attention Map

Figure 1. (a-d) Meaning and grasp map generation schematic. (a) Real-world scene. Raters saw the real-world scene and either a fine (inner) or coarse (outer) green circle indicating the origin of the scene patch under consideration. (b-c) Fine-scale (b) and coarse-scale (c) spatial grids used to create scene patches. (d) Examples of scene patches that were rated as low or high with respect to graspability and meaning. (e-h) Examples of saliency (e), meaning (f), graspability (g), and attention (h) maps for the scene shown in (a). The attention map (h) was empirically derived from viewer fixations.

Grasp Maps. Grasp maps were constructed from ratings in the same manner as meaning maps, with the critical exception that subjects rated each patch on how ‘graspable’ the region of the scene shown in the patch was. In the instructions, we defined ‘graspability’ as how easily an object depicted in the patch could be picked up or manipulated by hand. If a patch contained more than one object or only part of an object, raters were instructed to use the object or entity that occupied the most space in the patch as the basis for their rating. The remainder of the procedure was identical to the one used to generate meaning maps.

Raters were 84 subjects recruited from the Amazon Mechanical Turk crowdsourcing platform. Subjects were selected according to the same criteria that were used in the meaning map generation task, and received the same monetary compensation.

Each subject rated 300 random patches extracted from the 20 scenes, presented alongside the scene image for context. Subjects were instructed to rate how graspable each patch was using a 6-point Likert scale (‘very low’, ‘low’, ‘somewhat low’, ‘somewhat high’, ‘high’, ‘very high’). Prior to rating patches, subjects were given two examples of low-graspability and two examples of high-graspability scene patches in the instructions to ensure that they understood the task. Scene-patch pairs were presented in random order. Each unique patch was rated 3 times by 3 independent raters for a total of 24,480 ratings (see Figure 1d for example of patches rated low and high in graspability).

Grasp maps were generated in the same manner as the meaning maps: By averaging, smoothing, and combining the fine and coarse scale maps from the corresponding patch ratings. The ratings for each pixel at each scale in each scene were averaged, producing an average fine and coarse rating map for each scene. The fine and coarse maps were then averaged [(fine map + coarse map)/2]. This procedure was used for each scene. Because subjects in the eyetracking task showed a consistent center bias in their fixations, center bias was applied to the maps by downweighting the periphery of the maps. An example grasp map can be seen in Figure 1g.

Saliency Maps. Image-based saliency maps were constructed using the Graph-Based Visual Saliency (GBVS) toolbox in Matlab with the default parameters (Harel et al., 2006).

Prior to analyzing the fixation data, maps were normalized to a common scale using image histogram matching via the Matlab function 'imhistmatch' in the Image Processing Toolbox. For all map types (meaning, graspability, saliency), the corresponding attention map for each scene served as the reference image. See Henderson & Hayes (2017) for more details.

We then computed three correlations (R^2) across the maps of 15 scenes used in Experiment 1 to determine the degree to which saliency, meaning, and graspability overlap with one another. On average, the relationship between meaning and grasp maps accounted for 79% of the variance (*SD* = 0.18), and a t-test revealed that the correlation was significantly different from zero (*t*(14) = 17.26, p < 0.0001, 95% CI = [0.70 0.89]). The relationship between meaning and saliency (*M* = 0.53, *SD* = 0.15) and graspability and saliency (*M* = 0.51, *SD* = 0.17) accounted for less variance, though both relationships differed from zero (meaning and saliency: *t*(14) = 14.06, *p* < 0.001, 95% CI = [0.45 0.61]; graspability and saliency: *t*(14) = 11.83, p < 0.001, 95% CI = [0.42 0.61]). We can conclude that meaning and graspability overlap considerably, while saliency overlaps to a lesser extent with meaning and graspability, respectively. Graspability and saliency overlapped the least, likely because the edges that define object boundaries—which would be relevant for graspability—do not primarily drive image salience.

Table 1. *Descriptive Statistics for Correlations (R^2) between Maps used in Experiment 1*

	R^2	
Map Comparison	*M*	*SD*
Meaning vs. Graspability	0.794	0.178

Meaning vs. Saliency	0.532	0.147
Graspability vs. Saliency	0.513	0.168

Apparatus

Eye movements were recorded with an SR Research EyeLink 1000+ tower mount eyetracker (spatial resolution 0.01) at a sampling rate of 1000 Hz. Subjects sat 85 cm away from a 21” monitor such that scenes subtended approximately 33° x 25° visual angle. Head movements were minimized using a chin and forehead rest integrated with the eyetracker’s tower mount. Although viewing was binocular, eye movements were recorded from the right eye only. The experiment was controlled using SR Research Experiment Builder software. Audio was recorded digitally at a rate of 48 kHz using a Roland Rubix 22 USB audio interface and a Shure SM86 cardioid condenser microphone.

Procedure

A calibration procedure was conducted at the beginning of each session to map eye position to screen coordinates. Successful calibration required an average error of less than 0.49° and a maximum error below 0.99°. Fixations and saccades were parsed with EyeLink’s standard algorithm using velocity and acceleration thresholds (30°/s and 9500°/s^2; SR Research, 2017).

After successful calibration, subjects received the following instruction: “In this experiment, you will see a series of scenes. In each scene, think of the average person. Describe what the average person would be inclined to do in the scene. You will have 30 seconds to respond.” The instruction was followed by three practice trials that allowed subjects to familiarize themselves with the task and the duration of the response window. Subjects pressed any button on a button box to advance throughout the task.

The task instruction was repeated before subjects began the experimental block (Figure 2a). Within the block, each subject received a unique pseudo-random trial order that prevented two scenes of the same type (e.g., living room) from occurring consecutively. A trial proceeded as follows. First, a five-point fixation array was displayed to check calibration (Figure 2b). The subject fixated the center cross and the experimenter pressed a key to begin the trial if the fixation was stable, otherwise the experimenter reran the calibration procedure. The scene was then shown for a period of 30 seconds, during which time eye-movements and audio were simultaneously recorded (Figure 2c). Finally, after 30 seconds elapsed, subjects were instructed to press a button to proceed to the next trial (Figure 2d). The trial procedure repeated until all 30 trials were complete.

Eye movement data were imported offline into Matlab using the Visual EDF2ASC tool packaged with SR Research DataViewer software. The first fixation was excluded from analysis, as were saccade outliers (amplitude > 20°).

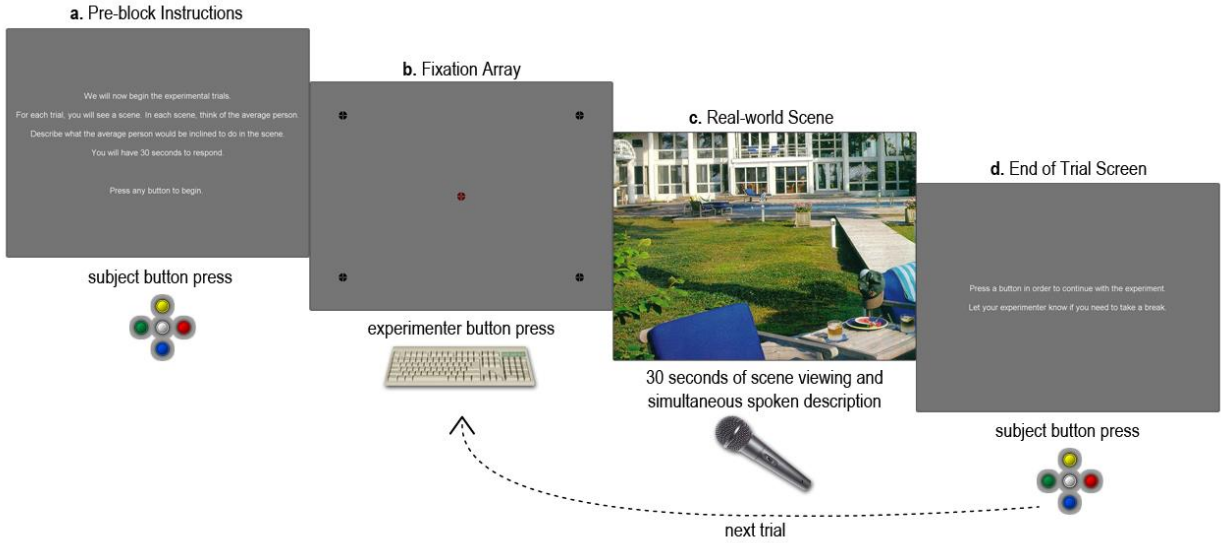


Figure 2. Trial procedure schematic. (a) Task instructions were reiterated to subjects prior to beginning the experimental trials. (b) A five point fixation array was used to assess calibration quality. (c) The real-world scene was shown for 30 seconds. Eye-movements and speech were recorded for the duration of the viewing period. (d) Subjects were instructed to press a button to initiate the next trial. After pressing the button, the trial procedure repeated (from b).

The onset and offset of speech was automatically identified from recordings using an algorithm developed in-house and implemented in Matlab. First, the absolute value of the waveform was computed and the signal was normalized to a minimum value of 0 and a maximum of 1. High frequency noise was omitted by smoothing the waveform using a running average computed over 6000 samples. Second, periods of the audio that contained speech were identified using a signal threshold (0.10 by default). Periods of silence that were shorter than 200 ms in duration were treated as speech. The smoothing and signal threshold parameters were adjusted to accommodate individual variation in speech as needed. Using this procedure, the onset of the first period of the waveform to contain speech and the offset of the final period of the waveform containing speech were identified with high accuracy: onset and offset identifications between a hand coder (the first author) and the algorithm were highly correlated for 90 randomly selected recordings (onsets: $r(88) = 0.903$, $p < .0001$; offsets: $r(88) = 0.985$, $p < .0001$).

Results

Speech

On average, speakers waited more than two seconds into the trial to begin their descriptions (*M* = 2,589 ms, *SD* = 1,309 ms; see Table 2), nearly one second longer than Griffin and Bock's speakers waited before speaking (*M* = 1,686 ms), and slightly longer than Gleitman et al.'s speakers required to produce either active (*M* = 2,076 ms) or passive (*M* = 2,324 ms) single sentence descriptions. Although we recorded speech for the duration of the trial (30 s), speakers finished talking 25,932 ms into the trial on average (*SD* = 5,726), and overall their descriptions required 78% of the time allotted (*M* = 23, 343 ms *SD* = 6,200 ms). See below for a participant's description of possible actions in the scene shown in Figure 1a:

> "Um, sit down in the chair and drink the drink on the table, or eat some of the watermelon, uh, go down the pathway and go for a swim, uh, sit in the other chair by the pool and tan, um, take a nap in the sun, swim laps, uh, go for another snack, uh, dry off, mm."

Table 2. *Descriptive Statistics for Onset and Offset of Speech in Experiment 1*

	Time into trial (ms)		
Measure	*M*	*SD*	*N*
Speech onset	2,589.41	1,308.79	450
Speech offset	25,932.08	5,726.47	450
Duration	23,342.66	6,199.93	450

Attention: Scene-Level Correlations

We correlated each feature map (graspability, meaning, saliency) and attention maps that were empirically derived from viewer fixations in order to determine the degree to which each feature guides visual attention (Figure 3). Squared linear and semipartial correlations (R^2) were computed for each of the 15 scenes, and the relationship between each feature map and visual attention was analyzed using paired t-tests. Cohen's *d* was computed to estimate effect size, and values were interpreted as small (*d* = 0.2 - 0.49), medium (*d* = 0.5 - 0.79), or large (*d* = 0.8+) following Cohen (1988).

Linear correlations.

Meaning explained 53% of the variance in attention on average (M = 0.53, SD = 0.13), while graspability explained 47% of the variance (M = 0.47, SD = 0.17; Figure 3a), and the advantage of meaning over graspability in accounting for variation in attention maps was significant (t(14) = 2.75, p = 0.02, 95% CI = [0.01 0.11], d = 0.39, d 95% CI = [0.09 0.68]). Image salience explained 37% of the variance (M = 0.37, SD = 0.13). Consistent with our previous work (Henderson & Hayes, 2017; 2018; Henderson et al., 2018), there was a reliable advantage of meaning over salience with a large effect size (t(14) = 5.84, p < .0001, 95% CI = [0.10 0.22], d = 1.24, d 95% CI = [0.66 1.82]). Graspability accounted for visual attention better than did image salience, and the effect size was medium (t(14) = 2.63, p = 0.02, 95% CI = [0.02 0.18], d = 0.64, d 95% CI = [0.09 1.20]).

Semipartial correlations.

We subsequently partialed out the shared variance explained by each feature map because the maps were correlated with one another (Table 1). When shared variance explained by salience was accounted for, meaning explained 18% of the remaining variance (M = 0.18, SD = 0.10) and salience accounted for only 2% of the variance (M = 0.02, SD = 0.02) after the contribution of meaning was accounted for, and the effect size was large (t(14) = 5.84, p < 0.001, 95% CI = [0.10 0.22], d = 2.51, d 95% CI = [0.72 4.30]). Similarly, when the variance explained by salience was partialed out, graspability explained 14% of the variance (M = 0.14, SD = 0.10) and salience only explained 4% of the variance (M = 0.04, SD = 0.06) after graspability was partialed out; this effect size was also large (t(14) = 2.63, p = 0.02, 95% CI = [0.02 0.18], d = 1.25, d 95% CI = [-0.05 2.55]). Meaning accounted for 8% of the remaining variance (M = 0.08, SD = 0.08) after the contribution of graspability was partialed out, and graspability explained only 2% of the variance (M = 0.02, SD = 0.02) after the variance explained by meaning was partialed out, and a large effect size was found (t(14) = 2.74, p = 0.02, 95% CI = [0.1 0.11], d = 1.07, d 95% CI = [0.07 2.08]). The unique variance explained by graspability, however, differed from zero (t(14) = 3.43, p = .004, 95% CI = [0.006 0.03])[1].

[1] The unique variance explained by salience when meaning was partialled out also differed from zero (t(14) = 3.47, p = 0.004, 95% CI = [0.008 0.03]).

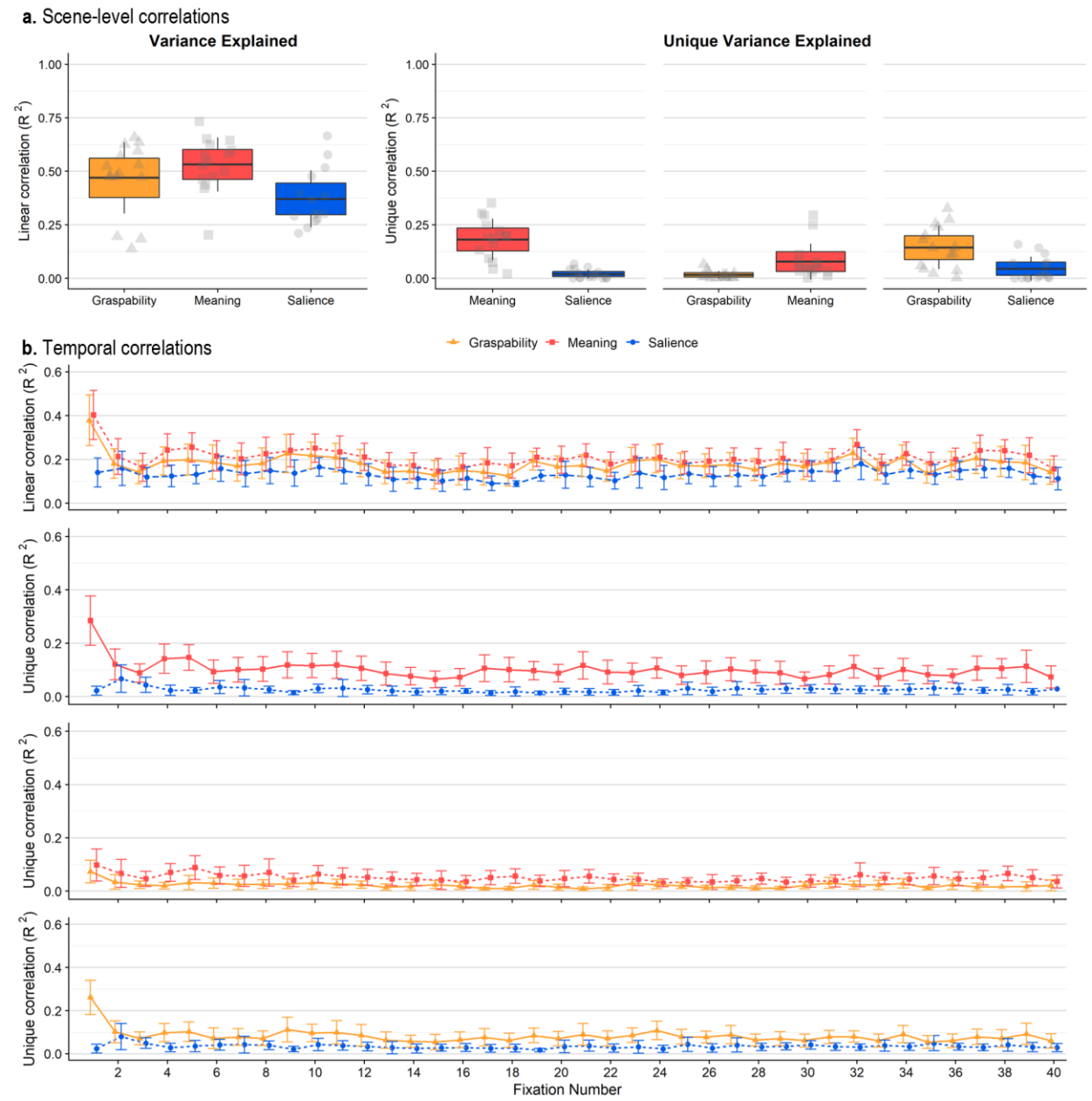


Figure 3. a) Box plots showing linear correlations (left) and semipartial correlations (right) between feature maps (graspability, meaning, saliency) and attention maps. The scatter box plots show the corresponding grand mean (black horizontal line), 95% confidence intervals (colored box), and 1 standard deviation (black vertical line) for meaning (red box, square points), graspability (orange box, triangular points), and salience (blue box, circular points) across 15 scenes. b) Line graphs showing linear correlations (row 1) and semipartial correlations (rows 2-4) between feature maps and attention maps for each fixation (1-40). Error bars indicate 95% confidence intervals.

To summarize, both meaning and graspability explained variance in attention better than salience did, and ultimately meaning outperformed graspability. Effect sizes for all comparisons of semipartial correlations were large (Cohen's $d > .8$).

Attention: Initial Fixations

To test for differences in attentional guidance at the beginning of the trial, we correlated each feature map (graspability, meaning, saliency) with attention maps at each fixation (Figure 3b). Because our primary interest was in what features influenced attention while speakers planned their descriptions, we focused our analysis on the first three fixations in each trial. Squared linear and semipartial correlations (R^2) were computed for each of these fixations, and the relationship between each feature map and visual attention was analyzed using paired t-tests.

Linear correlations.

During the first three fixations, on average meaning accounted for 40% (*M* = 0.40, *SD* = 0.22), 21% (*M* = 0.21, *SD* = 0.16), and 16% (*M* = 0.16, *SD* = 0.13) of the variance in attention. Graspability accounted for 38% of the variance during the first fixation (*M* = 0.38, *SD* = 0.23), 18% of the variance during the second fixation (*M* = 0.18, *SD* = 0.13), and 14% of the variance at the third fixation (*M* = 0.14, *SD* = 0.10). On average, saliency accounted for the smallest percentage of the variance for all three fixations, explaining 14% of the variance at the first fixation (*M* = 0.14, *SD* = 0.13), 16% of the variance at the second fixation (*M* = 0.16, *SD* = 0.15), and 12% of the variance during the third fixation (*M* = 0.12, *SD* = 0.09).

Meaning held a significant advantage over saliency during the first fixation with a large effect size (*t*(14) = 5.29, *p* = 0.0001, 95% CI = [0.16 0.37], *d* = 1.36, *d* 95% CI = [0.63 2.08]) but not for the second or third fixations (2: *t*(14) = 1.18, *p* = 0.26, 95% CI = [-0.04 0.15], *d* = 0.34, *d* 95% CI = [-0.27 0.95]; 3: *t*(14) = 1.62, *p* = 0.13, 95% CI = [-0.01 0.10], *d* = 0.39, *d* 95% CI = [-0.12 0.91]). In contrast, the difference in average variance explained by meaning compared to graspability was not significant for any of the three fixations, and effect sizes were small or medium (1: *t*(14) = 0.58, *p* = 0.57, 95% CI = [-0.07 0.12], *d* = 0.11, *d* 95% CI = [-0.28 0.50]; 2: *t*(14) = 0.99, *p* = 0.34, 95% CI = [-0.04 0.10], *d* = 0.21, *d* 95% CI = [-0.23 0.66]; 3: *t*(14) = 1.37, *p* = 0.19, 95% CI = [-0.01 0.06], *d* = 0.18, *d* 95% CI = [-0.09 0.45]). Graspability held a large advantage over saliency for the first fixation (*t*(14) = 5.04, *p* = 0.0002, 95% CI = [0.14 0.34], *d* = 1.17, *d* 95% CI = [0.55 1.78]), and did not explain the variance in attention better than saliency for the second and third fixations (2: *t*(14) = 0.47, *p* = 0.64, 95% CI = [-0.08 0.12], *d* = 0.15, *d* 95% CI = [-0.51 0.81]; 3: *t*(14) = 0.96, *p* = 0.35, 95% CI = [-0.03 0.0.07], *d* = 0.23, *d* 95% CI = [-0.27 0.73]), and the latter two effect size estimates were small.

Semipartial correlations.

As in the analysis by scene, we accounted for the shared variance explained by two features using semipartial correlations, which allowed us to determine how well one feature accounted for variance in attention. We will consider the analysis for each feature pair in turn.

When the variance explained by both meaning and salience was partialled out, during the first fixation meaning explained 28% of the variance on average (M = 0.28, SD = 0.18), 12% of the variance at the second fixation (M = 0.12, SD = 0.11), and 9% of the variance at the third fixation (M = 0.09, SD = 0.07). Salience accounted for 2% of the variance on average at the first fixation (M = 0.02, SD = 0.03), 7% of the variance at the second fixation (M = 0.07, SD = 0.10), and 4% of the variance at the third fixation (M = 0.04, SD = 0.06). The advantage of meaning over salience was significant for the first fixation ($t(14) = 5.30$, $p < 0.001$, 95% CI = [0.16 0.37], d = 2.09, d 95% CI = [0.65 3.53]), carrying a large effect size, not significant for the second and third fixations (2: $t(14) = 1.18$, $p = 0.26$, 95% CI = [-0.04 0.15], d = 0.49, d 95% CI = [-0.41 1.40]; 3: $t(14) = 1.62$, $p = 0.13$, 95% CI = [-0.01 0.10], d = 0.71, d 95% CI = [-0.30 1.72]).

After partialling out the variance explained by both meaning and graspability, on average meaning accounted for 10% of the variance during the first fixation (M = 0.10, SD = 0.12), and 7% of the variance during the second fixation (M = 0.07, SD = 0.10) and 5% during the third fixation (M = 0.05, SD = 0.06). Graspability accounted for 7% of the variance at the first fixation (M = 0.07, SD = 0.09), 3% at the second fixation (M = 0.03, SD = 0.05), and 2% during the third fixation (M = 0.02, SD = 0.03). The difference in means was not significant at any time steps, and all effect sizes were small or medium (1: $t(14) = 0.58$, $p = 0.52$, 95% CI = [-0.07 0.12], d = 0.24, d 95% CI = [-0.63 1.11]; 2: $t(14) = 0.98$, $p = 0.34$, 95% CI = [-0.04 0.10], d = 0.39, d 95% CI = [-0.46 1.24]; 3: $t(14) = 1.36$, $p = 0.19$, 95% CI = [-0.01 0.06], d = 0.50, d 95% CI = [-0.29 1.28]).

Finally, when the shared variance that both graspability and salience accounted for was partialled out, graspability explained on average 26% of the variance at the first fixation (M = 0.26, SD = 0.16), 10% at the second fixation (M = 0.10, SD = 0.10), and 7% for the third fixation (M = 0.07, SD = 0.06), while saliency accounted for 2% of the variance on average during the first fixation (M = 0.02, SD = 0.04), 8% of the variance at the second fixation (M = 0.08, SD = 0.12), and 5% at the third (M = 0.05, SD = 0.05). The difference at the first fixation was significant and the effect size was large ($t(14) = 5.05$, $p = 0.03$, 95% CI = [0.14 0.34], d = 2.31, d 95% CI = [0.52 4.10]), whereas differences at the second and third fixations were not significant (2: $t(14) = 0.47$, $p = 0.27$, 95% CI = [-0.08 0.12], d = 0.20, d 95% CI = [-0.67 1.06]; 3: $t(14)$ =

0.96, p = 0.11, 95% CI = [-0.03 0.07], d = 0.42, d 95% CI = [-0.52 1.35]), and effect sizes for both of the latter differences were small.

In sum, when we correlated attention maps constructed using individual fixations with feature maps and focused on the earliest (most informative) fixations, meaning outperformed salience at the first fixation, and the effect size was large. Graspability also outperformed salience at the first fixation. However, meaning did not significantly outperform graspability, nor did any of the comparisons differ significantly beyond the first fixation. This was true across both linear and semipartial correlations. The results are partially consistent with the scene-level analyses in that meaning and graspability both outperformed salience, but in the scene-level analyses meaning significantly outperformed graspability. While the numerical differences followed the same pattern when we analyzed early fixations, the effects were less robust.

Discussion

In Experiment 1, the duration of the pre-speech interval was over two seconds, only slightly longer than the intervals Griffin and Bock (2000) and Gleitman et al. (2007) reported for single sentence descriptions. However, descriptions in our task went on for over twenty seconds and included multiple clauses. We can conclude that speakers do not plan the entirety of the description during this time, but rather engage in macroplanning to decide where to begin and what to talk about in the scene in a very broad sense. Then they engage in incremental planning as the description unfolds, allocating attention to regions of the scene that they wish to talk about.

Meaning explained visual attention better than image salience, which is consistent with our previous work (Henderson & Hayes, 2017; 2018; Henderson et al., 2018; Peacock et al., 2019a). Novel to this experiment, we found that graspability explained visual attention better than image salience did in the scene-level analysis. We were primarily interested in whether graspability preferentially guided attention over meaning because it is task-relevant. We found the opposite pattern: meaning explained visual attention better than graspability did in the scene-level analysis, though a small amount of the unique variance was explained by graspability. Graspability outperformed saliency, which suggests that the information tapped by graspability was more relevant than image salience. When we analyzed only the first three fixations, we found similar numerical patterns in the data, but the effects were only significant with respect to the advantage of meaning over salience.

Our results again confirm that the advantage of meaning over salience is robust. To our surprise, graspability did not account for attention better or even as well as than meaning did. Finally, although graspability could not compete with meaning, it did outperform image salience.

Taken together, the findings suggest that graspability may tap a subcomponent of meaning, but does not capture enough semantic information to account for semantic guidance of attention in scenes. We conducted a second experiment using greater power to further address whether graspability influences visual attention.

Experiment 2

We conducted a second experiment to replicate the results of Experiment 1. In Experiment 2, 40 subjects described 20 scenes. The methods were identical to those used in Experiment 1, with the exceptions mentioned below.

Subjects

Forty-eight undergraduate students enrolled at the University of California, Davis participated in exchange for course credit. Data from one subject were excluded from analysis due to a behavioral issue that was unrelated to the task. Data from an additional 3 subjects were excluded due to an audio recording equipment failure, and another 5 subjects who could not be accurately eye tracked. Data from the remaining 40 subjects were analyzed.

Stimuli

Twenty digitized (1024x768) and luminance-matched photographs of real-world scenes depicting indoor and outdoor environments were presented. Three scenes depicted outdoor environments (2 street views), and 17 depicted indoor environments (2 kitchens, 3 living rooms, 3 desk areas). One scene showed people in the background of the image, but people were not present in any of the other 19 scenes. Meaning, graspability, and saliency maps for these scenes shared similar overlap to those tested in Experiment 1 (Table 3), and all correlations were significantly different from zero (all *ps* < .001).

Table 3. *Descriptive Statistics for Correlations (R^2) between Maps used in Experiment 2*

	R^2	
Map Comparison	*M*	*SD*
Meaning vs. Graspability	0.811	0.157
Meaning vs. Saliency	0.547	0.144
Graspability vs. Saliency	0.526	0.154

Apparatus. Subjects sat 83 cm away from a 24.5” monitor such that scenes subtended approximately 26° x 19° visual angle at a resolution of 1024 x 768 pixels, presented in 4:3

aspect ratio. Audio was recorded digitally at a rate of 48 kHz using a Shure SM86 cardioid condenser microphone. Data were collected on two separate systems that were identical with two exceptions. First, audio input was preamplified using a Roland Rubix 22 audio interface on one system, and was not preamplified on the other. Second, the operating system for the subject computer in one system was Windows 10, and Windows 7 on the other.

Procedure

The experimental procedure was identical to that of Experiment 1, except that there were 20 experimental trials in Experiment 2.

Results

Speech

On average, the pre-speech interval was over two seconds long (*M* = 2,092 ms, *SD* = 1,568 ms; see Table 4), which is comparable to the pre-speech interval duration we found in Experiment 1. The duration is also consistent with the pre-speech intervals reported by Griffin and Bock (2000) and Gleitman et al. (2007) despite the fact that speakers' utterances were longer and more complex in our task. Speakers finished talking on average 29,297 ms into the trial (*SD* = 1,514 ms), and used over 90% of the allotted time (*M* = 27,205 ms, *SD* = 2,354 ms).

Table 4. *Descriptive Statistics for Onset and Offset of Speech in Experiment 2*

	Time into trial (ms)		
Measure	*M*	*SD*	*N*
Speech onset	2,091.97	1,568.17	800
Speech offset	29,296.92	1,514.91	800
Duration	27,204.95	2,353.57	800

Attention: Scene-Level Correlations

Attention maps were related to feature maps (meaning, saliency, graspability) in the same manner as in Experiment 1. We correlated each feature map (graspability, meaning, saliency) and attention maps that were empirically derived from viewer fixations in order to determine the degree to which each feature guides visual attention (Figure 4). Squared linear and semipartial correlations (R^2) were computed for each of the 20 scenes, and the relationship between each feature map and visual attention was analyzed using paired t-tests.

Linear correlations.

Meaning explained 49% of the variance in attention on average (M = 0.49, SD = 0.13, while graspability explained 43% of the variance (M = 0.43, SD = 0.16; see Figure 4a). Image salience explained only 30% of the variance (M = 0.30, SD = 0.11). Once again we replicated the advantage of meaning over salience, which was a large effect (t(19) = 7.18, p < .0001, 95% CI = [0.13 0.24], d = 1.50, d 95% CI = [0.88 2.12]). Graspability again explained visual attention better than image salience did, and this was also a large effect (t(19) = 4.08, p < 0.001, 95% CI = [0.06 0.19], d = 0.85, d 95% CI = [0.36 1.34]). The advantage of meaning over graspability in accounting for the variance in attention maps was significant, though the effect size for this comparison was small (t(19) = 2.59, p = 0.02, 95% CI = [0.01 0.11], d = 0.40, d 95% CI = [0.08 0.73]).

Semipartial correlations.

We subsequently partialed out the shared variance explained by each feature map because the maps were correlated with one another (Table 3). When shared variance explained by salience was accounted for, meaning explained 20% of the remaining variance (M = 0.20, SD = 0.11) and salience accounted for only 1% of the variance (M = 0.01, SD = 0.01) after the contribution of meaning was accounted for, and the effect size for this comparison was large (t(19) = 7.19, p < 0.001, 95% CI = [0.13 0.24], d = 2.67, d 95% CI = [1.06 4.28]). Similarly, when the variance explained by salience was partialed out, graspability explained 16% of the variance (M = 0.16, SD = 0.10) and salience only explained 3% of the variance (M = 0.03, SD = 0.05) after graspability was partialed out, and the size of this effect was large (t(19) = 4.09, p < 0.001, 95% CI = [0.06 0.19], d = 1.66, d 95% CI = [0.39 2.92]). Meaning accounted for 8% of the remaining variance (M = 0.08, SD = 0.11) after the contribution of graspability was partialed out, and graspability explained only 2% of the variance (M = 0.02, SD = 0.02) after the variance explained by meaning was partialed out, and this comparison again carried a medium effect size (t(19) = 2.59, p = 0.02, 95% CI = [0.01 0.11], d = 0.80, d 95% CI = [0.08 1.52]). However, the unique variance captured by graspability differed from zero (t(19) = 3.98, p < .001, 95% CI = [0.009 0.03])[2].

[2] The unique variance explained by salience when meaning was partialled out was also different from zero (t(19) = 4.24, p < .001, 95% CI = [0.005 0.02]).

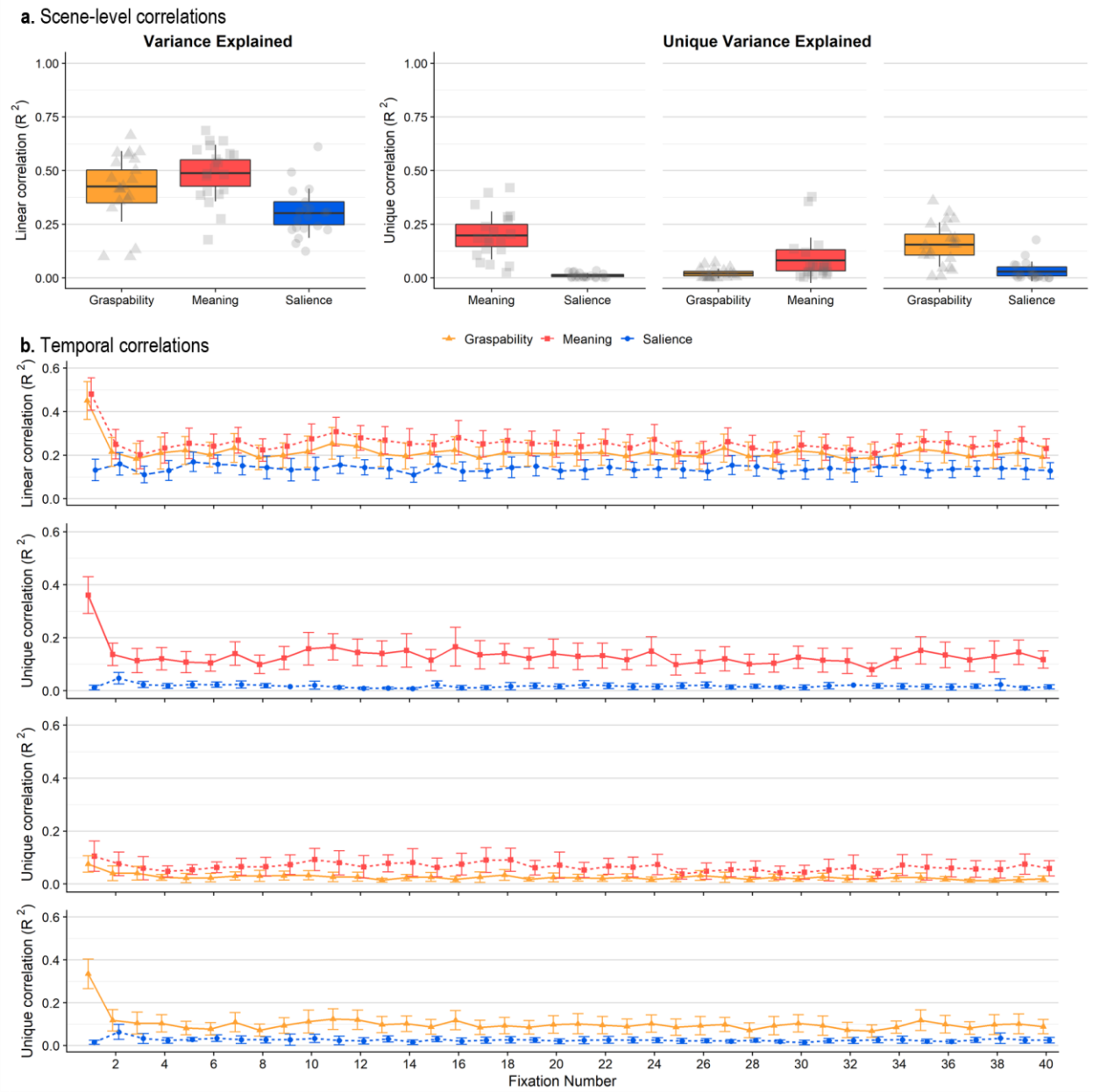


Figure 4. a) Box plots showing linear correlations (left) and semipartial correlations (right) between feature maps (graspability, meaning, saliency) and attention maps for each scene. The scatter box plots show the corresponding grand mean (black horizontal line), 95% confidence intervals (colored box), and 1 standard deviation (black vertical line) for meaning (red box, square points), graspability (orange box, triangular points), and salience (blue box, circular points) across 20 scenes. b) Line graphs showing linear correlations (row 1) and semipartial correlations (rows 2-4) between feature maps and attention maps for each fixation (1-40). Error bars indicate 95% confidence intervals.

Consistent with the same analysis in Experiment 1, both meaning and graspability explained variance in attention better than image salience did, and effect sizes were large for these comparisons. Meaning significantly outperformed graspability again, though the size of the effect was smaller, and graspability did capture a small amount (but not zero) of the unique variance in attention.

Attention: Initial Fixations

To test for changes in attentional guidance at the beginning of the trial, we again correlated each feature map (graspability, meaning, saliency) with attention maps at each fixation, and focused on the first three fixations in our analysis (Figure 4b). Squared linear and semipartial correlations (R^2) were computed for each fixation, and the relationship between each feature map and visual attention was analyzed using paired t-tests.

Linear correlations.

During early fixations, on average meaning accounted for 48% of the variance in attention (*M* = 0.48, *SD* = 0.17), 25% of the variance during the second fixation (*M* = 0.25, *SD* = 0.16), and 21% during the third fixation (*M* = 0.20, *SD* = 0.15). Graspability accounted for 45% of the variance during the first fixation (*M* = 0.45, *SD* = 0.20), 21% of the variance at the second fixation (*M* = 0.21, *SD* = 0.15), and 18% of the variance at the third fixation (*M* = 0.18, *SD* = 0.16). On average, saliency accounted for the smallest percentage of the variance at all three fixations, explaining 13% of the variance at the first fixation (*M* = 0.13, *SD* = 0.11), and 16% of the variance at the second fixation (*M* = 0.16, *SD* = 0.12), and 11% at the third fixation (*M* = 0.11, *SD* = 0.09).

Meaning held a significant advantage over saliency during all three fixations, and all effects were large or medium (1: $t(19) = 9.69$, $p < 0.001$, 95% CI = [0.27 0.42], $d = 2.37$, d 95% CI = [1.40 3.33]; 2: $t(19) = 3.47$, $p = 0.003$, 95% CI = [0.04 0.14], $d = 0.63$, d 95% CI = [0.23 1.03]; 3: $t(19) = 3.49$, $p = 0.002$, 95% CI = [0.04 0.14], $d = 0.69$, d 95% CI = [0.24 1.13]). The difference in the average variance explained by meaning compared to graspability was not significant during the first three fixations and effect sizes were small (1: $t(19) = 0.82$, $p = 0.42$, 95% CI = [-0.05 0.11], $d = 0.16$, d 95% CI = [-0.24 0.55]; 2: $t(19) = 1.21$, $p = 0.24$, 95% CI = [-0.03 0.10], $d = 0.23$, d 95% CI = [-0.16 0.62]; 3: $t(19) = 0.65$, $p = 0.52$, 95% CI = [-0.04 0.08], $d = 0.12$, d 95% CI = [-0.25 0.49]). Graspability explained the variance in attention better than saliency did for the first and third fixations, carrying large and immediate effects, respectively (1: $t(19) = 8.27$, $p < 0.001$, 95% CI = [0.24 0.40], $d = 1.85$, d 95% CI = [1.11 2.60]; 3: $t(19) = 2.21$, $p = 0.04$, 95% CI = [0.004 0.14], $d = 0.53$, d 95% CI = [0.01 1.04]), but the difference was not

significant during the second fixation, and the effect size was small ($t(19) = 1.55$, $p = 0.14$, 95% CI = [-0.02 0.13], $d = 0.40$, d 95% CI = [-0.14 0.94]).

Semipartial correlations.

As in the analysis by scene, we accounted for the shared variance explained by two features using semipartial correlations, which allowed us to determine how well one feature accounted for variance in attention. We will consider the analysis for each feature pair in turn.

When the variance explained by both meaning and salience was partialled out, meaning explained 36% of the variance on average during the first ($M = 0.36$, $SD = 0.16$), 14% during the second fixation ($M = 0.14$, $SD = 0.10$), and 11% of the variance at the third fixation ($M = 0.11$, $SD = 0.10$). Salience accounted for 1% of the variance on average at the first fixation ($M = 0.01$, $SD = 0.02$), 5% of the variance during the second fixation ($M = 0.05$, $SD = 0.05$), and 2% during the third fixation ($M = 0.02$, $SD = 0.03$). As was the case for the linear correlations, the advantage of meaning over salience was significant for all fixations, and all effect sizes were large (1: $t(19) = 9.69$, $p < 0.001$, 95% CI = [0.27 0.42], $d = 3.24$, d 95% CI = [1.55 4.93]; 2: $t(19) = 3.48$, $p = 0.003$, 95% CI = [0.04 0.14], $d = 1.18$, d 95% CI = [0.29 2.07]; 3: $t(19) = 3.48$, $p = 0.002$, 95% CI = [0.04 0.14], $d = 1.26$, d 95% CI = [0.28 2.24]).

After partialling out the variance explained by both meaning and graspability, on average meaning accounted for 11% of the variance during the first fixation ($M = 0.11$, $SD = 0.13$), 8% of the variance during the second fixation ($M = 0.08$, $SD = 0.10$), and 6% of the variance during the third ($M = 0.06$, $SD = 0.10$). Graspability accounted for 8% of the variance at the first fixation ($M = 0.08$, $SD = 0.07$) and 4% at the second and third fixations (2: $M = 0.04$, $SD = 0.06$; 3: $M = 0.04$, $SD = 0.06$). The difference in means was not significant for any fixations, and all effect sizes were small (1: $t(19) = 0.81$, $p = 0.43$, 95% CI = [-0.05 0.10], $d = 0.28$, d 95% CI = [-0.44 1.01]; 2: $t(19) = 1.21$, $p = 0.24$, 95% CI = [-0.03 0.10], $d = 0.41$, d 95% CI = [-0.31 1.14]; 3: $t(19) = 0.65$, $p = 0.52$, 95% CI = [-0.04 0.08], $d = 0.23$, d 95% CI = [-0.48 0.93]).

Lastly, when the shared variance that both graspability and saliency accounted for was partialled out, graspability explained on average 33% of the variance at the first fixation ($M = 0.33$, $SD = 0.16$), 12% during the second fixation ($M = 0.12$, $SD = 0.11$), and 11% at the third fixation ($M = 0.11$, $SD = 0.12$). Saliency accounted for 2% of the variance on average during the first fixation ($M = 0.02$, $SD = 0.02$), 6% at the second fixation ($M = 0.06$, $SD = 0.08$), and 3% at the third ($M = 0.03$, $SD = 0.05$). The difference at the first fixation was significant with a large effect size ($t(19) = 8.26$, $p < 0.001$, 95% CI = [0.24 0.40], $d = 3.33$, d 95% CI = [1.24 5.42]) and the differences at the second and third fixations were not significant, but carried medium and

large effect sizes (2: $t(19) = 1.55$, $p = 0.27$, 95% CI = [-0.02 0.13], $d = 0.55$, d 95% CI = [-0.22 1.33]; 3: $t(19) = 2.22$, $p = 0.85$, 95% CI = [0.004 0.14], $d = 0.82$, d 95% CI = [-0.05 1.68]).

In sum, when we analyzed the first three fixations, meaning outperformed salience, and effect sizes for all comparisons were large. Meaning numerically outperformed graspability, but the advantage of meaning was not significant during early viewing in either analysis. For both linear and semipartial correlations, we found a consistent and significant advantage of graspability over salience at all fixations, bearing medium sized effects. These findings replicate what we found in Experiment 1.

Discussion

The onset of speech in Experiment 2 was comparable to what we found in Experiment 1, and to the production literature (Griffin & Bock, 2000; Gleitman et al., 2007): the duration of the pre-speech interval just exceeded two seconds, and speakers took almost all of the remaining time to describe the scene's possible actions. The results provide further evidence that speakers do not plan an entire utterance before speaking, but instead form a general strategy, then plan and execute the utterance incrementally.

Consistent with our previous work and with the findings from Experiment 1, meaning explained visual attention better than image salience (Henderson & Hayes, 2017; 2018; Henderson et al., 2018; Peacock et al., 2019a). We found that graspability explained visual attention better than image salience did in all analyses as well. Meaning explained visual attention better than graspability did in all scene-level analyses, but not during early viewing. This general pattern of results was observed over both experiments.

Experiment 3

Overall, the results of Experiments 1 and 2 suggest that, counter to our initial predictions and previous literature, the relationship between graspability and attention does not compete with that between meaning and attention. However, it could also be the case that graspability played a lesser role because very few of the scenes depicted reachable spaces: Most of the objects that were graspable were too far away from the viewpoint of the photograph for a person, standing at the viewpoint, to reach them, in which graspability may have been rendered less relevant (Josephs & Konkle, 2019; Bonner & Epstein, 2017; 2018; Gomez et al., 2018). Additionally, not all of the scenes depicted graspable objects, regardless of distance from the viewpoint. We conducted a third experiment to determine whether the stimuli used in Experiments 1 and 2 inadvertently put graspability at a disadvantage. In Experiment 3 we tested a new stimulus set intended to emphasize the role of graspability. Each of the 20 scenes contain graspable objects, both in the foreground of the image (within reach of the viewpoint)

and the background, as well as objects that are meaningful but not graspable. In addition to mapping the new scenes for image salience, meaning, and graspability, we introduced reach maps to define the reachable space in each scene and constructed reach-weighted graspability maps to determine whether graspability draws attention specifically for objects that would be within reach.

Methods

In this experiment, 40 subjects viewed a new set of 20 scenes and described what actions they would carry out if they were in the scene. The goal of Experiment 3 was to determine whether graspability would preferentially guide visual attention more than meaning for scenes that depict graspable objects within reach. If the semantic information captured by meaning maps is reducible to grasping affordances for the new scenes, we expect meaning and graspability to explain variance in attention equally well. If grasping affordances are particularly important for objects that are within reach of the viewer, we anticipate reach-weighted grasp maps to explain variance in attention as well as graspability, if not better.

Subjects

Forty-nine undergraduate students enrolled at the University of California, Davis participated in exchange for course credit. Data collection for four subjects was incomplete due to a software error, and therefore data for these subjects were not analyzed. Data from another five subjects who could not be accurately eye tracked were also excluded. Data from the remaining 40 subjects were analyzed.

Stimuli

Twenty digitized (1024x768) and luminance-matched photographs of real-world scenes depicting indoor and outdoor environments were presented. Fifteen of the scenes were photographed by the first and second author. For those 15 scenes, the authors confirmed that objects in the foreground of the scene were within reach of the scene's viewpoint. The remaining scenes were drawn from other studies: 4 from Xu et al. (2014) and 1 from Cullimore, Rehrig, Henderson, and Ferreira (2018). Three scenes depicted outdoor environments (two beach scenes), and 17 depicted indoor environments (eight kitchens, five dining areas). Faces were not present in any of the scenes. Text was removed from each scene using the clone stamp and patch tools in Adobe Photoshop CS4.

Meaning, Grasp, Reach, and Saliency Map Generation

Meaning Maps

Meaning maps for the Experiment 3 stimuli were constructed from ratings in a similar fashion as the maps used in Experiments 1 and 2. Each 1024 x 768 pixel scene was

decomposed into a series of partially overlapping circular patches at fine and coarse spatial scales (Figure 1b&c). The decomposition resulted in 7,500 unique fine-scale patches (93 pixel diameter) and 2,700 unique coarse-scale patches (217 pixel diameter), totaling 10,200 patches across scales. Subjects viewed scene patches presented alongside an image of the scene. A superimposed green circle indicated what region the patch came from to provide context (Figure 1a).

Raters were 124 undergraduates enrolled at UC Davis who participated through Sona. Students received credit toward a course requirement for participating. Subjects were at least 18 years old, had normal or corrected-to-normal vision, and had normal color vision.

Each subject rated 300 random patches extracted from 25 scenes, presented alongside the scene image for context. In each survey, 20 catch patches that showed solid surfaces (e.g., a wall) served as an attention check. Subjects were instructed to rate how meaningful each patch was using a 6-point Likert scale (‘very low’, ‘low’, ‘somewhat low’, ‘somewhat high’, ‘high’, ‘very high’). Prior to rating patches, subjects were given two examples of low-meaning and two examples of high-meaning scene patches in the instructions to ensure that they understood the task. Scene-patch pairs were presented in random order. Ratings from 20 subjects that scored below 85% on the catch patches were excluded. Each unique patch was rated 3 times by 3 independent raters for a total of 30,600 ratings.

Meaning maps were generated from ratings in the same manner as in Experiment 1. Because subjects in the eyetracking task showed a consistent center bias in their fixations, center bias was applied to the maps by downweighting the periphery of the maps.

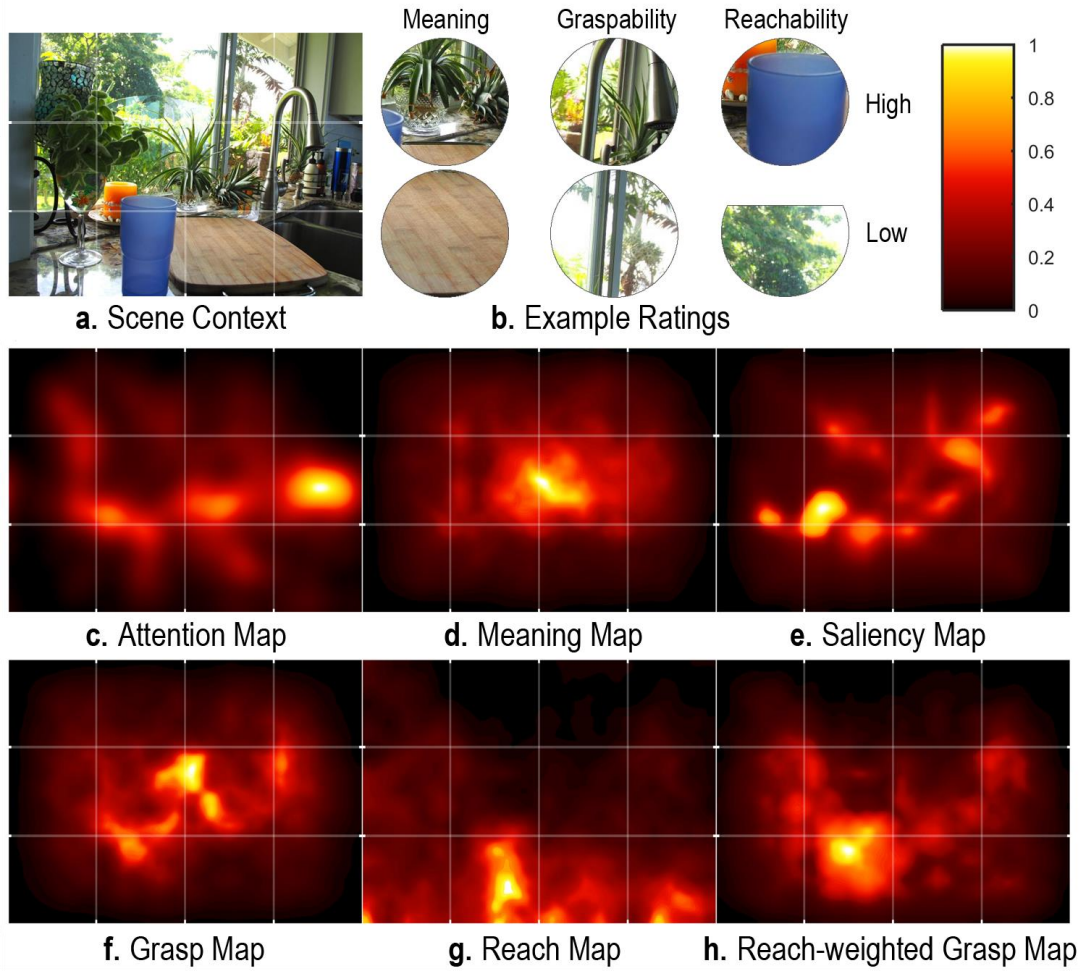


Figure 5. (a) Real-world scene. Raters saw the real-world scene and either a fine- or coarse-scale green circle indicating the origin of the scene patch (see Figure 1 for an example and spatial grids). (b) Examples of scene patches that were rated as low or high with respect to meaning, graspability, and reachability. (c-h) examples of attention (c), meaning (d), saliency (e), grasp (f), reach (g) and reach-weighted grasp (h) maps for the scene shown in (a). The attention map (c) was empirically derived from viewer fixations. The reach-weighted grasp map (h) is the product of element-wise matrix multiplication between the grasp map (f) and reach map (g).

Grasp Maps

Grasp maps for the Experiment 3 stimuli were constructed from ratings in the same way that the Experiment 1 and 2 grasp maps were generated.

Raters were 128 undergraduates enrolled at UC Davis who participated through Sona. Students received credit toward a course requirement for participating. Subjects were at least 18 years old, had normal or corrected-to-normal vision, and had normal color vision.

Each subject rated 300 random patches extracted from the 25 scenes, presented alongside the scene image for context. In each survey, 20 catch patches that showed solid

surfaces (e.g., a wall) served as an attention check. Subjects were instructed to rate how graspable each patch was using a 6-point Likert scale (‘very low’, ‘low’, ‘somewhat low’, ‘somewhat high’, ‘high’, ‘very high’). Prior to rating patches, subjects were given two examples of low-graspability and two examples of high-graspability scene patches in the instructions to ensure that they understood the task. Scene-patch pairs were presented in random order. Each unique patch was rated 3 times by 3 independent raters for a total of 30,600 ratings (see Figure 5b for example of patches rated low and high in graspability). Ratings from 24 subjects that scored below 85% on the catch patches were excluded. Each unique patch was rated 3 times by 3 independent raters for a total of 30,600 ratings.

Because subjects in the eyetracking task showed a consistent center bias in their fixations, center bias was applied to the maps by downweighting the periphery of the maps. See Figure 5b for example patch ratings and Figure 5f for an example grasp map.

Reach Maps

Reach maps were constructed from ratings in the same manner as meaning and grasp maps, with the critical exception that subjects rated each patch on how ‘reachable’ the region of the scene shown in the patch was. In the instructions, we defined ‘reachability’ as how easily they could naturally reach what is shown in the scene patch, using their arms and hands only, if they were standing at the camera’s viewpoint. The remainder of the procedure was identical to the one used to generate meaning and grasp maps.

Raters were 212 undergraduates enrolled at UC Davis who were recruited through Sona (N = 209) or were research assistants who were naive to the research question (N = 3). Subjects were at least 18 years old, had normal or corrected-to-normal vision, and had normal color vision.

Each subject rated 300 random patches extracted from the 25 scenes, presented alongside the scene image for context. In each survey, 20 catch patches that showed solid surfaces (e.g., a wall) served as an attention check. Subjects were instructed to rate how reachable each patch was using a 6-point Likert scale (‘very low’, ‘low’, ‘somewhat low’, ‘somewhat high’, ‘high’, ‘very high’). Prior to rating patches, subjects were given two examples of low-reachability and two examples of high-reachability scene patches in the instructions to ensure that they understood the task. Scene-patch pairs were presented in random order. Ratings from 108 subjects that scored below 85% on the catch patches were excluded. Each unique patch was rated 3 times by 3 independent raters for a total of 30,600 ratings.

Reach maps were generated from ratings in the same manner as meaning and graspability maps were in Experiment 1. No center bias adjustment was applied (see Reach-weighted Grasp Maps). See Figure 5b for example patch ratings and Figure 5g for an example reach map.

Reach-weighted Grasp Maps

To determine whether the influence of graspability on attention is amplified for the region of the scene that is within reach, we created reach-weighted grasp maps. Reach-weighted grasp maps were created by converting grasp maps and reach maps to grayscale image matrices in Matlab, then performing element-wise matrix multiplication between the two maps. Multiplication both weighted the grasp map by the reach map and carried over the peripheral downweighting from the grasp map to the resulting reach-weighted grasp map. See Figure 5h for an example reach-weighted grasp map.

Meaning, grasp, reach-weighted grasp, and saliency maps for these scenes shared some overlap (Table 3), and all correlations were significantly different from zero (all *ps* < .001). Note that the new meaning and grasp maps were not only highly correlated with one another on average (*M* = 0.77), there was also little variance in the relationship across scenes (*SD* = 0.07).

Table 5. *Descriptive Statistics for Correlations (R^2) between Maps used in Experiment 3*

	R^2	
Map Comparison	*M*	*SD*
Meaning vs. Grasp	0.767	0.067
Meaning vs. Reach-weighted Grasp	0.282	0.182
Meaning vs. Saliency	0.490	0.121
Reach-weighted Grasp vs. Saliency	0.192	0.112
Grasp vs. Reach-weighted Grasp	0.516	0.148
Grasp vs. Saliency	0.491	0.147

Apparatus. Subjects sat 83 cm away from a 24.5” monitor such that scenes subtended approximately 26° x 19° visual angle at a resolution of 1024 x 768 pixels, presented in 4:3 aspect ratio. The experiment was controlled using SR Research Experiment Builder software.

Though viewing was binocular, eye movements were recorded from the right eye only. Audio was recorded digitally at a rate of 48 kHz using a Shure SM86 cardioid condenser microphone and was preamplified using an InnoGear IG101 phantom power preamplifier.

Procedure

The experimental procedure was identical to that of Experiments 1 and 2, with the following exception. Rather than instructing subjects to think of the average person, subjects were instructed as follows: "In this experiment, you will see a series of scenes. For each scene, describe what you would do in the scene. You will have 30 seconds to respond."[3] Nearly half of audio files (N = 341) foiled the speech onset and offset detection algorithm due to background noise (e.g., sniffling, coughing) or whispering. For these files, speech onset and offset were hand-coded in Praat by the first author.

Results

Speech

Speakers waited just under three seconds into the trial on average to begin their descriptions (*M* = 2,699 ms, *SD* = 1,811 ms; see Table 6), which was longer than participants waited in Experiments 1 and 2 to begin speaking. Speakers finished talking 23,791 ms into the trial on average (*SD* = 6,643), and overall spoke for 70% of the viewing period (*M* = 21,092 ms *SD* = 7,428 ms).

Table 6. *Descriptive Statistics for Onset and Offset of Speech in Experiment 3*

	Time into trial (ms)		
Measure	*M*	*SD*	*N*
Speech onset	2,699.10	1,810.94	800
Speech offset	23,791.18	6,642.97	800
Duration	21,092.08	7,428.27	800

Attention: Scene-Level Correlations

Once again, attention maps were related to feature maps (meaning, saliency, graspability) in the same manner as the previous two experiments, except that we additionally

[3] Note that because most subjects in Experiment 1 and 2 described actions from a first-person perspective, the instruction change merely confirmed that perspective as the expectation.

compared attention maps to grasp maps that were weighted by reach maps, highlighting graspable scene regions that were within reach. We correlated each feature map (meaning, saliency, graspability, reach-weighted graspability) and attention maps that were empirically derived from viewer fixations in order to determine the degree to which each feature guides visual attention (Figure 5). Squared linear and semipartial correlations (R^2) were computed for each of the 20 scenes, and the relationship between each feature map and visual attention was analyzed using paired t-tests.

Linear correlations.

Meaning and graspability both explained 36% of the variance in attention on average (meaning: *M* = 0.36, *SD* = 0.15; graspability: *M* = 0.36, *SD* = 0.16; see Figure 6). The relationship between meaning and graspability with attention did not differ (*t*(19) = -0.12, *p* = 0.91, 95% CI = [-0.05 0.05], *d* = -0.02, *d* 95% CI = [-0.33 0.29]). Image salience explained 28% of the variance (*M* = 0.28, *SD* = 0.13). Once again we replicated the advantage of meaning over salience, which was a medium effect (*t*(19) = 2.66, *p* = 0.02, 95% CI = [0.02 0.15], *d* = 0.59, *d* 95% CI = [0.10 1.08]). Graspability again explained visual attention better than image salience did, and this was also a medium-sized effect (*t*(19) = 2.74, *p* = 0.01, 95% CI = [0.02 0.15], *d* = 0.60, *d* 95% CI = [0.12 1.08]). Reach-weighted graspability explained only 29% of the variance in attention on average (*M* = 0.29, *SD* = 0.13). The relationships between reach-weighted graspability versus salience with attention did not differ (*t*(19) = -0.20, *p* = 0.83, 95% CI = [-0.08 0.07], *d* = -0.-5, *d* 95% CI = [-0.61 0.51]). Both meaning (*t*(19) = 2.28, *p* = 0.03, 95% CI = [0.006 0.15], *d* = 0.54, *d* 95% CI = [0.03 1.05]) and graspability (*t*(19) = 3.85, *p* = 0.001, 95% CI = [0.04 0.12], *d* = 0.53, *d* 95% CI = [0.23 0.83]) significantly outperformed reach-weighted graspability in accounting for attention with medium-sized effects.

Semipartial correlations.

We subsequently partialed out the shared variance explained by each feature map because the maps were correlated with one another (Table 5). When shared variance explained by salience was accounted for, meaning explained 13% of the remaining variance (*M* = 0.13, *SD* = 0.10) and salience accounted for only 4% of the variance (*M* = 0.04, *SD* = 0.05) after the contribution of meaning was accounted for, and the effect size for this comparison was large (*t*(19) = 2.67, *p* = 0.02, 95% CI = [0.02 0.15], *d* = 1.11, *d* 95% CI = [0.04 2.18]). Similarly, when the variance explained by salience was partialed out, graspability explained 13% of the variance (*M* = 0.13, *SD* = 0.11) and salience only explained 4% of the variance (*M* = 0.04, *SD* = 0.05) after graspability was partialed out, and the size of this effect was also large (*t*(19) = 2.74, *p* = 0.01, 95% CI = [0.02 0.5], *d* = 1.09, *d* 95% CI = [0.07 2.10]). Meaning accounted for 4% of the

remaining variance (*M* = 0.04, *SD* = 0.04) after the contribution of graspability was partialed out, and graspability explained 5% of the variance (*M* = 0.05, *SD* = 0.08) after the variance explained by meaning was partialed out, and this comparison again was not significant and carried a small effect size (*t*(19) = -0.12, *p* = 0.91, 95% CI = [-0.05 0.05], *d* = -0.04, *d* 95% CI = [-0.80 0.71]). When the contribution of reach-weighted graspability was partialled out, meaning explained 14% of the average variance in attention maps (*M* = 0.14, *SD* = 0.10) and reach-weighted graspability explained 6% (*M* = 0.06, *SD* = 0.09). The difference in means was significant and the effect size was large (*t*(19) = 2.28, *p* = 0.03, 95% CI = [0.006 0.15], *d* = 0.81, *d* 95% CI = [-0.02 1.64]). Similarly, when the correlation with reach-weighted graspability was accounted for, graspability explained 10% of the average variance (*M* = 0.10, *SD* = 0.07) while reach-weighted graspability only explained 2% (*M* = 0.02, *SD* = 0.03), and the advantage of graspability over reach-weighted graspability was significant, with a large effect size (*t*(19) = 3.83, *p* = 0.001, 95% CI = [0.04 0.12], *d* = 1.46, *d* 95% CI = [0.35 2.57]. Finally, when the contribution of reach-weighted graspability was partialled out, salience (*M* = 0.10, *SD* = 0.08) and reach-weighted graspability (*M* = 0.11, *SD* = 0.10) explained a comparable portion of the variance on average, and the difference was not significant (*t*(19) = -0.21, *p* = 0.84, 95% CI = [-0.08 0.07], *d* = -0.08, *d* 95% CI = [-0.90 0.73]).

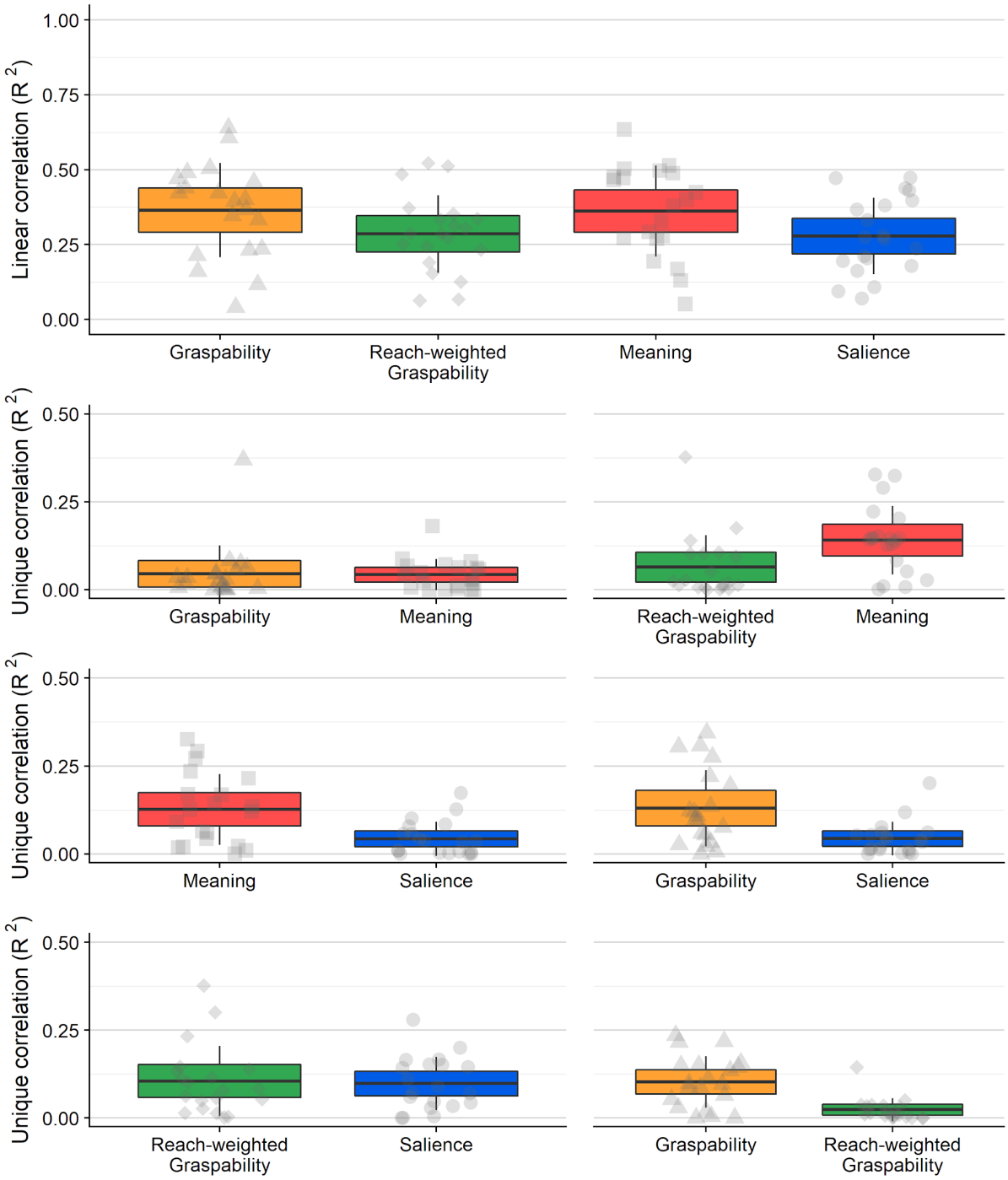


Figure 6. Box plots showing linear correlations (row 1) and semipartial correlations (rows 2-4) between feature maps (graspability, reach-weighted graspability, meaning, and salience) and attention maps for each scene. The scatter box plots show the corresponding grand mean (black horizontal line), 95% confidence intervals (colored box), and 1 standard deviation (black vertical line) for meaning (red box, square points), graspability (orange box, triangular points), reach-weighted graspability (green box, diamond points), and salience (blue box, circular points) across 20 scenes.

In sum, when we correlated attention maps with feature maps, meaning explained greater variance in attention maps than salience, as did graspability, and both effect sizes were medium. Meaning and graspability did not differ with respect to how much variance each explained in attention maps, and this was true across both linear and semipartial correlations. Reach-weighted graspability explained significantly less variance in attention than either meaning or graspability did, and it did not differ from salience. While the advantage of meaning and graspability over salience is consistent with our findings from the first two experiments, the finding that meaning and graspability are relatively equal, and both explain similar amounts of unique variance, is new. Also novel to the current experiment is the finding that reach-weighted graspability did not explain more variance in attention than salience did.

Attention: Initial Fixations

To test for changes in attentional guidance at the beginning of the trial, we again correlated each feature map (meaning, saliency, graspability, reach-weighted graspability) with attention maps at each fixation, and focused on the first three fixations in our analysis (Figure 7). Squared linear and semipartial correlations (R^2) were computed for each fixation, and the relationship between each feature map and visual attention was analyzed using paired t-tests.

Linear correlations.

Graspability explained the most variance on average at each of the three earliest time points: 36% during the first fixation (*M* = 0.36, *SD* = 0.20), 22% during the second (*M* = 0.22, *SD* = 0.16), and 20% during the third (*M* = 0.20, *SD* = 0.16). Meaning explained 29% of the average variance at the first fixation (*M* = 0.29, *SD* = 0.22), 18% at the second (*M* = 0.18, *SD* = 0.16), and 17% at the third (*M* = 0.17, *SD* = 0.15). Salience explained the least variance on average during early viewing, accounting for 9% of the average variance at the first fixation (*M* = 0.09, *SD* = 0.12), 14% at the second (*M* = 0.14, *SD* = 0.12), and 16% at the third (*M* = 0.16, *SD* = 0.15). Reach-weighted graspability explained 27% of the average variance during the first and second fixations (1: *M* = 0.27, *SD* = 0.17; 2: *M* = 0.27, *SD* = 0.17), and 21% at the third fixation (*M* = 0.21, *SD* = 0.15).

Meaning held a significant advantage over saliency during the first fixation with a large effect size (*t*(19) = 3.50, *p* = 0.002, 95% CI = [0.08 0.32], *d* = 1.13, *d* 95% CI = [0.29 1.96]) but not for the second or third fixations (2: *t*(19) = 0.83, *p* = 0.42, 95% CI = [-0.05 0.11], *d* = 0.22, *d* 95% CI = [-0.33 0.78]; 3: *t*(19) = 0.29, *p* = 0.77, 95% CI = [-0.06 0.08], *d* = 0.07, *d* 95% CI = [-0.39 0.53]). In contrast, the difference in average variance explained by graspability compared to meaning was not significant for any of the three fixations, and effect sizes were small (1: *t*(19) = -1.61, *p* = 0.12, 95% CI = [-0.16 0.02], *d* = -0.34, *d* 95% CI = [-0.78 0.10]; 2: *t*(19) = -1.55, *p* =

0.14, 95% CI = [-0.10 0.02], *d* = -0.27, *d* 95% CI = [-0.63 0.09]; 3: *t*(19) = -1.13, *p* = 0.27, 95% CI = [-0.09 0.03], *d* = -0.19, *d* 95% CI = [-0.53 0.15]). Graspability held a large advantage over saliency for the first fixation (*t*(19) = 4.75, *p* = 0.0001, 95% CI = [0.15 0.39], *d* = 1.69, *d* 95% CI = [0.57 2.81]), a marginal advantage at the second fixation (*t*(19) = 2.03, *p* = 0.06, 95% CI = [-0.002 0.15], *d* = 0.52, *d* 95% CI = [-0.03 1.06]), and did not explain the variance in attention better than saliency for the third fixation (*t*(19) = 1.08, *p* = 0.29, 95% CI = [-0.04 0.0.12], *d* = 0.26, *d* 95% CI = [-0.23 0.75]), and the latter two effect size estimates were medium and small, respectively. Reach-weighted graspability explained greater variance in attention than saliency during the first two fixations (1: *t*(19) = 3.73, *p* = 0.001, 95% CI = [0.08 0.28], *d* = 1.24, *d* 95% CI = [0.34, 2.14]; 2: *t*(19) = 3.39, *p* = 0.003, 95% CI = [0.05 0.21], *d* = 0.86, *d* 95% CI = [0.26 1.45]), and the effect sizes were large. The average variance explained by reach-weighted graspability and salience did not differ at the third fixation (*t*(19) = .29, *p* = 0.21, 95% CI = [-0.03 0.14], *d* = 0.35, *d* 95% CI = [-0.22 0.92]). During the second fixation only, reach-weighted graspability held a medium-sized, significant advantage over meaning (*t*(19) = 2.15, *p* = 0.045, 95% CI = [0.002 0.19], *d* = 0.59, *d* 95% CI = [-0.01 1.20]), but had no such advantage during the first and third fixations (1: *t*(19) = -0.28, *p* = 0.78, 95% CI = [-0.16 0.12], *d* = -0.09, *d* 95% CI = [-0.78 0.59]; 3: *t*(19) = 1.01, *p* = 0.32, 95% CI = [-0.04 0.13], *d* = 0.28, *d* 95% CI = [-0.29 0.84]). Similarly, during the second fixation reach-weighted graspability held a marginal advantage over graspability with a small effect size (*t*(19) = 1.84, *p* = 0.08, 95% CI = [-0.007 0.12], *d* = 0.32, *d* 95% CI = [-0.04 0.69]), and the two explained similar variance during the first and third fixations (1: *t*(19) = -1.63, *p* = 0.12, *d* = -049, *d* 95% CI = [-1.14 0.15]; 3: *t*(19) = 0.47, *p* = 0.64, 95% CI = [-0.04 0.06], *d* = 0.07, *d* 95% CI = [-0.24 0.39]).

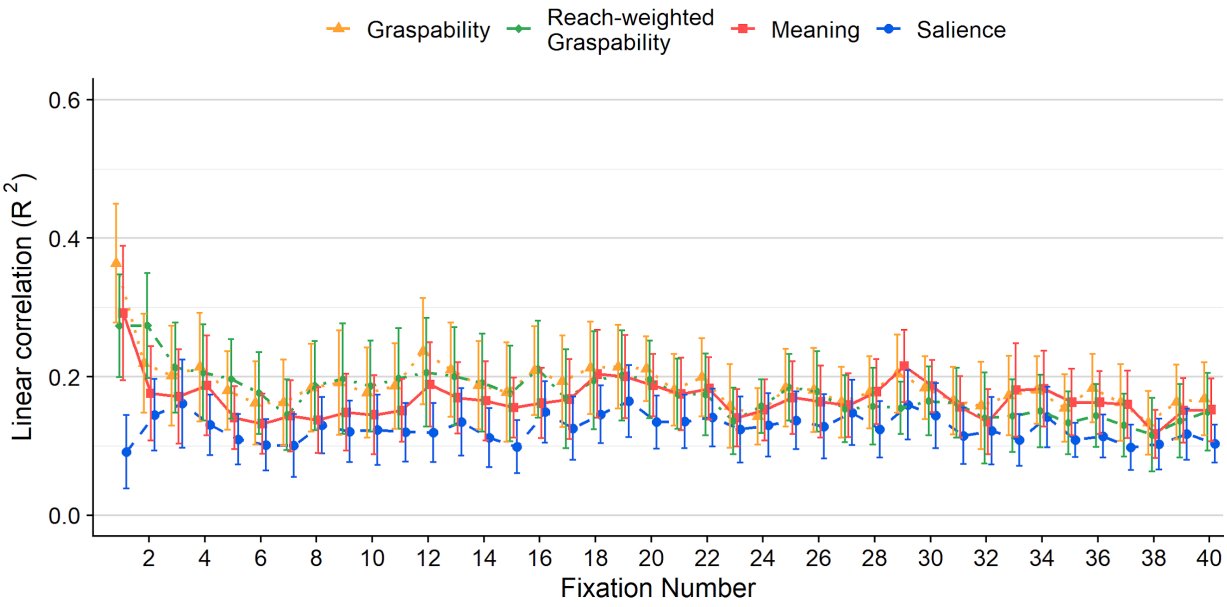


Figure 7. Line graphs showing linear correlations between feature maps (meaning, saliency, graspability, and reach-weighted graspability) and attention maps for each fixation (1-40). Error bars indicate 95% confidence intervals.

Semipartial correlations.

As in the analysis by scene, we accounted for the shared variance explained by two features using semipartial correlations, which allowed us to determine how well one feature accounted for variance in attention. We will consider the analysis for each feature pair in turn.

When the variance explained by both meaning and salience was partialled out, during the first fixation meaning explained 24% of the variance on average (*M* = 0.24, *SD* = 0.23), 10% of the variance at the second fixation (*M* = 0.10, *SD* = 0.12), and 8% of the variance at the third fixation (*M* = 0.08, *SD* = 0.10; see Figure 8). Salience accounted for 4% of the variance on average at the first fixation (*M* = 0.04, *SD* = 0.07), 7% of the variance at the second fixation (*M* = 0.07, *SD* = 0.09), and 7% of the variance at the third fixation (*M* = 0.07, *SD* = 0.08). The advantage of meaning over salience was significant for the first fixation (*t*(19) = 3.50, *p* = 0.002, 95% CI = [0.08 0.32], *d* = 1.28, *d* 95% CI = [0.28 2.28]), carrying a large effect size, and was not significant for the second and third fixations (2: *t*(19) = 0.82, *p* = 0.42, 95% CI = [-0.05 0.11], *d* = 0.29, *d* 95% CI = [-0.44 1.02]; 3: *t*(19) = 0.30, *p* = 0.77, 95% CI = [-0.06 0.08], *d* = 0.11, *d* 95% CI = [-0.64 0.86]).

After partialling out the variance explained by both meaning and graspability, on average meaning accounted for 7% of the variance during the first fixation (*M* = 0.07, *SD* = 0.11), and 4% of the variance during the second fixation (*M* = 0.04, *SD* = 0.06) and 3% during the third fixation (*M* = 0.03, *SD* = 0.03). Graspability accounted for 15% of the variance at the first fixation

(*M* = 0.15, *SD* = 0.11), 8% at the second fixation (*M* = 0.08, *SD* = 0.09), and 6% during the third fixation (*M* = 0.06, *SD* = 0.11). The difference in means was not significant at any time steps, and all effect sizes were small or medium (1: *t*(19) = -1.60, *p* = 0.13, 95% CI = [-0.16 0.02], *d* = -0.63, *d* 95% CI = [-1.51 0.24]; 2: *t*(19) = -1.56, *p* = 0.14, 95% CI = [-0.10 0.01], *d* = -0.57, *d* 95% CI = [-1.37 0.23]; 3: *t*(19) = -1.13, *p* = 0.27, 95% CI = [-0.09 0.03], *d* = -0.39, *d* 95% CI = [-1.12 0.34]).

When the shared variance that both graspability and salience accounted for was partialled out, graspability explained on average 31% of the variance at the first fixation (*M* = 0.31, *SD* = 0.20), 14% at the second fixation (*M* = 0.14, *SD* = 0.13), and 10% for the third fixation (*M* = 0.10, *SD* = 0.13), while saliency accounted for 4% of the variance on average during the first fixation (*M* = 0.04, *SD* = 0.09), 6% of the variance at the second fixation (*M* = 0.06, *SD* = 0.07), and 6% at the third (*M* = 0.06, *SD* = 0.08). The difference at the first fixation was significant and the effect size was large (*t*(19) = 4.76, *p* = 0.002, 95% CI = [0.15 0.39], *d* = 1.82, *d* 95% CI = [0.56 3.09]), whereas differences at the second and third fixations were marginal and not significant, respectively (2: *t*(19) = 2.03, *p* = 0.06, 95% CI = [-0.002 0.15], *d* = 0.72, *d* 95% CI = [-0.09 1.53]; 3: *t*(19) = 1.08, *p* = 0.22, 95% CI = [-0.04 0.12], *d* = 0.39, *d* 95% CI = [-0.37 1.15]), and effect sizes for the latter differences were medium and small.

When the contribution of reach-weighted graspability was accounted for, graspability explained 16% of the remaining average variance at the first fixation (*M* = 0.16, *SD* = 0.20), and 4% at the second and third fixations (1: *M* = 0.04, *SD* = 0.06; 3: *M* = 0.04, *SD* = 0.06). In turn, reach-weighted graspability accounted for 7% of the variance on average at the first fixation (*M* = 0.07, *SD* = 0.09), 10% of the variance at the second fixation (*M* = 0.10, *M* = 0.11), and 6% of the variance at the third fixation (*M* = 0.06, *M* = 0.07). The difference in means was marginally significant with a medium effect size at the second fixation (*t*(19) = 1.84, *p* = 0.08, 95% CI = [0.008 0.12], *d* = 0.64, *d* 95% CI = [-0.13 1.41]) and was not significant during the first and third fixations (1: *t*(19) = -1.63, *p* = 0.12, 95% CI = [0.21 0.03], *d* = -0.61, *d* 95% CI = [-1.43 0.21]; 3: *t*(19) = 0.48, *p* = 0.64, 95% CI = [0.04 0.06], *d* = 0.17, *d* 95% CI = [-0.57 0.91]).

During the first fixation, meaning accounted for 16% of the average variance after reach-weighted graspability was partialled out (*M* = 0.16, *SD* = 0.18), and only accounted for 6% of the average variance in attention during the second and third fixations (2: *M* = 0.06, *SD* = 0.08; 3: *M* = 0.06, *SD* = 0.09). Reach-weighted graspability explained 14% of the remaining variance at the first fixation (*M* = 0.14, *SD* = 0.15), 16% at the second (*M* = 0.16, *SD* = 0.17), and 11% at the third (*M* = 0.11, *SD* = 0.14). The advantage of reach-weighted graspability over meaning was significant at the second fixation with a medium effect size (*t*(19) = 2.15, *p* = 0.04, 95% CI =

[0.002 0.19], *d* = 0.76, *d* 95% CI = [-0.05 1.58]), and the difference in means was not significant during the other two time points (1: *t*(19) = -0.28, *p* = 0.78, 95% CI = [-0.16 0.12], *d* = -0.11, *d* 95% CI = [-0.92 0.70]; 3: *t*(19) = 1.01, *p* = 0.32, 95% CI = [-0.04 0.13], *d* = 0.36, *d* 95% CI = [-0.38 1.11]).

Lastly, when the contribution of reach-weighted graspability was accounted for, salience explained 4% of the average variance at the first fixation (*M* = 0.04, *SD* = 0.07), 6% at the second (*M* = 0.06, *SD* = 0.07), and 8% at the third (*M* = 0.08, *SD* = 0.08). Reach-weighted graspability explained more of the remaining variance at each time step: 23% at the first fixation (*M* = 0.23, *SD* = 0.18), 19% at the second (*M* = 0.19, *SD* = 0.13), and 13% at the third (*M* = 0.13, *SD* = 0.12). The advantage of reach-weighted graspability was significant at the first two fixations, with large effect sizes (1: *t*(19) = 3.72, *p* = 0.001, 95% CI =[0.08 0.28], *d* = 1.41, *d* 95% CI = [0.33 2.48]; 2: *t*(19) = 3.38, *p* = 0.003, 95% CI =[0.05 0.2], *d* = 1.31, *d* 95% CI = [0.24 2.38]), and was not significant at the third (*t*(19) = 1.29, *p* = 0.21, 95% CI =[-0.03 0.14], *d* = 0.50, *d* 95% CI =[-0.34 1.35]).

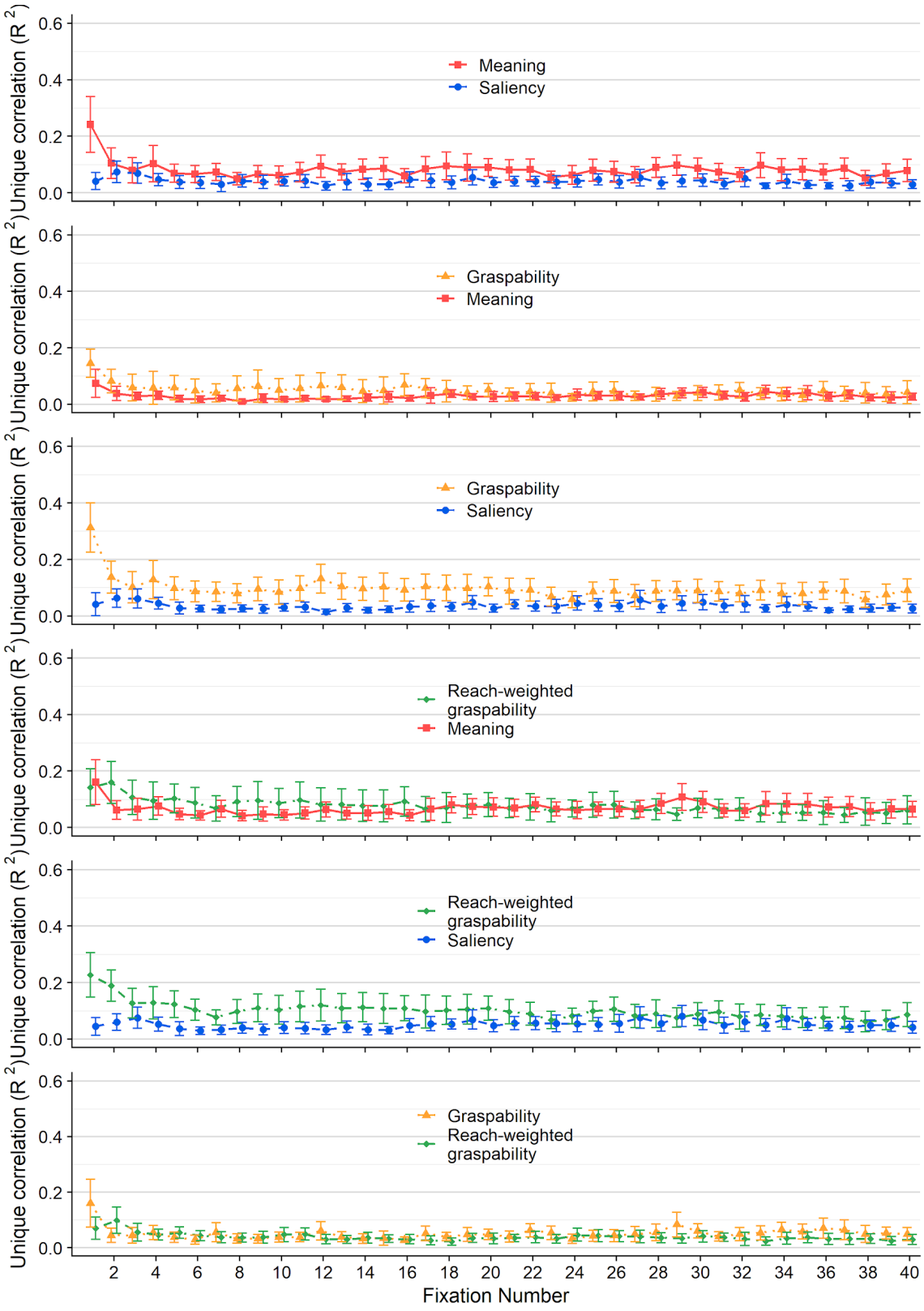


Figure 8. Line graphs showing semipartial between feature maps and attention maps for each fixation (1-40). Error bars indicate 95% confidence intervals.

In sum, when we analyzed the first three fixations, meaning outperformed salience at the first fixation, and effect sizes for all comparisons were large. Graspability explained greater variance in attention maps better than meaning did at all three time points, but the advantage of graspability was not significant, and all effect sizes were medium or small. For both linear and semipartial correlations, we found a consistent and significant advantage of graspability over salience at the first fixation for both linear and semipartial correlations, and a marginal effect at the second fixation, bearing large and medium sized effects. We further found an advantage of reach-weighted graspability over salience during the first two fixations, over meaning during the second fixation, and a marginal advantage over graspability at the second fixation. Our results are largely consistent with the findings from Experiments 1 and 2 except that graspability held a numerical advantage over meaning, though the advantage was not significant. Our results for early viewing differ from the overall analysis with respect to reach-weighted graspability, which outperformed all other features during the second fixation, but did not outperform any of the other features overall.

Discussion

The onset of speech in Experiment 3 was comparable to what we found in Experiment 1 and 2, and in our prior work. The consistency supports the idea that speakers form a general speech plan during the interval prior to speaking, which is then refined incrementally.

Once again, we found a reliable advantage of meaning over image salience in explaining the variance in attention maps (Henderson & Hayes, 2017; 2018; Henderson et al., 2018; Peacock et al., 2019a; 2019b). Consistent with Experiments 1 and 2, graspability explained variance in attention better than image salience did. Our findings for the current experiment diverged from those of Experiments 1 and 2 with respect to meaning and graspability, the key comparison of interest. In the prior two experiments, meaning explained greater variance in attention than did graspability, and the advantage was significant in at least one analysis per experiment. In the current experiment, meaning and graspability fared equally well, and indeed graspability outperformed meaning numerically during early fixations, though these numerical advantages were not significant. Reach-weighted graspability, intended to determine whether graspable objects within reach are particularly relevant, did not compete when the full trial period was considered, but showed an advantage over all other features during early viewing.

Unlike Experiments 1 and 2, graspability and meaning explained variance in attention comparably well. Given the consistency of our prior results, we conclude that the difference in Experiment 3 must be attributable to differences in the stimulus set and task instructions, which were selected to more optimally test for effects of graspability. Graspability and meaning indeed appear to tap into highly correlated aspects of the scene when graspable objects are present and close to the viewpoint, and both explain variance in attention well. Reach-weighted graspability appears to be most relevant during early scene viewing, and is otherwise as relevant as image salience.

General Discussion

The current study investigated whether speakers allocate attention using object affordances in scenes, as measured by grasp maps, when planning to describe and then subsequently describing the actions one could perform in a scene. We also quantified image salience and scene meaning for each of our scenes and determined which of the three features (meaning, graspability, or salience) was most related to the allocation of attention in the scenes. We predicted that meaning would account for visual attention in scenes better than image salience, based on a now large body of research that suggests that scene semantics guide attention over image salience (Henderson & Hayes, 2017; 2018; Henderson et al., 2018; Peacock et al., 2019a; Ferreira & Rehrig, 2019). If meaning is reducible to object affordances (Altmann & Kamide, 2007), then we expected meaning and graspability to share considerable overlap, and to explain variance in attention equally well. We further predicted that graspability, which we expected to be task-relevant, would be as related if not more related to attention in scenes than scene meaning, based on recent evidence that object affordances constrain visual attention in scenes (Malcolm & Shomstein, 2015; Castelhano & Witherspoon, 2016; Gomez & Snow, 2017; Gomez et al., 2018). In keeping with these specific predictions, we also expected graspability to explain visual attention in scenes better than image salience. Finally, we expected grasping affordances to be most relevant for graspable objects within reach of the scene's viewpoint.

The results for all experiments were consistent with two of the predictions, and partially consistent with the third. Meaning was indeed more related to visual attention than image salience, and graspability also explained attention better than image salience did. In the first two experiments, we found that meaning was more related to visual attention than graspability, a result that challenged our expectation that graspability would either fare as well as meaning, or would be more task-relevant than meaning to our speakers, and therefore would surpass meaning. However, the stimulus sets used in both experiments did not contain many objects

that would afford grasping interactions, and those objects were not positioned within reach of the viewpoint. In the third experiment, we tested a new set of stimuli selected such that each image contained graspable objects near the scene's viewpoint. Results for the latter experiment showed meaning and graspability performed equally well, and both captured little unique variance in attention, consistent with the hypothesis that meaning is co-extensive with object affordances, at least for an ideal scene set (Altmann & Kamide, 2007). Overall, reach-weighted graspability was not more related to attention than graspability or meaning were, and was on par with image salience; however, reach-weighted graspability did perform almost as well as graspability during the first fixation, and outperformed graspability, meaning, and image salience at the second fixation, which is partially consistent with the prediction that graspable objects within reach would be particularly important. The latter results suggest that attention during early scene viewing is sensitive not only to object affordances but also to how readily the action (in this case, grasping) could be carried out on the object (e.g., whether the object is in reach or not), consistent with Borghi and Riggio (2009).

Why did the results differ with respect to our primary comparison of interest between the first two experiments and the third? We suspect that grasping affordances were not as task-relevant in Experiments 1 and 2 because the scenes did not depict enough graspable objects generally, or in reachable space specifically. Because reach-weighted graspability did not rival graspability overall, we suspect that increasing the number of graspable objects alone may have been sufficient for grasping affordances to exert greater influence on attention. Future work could be conducted to systematically vary the viewpoint of the scene and the number of graspable objects in the scene to determine which factor is more strongly tied to attention.

Our findings in Experiment 3 are consistent with what has been reported recently in the literature on affordance-guided attention (Malcolm & Shomstein, 2015; Castelhano & Witherspoon, 2016; Gomez & Snow, 2017; Gomez et al., 2018). Grasping affordances explained variance in attention as well as, and sometimes better than, general scene meaning. However, whether affordances guided attention well was dependent on the scene set. Our results highlight the need to quantify both informativeness and affordances, operationalized in a task-relevant manner, in future studies of affordance-guided gaze behavior, and to carefully control for the scene viewpoint and content.

An important limitation of the current study is that our findings are restricted to attention in 2D scenes presented on a computer monitor. Previous work on the relationship between object affordances and visual attention suggests that the relationship between the two is stronger for 3D object representations, and stronger still when objects are physically present in

the environment (Gomez, Skiba, & Snow, 2018). The current study is unable to speak to the relevance of meaning, graspability, and salience to visual attention in natural 3D environments; however, we expect that graspability would be more relevant in 3D space than it was in our 2D scenes.

Based on prior scene processing work (Josephs & Konkle, 2019; Bonner & Epstein, 2017; 2018; Gomez et al., 2018), we expected graspable objects within reach to be more strongly related to visual attention than graspable objects elsewhere in the scene—in terms of our variables in Experiment 3, we expected reach-weighted graspability to explain variance in attention well, and to fare better than graspability more broadly. Partially consistent with that prediction, while reach-weighted graspability predicted attention throughout the trial period poorly, it predicted attention as well as graspability during early viewing. A compelling explanation for the early advantage of reach-weighted graspability is the foreground bias reported in a recent study by Fernandes and Castelhano (2019). The authors constructed computer-generated chimera scenes in which the foreground and background were consistent with different scene categories (e.g., a kitchen background and an office foreground). In a series of tasks, observers were biased toward the scene category consistent with the foreground of the image during the first 100 ms of scene viewing. Because the foreground of our Experiment 3 scenes depicted reachable spaces, foreground bias may account for the relatively strong relationship between reach-weighted graspability and attention during early viewing only, and we would not expect foreground bias to persist over the duration of the long viewing period we tested (30 s), consistent with reach-weighted graspability's weak performance overall. We believe investigating the relationship between reachable spaces and foreground bias is an exciting opportunity for future work. A somewhat less exciting explanation for the early performance of reach-weighted graspability is that reach-weighting the maps may have amplified the influence of center bias, which exerts a strong influence during early viewing (Hayes & Henderson, 2019a). While we prefer foreground bias as an explanation for our findings, further research is needed to determine the relative contributions of both foreground bias and center bias during initial scene viewing.

What do our results say about the nature of meaning? The findings from the third experiment point to the possibility that object affordances are co-extensive with scene meaning (Altmann & Kamide, 2007), at least to the extent that our graspability operationalization captured object affordances in our scenes. Meaning and grasp maps were highly correlated with one another in all experiments, but most consistently in Experiment 3 (see Table 6), in which meaning and graspability each explained little variance in attention after their shared variance

was partialled out. Overall, our findings suggest that scene meaning is co-extensive with object affordances when objects that afford interaction are present in the scene, and the two dissociate when such objects are absent. It is worth noting that the role of graspability was likely amplified both by the new stimuli and task demands (i.e., the requirement to describe actions; see Ostarek & Huettig, 2019). It remains to be seen whether graspability would explain variance in attention less well compared to meaning for equivalent stimuli using a more general task (e.g., describe the scene itself, or memorize it). We leave determining whether this holds for object affordances other than grasping interactions, and more general tasks, to future work.

What guides attention during language planning and production? In both experiments, and all analyses, image salience accounted for the least variance in attention of the features we tested, and rivaled only reach-weighted graspability in Experiment 3. This was not surprising in the context of our previous work (Henderson & Hayes, 2017; 2018; Henderson et al., 2018; Peacock et al., 2019a; Ferreira & Rehrig, 2019), but it does contradict saliency-based theories of attentional guidance in scene processing (e.g., Parkhurst et al., 2002) and speculation about the influence of image salience in language production (e.g., Gleitman et al., 2007; Myachykov et al., 2011; Vogels et al., 2013).

It remains difficult to compare our findings, in some respects, to results reported in the literature from studies that did not quantify the same scene features that we tested. Meaning maps are still relatively new to vision science, and we have just now introduced grasp and reach-weighted grasp maps as a way to evaluate object affordances in scenes. Methodologically speaking, our findings indicate the importance of defining and measuring relevant scene properties of our visual stimuli. We recommend that others move away from approaches that rely on researcher intuitions about scene features and instead quantify properties like meaning and graspability, which models cannot automatically quantify currently. It is also difficult to compare our findings to the literature on vision-language interactions because the practice of measuring image salience for visual stimuli using readily available tools (like GBVS; Harel et al., 2007) has not caught on in psycholinguistics. We encourage psycholinguists to begin measuring the image statistics of their visual stimuli, especially when those image properties are directly relevant to the question at hand (see Henderson & Ferreira, 2004 and Ferreira & Rehrig, 2019 for elaboration).

Consistent with our previous work (Henderson et al., 2018; Ferreira & Rehrig, 2019), speakers waited about two seconds before speaking. The pre-speech interval duration was also consistent with the time speakers needed in Griffin and Bock (2000) and Gleitman et al. (2007) before producing single sentence descriptions. Given that both the stimuli used in the current

study and the descriptions they elicited were more complex, we can again conclude that the pre-speech interval reflects a macroplanning process that speakers engage in.

Our results support the cognitive guidance theory of visual attention in scenes. The semantic content of the scene, as measured by meaning and grasp maps, explained variance in attention better than image salience, contributing to a now large body of evidence showing that attention is not pulled to a region that stands out visually, but rather the cognitive system pushes attention to informative areas. Novel to the current study, we measured semantic information in two different ways: first with respect to general meaning (informativeness and recognizability) and second to graspability, which taps knowledge about how we interact with objects. Cognitive guidance theory predicts that the most task-relevant information gets priority for attention. Meaning outperformed graspability in Experiments 1 and 2 in which the scenes did not depict reachspaces or consistently contain graspable objects, whereas meaning and graspability performed equally well in guiding attention for the scenes tested in Experiment 3, which were designed to address the stimulus limitations of the other two experiments. The strong early performance of reach-weighted graspability could also be explained under cognitive guidance theory, if nearby objects are most task-relevant early on. Furthermore, our results suggest rational behavior consistent with cognitive guidance theory. The attentional system may optimize for the information available in the scene, prioritizing the semantic information that is most task-relevant when it is available (as in Experiment 3), and using information content more broadly to guide attention when there is not enough task-relevant semantic information in the scene (as in Experiments 1 and 2).

Conclusion

In the current study, we showed that speakers use broadly defined scene meaning and object affordances, operationalized as graspability, to find relevant regions in a scene to talk about, rather than attending to regions that stand out due to their image features. Our findings suggest that semantic information in scenes is co-extensive with object affordances as captured by graspability when scenes contain multiple objects that afford grasping, and otherwise dissociate from one another. Importantly, we did not find evidence to support the claim that image salience plays a central role in vision-language interactions. We conclude that human cognitive systems make use of both object affordances, when available, and general informativeness to satisfy attentional, visual, and linguistic demands.

Acknowledgments

We thank the reviewers for their suggestions, which motivated the third experiment and greatly improved the manuscript. This research was supported by the University of California,

Davis, the National Eye Institute of the National Institutes of Health under award number R01EY027792 awarded to John M. Henderson, and National Science Foundation grant BCS-1650888 awarded to Fernanda Ferreira. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the NSF.

References

Allopenna, P., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye-movements: evidence for continuous mapping models. *Journal of Memory and Language, 38,* 419-439.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition, 73,* 247-264.

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language, 57*(4), 502-518.

Altmann, G. T. M., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition, 111,* 55-71.

Bonner, M. F., & Epstein, R. A. (2017). Coding of navigational affordances in the human visual system. *PNAS*, *114*(18), 4793-4798.

Bonner, M. F., & Epstein, R. A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLOS Computational Biology, 14*(4), e1006111.

Borghi, A. M., & Riggio, L. (2009). Sentence comprehension and simulation of object temporary, canonical and stable affordances. *Brain Research, 1253*, 117-128.

Borghi, A. M. (2012). Language comprehension: Action, affordances and goals. In *Language and Action in Cognitive Neuroscience* (pp. 143-162). Psychology Press.

Castelhano, M. S., & Witherspoon, R. L. (2016). How you use it matters: Object function guides attention during visual search in scenes. *Psychological Science, 27*(5), 606-621.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language, 47*(1), 30-49.

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(3), 687.

Chambers, C. (2016). The role of affordances in visually situated language comprehension. *Visually Situated Language Comprehension, 12*, 205-226.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* New York, NY: Routledge Academic.

Cohn, N., Coderre, E., O'Donnell, E., Osterby, A., & Loschky, L. C. (2018). The cognitive systems of visual and multimodal narratives. Symposium held July, 2018 at the *40th*

Annual Cognitive Science Society Meeting, Monona Terrace Community and Convention Center, Madison, WI.

Cullimore, R., Rehrig, G., Henderson, J. M., & Ferreira, F. (2018). When less is not more: Violations of a Gricean maxim facilitate visual search. Poster presented July 26th 2018 at the *40th Annual Meeting of the Cognitive Science Society*, Monona Terrace Community and Convention Center, Madison, WI.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision, 8*(14), 18-18.

Fernandes, S., & Castelhano, M. (2019). The foreground bias: Initial scene representations across the depth plane. *PsyArXiv* preprint. https://doi.org/10.31234/osf.io/s32wz

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

Feven-Parsons, I. M., & Goslin, J. (2018). Electrophysiological study of action-affordance priming between object names. *Brain and Language, 184,* 20-31.

Gleitman, L. R., January, D., Napa, 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.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. Psychonomic bulletin & review, 9(3), 558-565.

Glenberg, A. M., Becker, R., Klötzer, S., Kolanko, L., Müller, S., & Rinck, M. (2009). Episodic affordances contribute to language comprehension. *Language and Cognition, 1*(1), 113-135.

Gomez, M. A., & Snow, J. C. (2017). Action properties of object images facilitate visual search. *Journal of Experimental Psychology: Human Perception and Performance, 43*(6), 1115-1124.

Gomez, M. A., Skiba, R. M., & Snow, J. C. (2018). Graspable objects grab attention more than images do. *Psychological Science, 29*(2), 206-218.

Grafton, S. T., Fadiga, L., Arbib, M. A., & Rizzolatti, G. (1997). Premotor cortex activation during observation and naming of familiar tools. *Neuroimage, 6*(4), 231-236.

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology: General, 145*(1), 82-94.

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

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. In *Advances in neural information processing systems* (pp. 545-552). Cambridge, MA: MIT Press.

Harptaintner, M., Sim, E.-J., Trumpp, N. M., Ulrich, M., & Kiefer, M. (2020). The grounding of abstract concepts in the motor and visual system: An fMRI study. *Cortex, 124,* 1-22.

Hayes, T. R., & Henderson, J. M. (2019a). Center bias outperforms image salience but not semantics in accounting for attention during scene viewing. *Attention, Perception, & Psychophysics*, 1–10.

Hayes, T. R., & Henderson, J. M. (2019b). Scene semantics involuntarily guide attention during visual search. *Psychonomic Bulletin & Review, 26*, 1683–1689.

Henderson, J. M., & Ferreira, F. (2004). Scene Perception for Psycholinguists. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 1-58). New York, NY, US: Psychology Press.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Image salience versus cognitive control of eye movements in real-world scenes: Evidence from visual search. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movement research: Insights into mind and brain* (pp. 537-562). Oxford, England: Elsevier.

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science, 16,* 219-222.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review, 16*(5), 850-856.

Henderson, J. M. (2017). Gaze control as prediction. *Trends in Cognitive Science, 21*(1), 15-23.

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. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision, 18*(6), 10-10.

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*, 19.

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

Itti, L., & Koch, C. (2001). Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging, 10*(1), 161-170.

Josephs, E. L., & Konkle, T. (2019). Perceptual dissociations among views of objects, scenes, and reachable spaces. J*ournal of Experimental Psychology: Human Perception and Performance, 45*(6), 715–728.

Kako, E., & Trueswell, J. C. (2000). Verb meanings, object affordances, and the incremental restriction of reference. *Proceedings of the Annual Meeting of the Cognitive Science Society, 22*(22).

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49,* 133-156.

Kaschak, M. P., & Glenberg, A. M. (2000). Constructing meaning: The role of affordances and grammatical constructions in sentence comprehension. *Journal of Memory and Language, 43,* 508-529.

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology, 102,* 59-70.

Malcolm, G. L., & Shomstein, S. (2015). Object-based attention in real-world scenes. *Journal of Experimental Psychology: General, 144*(2), 257-263.

Martin, A. (2007). The representation of object concepts in the brain. *Annu. Rev. Psychol., 58,* 25-45.

Myachykov, A., Thompson, D., Scheepers, C., & Garrod, S. (2011). Visual attention and structural choice in sentence production across languages. *Language and Linguistics Compass, 5*(2), 95-107.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision, 10*(8), 20-20.

Ostarek, M., & Huettig, F. (2019). Six challenges for embodiment research. *Current Directions in Psychological Science, 28*(6), 593-599.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*(1), 107-123.

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019a). Meaning guides attention during scene viewing, even when it is irrelevant. *Attention, Perception, & Psychophysics*, *81*, 20-34.

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019b). The role of meaning in attentional guidance during free viewing of real-world scenes. *Acta Psychologica,* 198, 102889.

Salverda, A. P., Brown, M, & Tanenhaus, M. K. (2011). A goal-based perspective on eye movements in visual world studies. *Acta Psychologica, 137*(2), 172-180.

SR Research (2017). *EyeLink 1000 Plus User Manual, Version 1.0.2*. Mississauga, ON: SR Research Ltd.

Tanenhaus, M. K., Magnuson, J. S., Dahan, D., & Chambers, C. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research, 29*(6), 557-580.

Vogels, J., Krahmer, E., & Maes, A. (2013). Who is where referred to how, and why? The influence of visual saliency on referent accessibility in spoken language production. *Language and Cognitive Processes, 28*(9), 1323-1349.

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour, 1*(3), 0058.

Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S., & Zhao, Q. (2014). Predicting human gaze beyond pixels. *Journal of Vision, 14*(1):28.