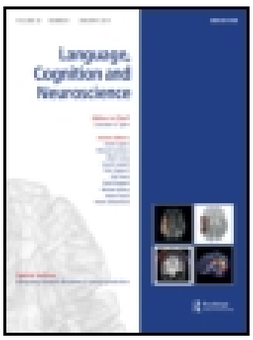

Routledge
Taylor & Francis Group

Language, Cognition and Neuroscience

ISSN: 2327-3798 (Print) 2327-3801 (Online) Journal homepage: https://www.tandfonline.com/loi/plcp21

Linearisation during language production: evidence from scene meaning and saliency maps

Fernanda Ferreira & Gwendolyn Rehrig

To cite this article: Fernanda Ferreira & Gwendolyn Rehrig (2019): Linearisation during language production: evidence from scene meaning and saliency maps, Language, Cognition and Neuroscience, DOI: 10.1080/23273798.2019.1566562

To link to this article: https://doi.org/10.1080/23273798.2019.1566562

LANGUAGE, COGNITION AND NEUROSCIENCE
https://doi.org/10.1080/23273798.2019.1566562

Routledge
Taylor & Francis Group

REGULAR ARTICLE

Linearisation during language production: evidence from scene meaning and saliency maps

Fernanda Ferreira and Gwendolyn Rehrig

Department of Psychology, University of California, Davis, Davis, CA, USA

ABSTRACT

Speaking (1989) inspired research on topics such as word selection, syntactic formulation, and dialogue, but an issue that remains understudied is linearisation: the question of how speakers organise a series of utterances into a coherent sequence, allowing both speaker and listener to keep track of what has been said and what will come next. In this paper we describe a new line of research investigating linearisation during scene description tasks, and we argue that, as Pim Levelt suggested in 1981 and in the 1989 book, the need to linearise arises from attentional constraints in the language system. Our work shows that attentional, visual, and linguistic processes are flexibly coordinated during scene descriptions, and that speakers not only respond to what the eye sees, but also to what the mind anticipates finding in the visual world.

ARTICLE HISTORY
Received 29 August 2018
Accepted 19 December 2018

KEYWORDS
Language production; linearisation; scene semantics; eye movements

In *Speaking* (Levelt, 1989), Pim Levelt observed that research in psycholinguistics had been heavily weighted towards the study of language understanding. Language production was certainly a topic of interest, but the literature was far sparser than for comprehension. The general sense among psycholinguists was that speaking was so difficult to study it was probably best to stick to a domain like comprehension in which empirical methods were already in general use and understood (or at least assumed to be). But with the publication of *Speaking*, language production researchers finally had what they'd needed to allow their field to take its place as an equal psycholinguistic partner: a comprehensive, sophisticated, and engaging overview of what was known up to that point about the processes that support speaking, and a coherent theoretical perspective from which to understand those findings and derive predictions to motivate new studies. To be fair, many excellent summaries and theoretical pieces already existed at that time (e.g. Bock, 1987; Fromkin, 1973; Garrett, 1988), but the scope and depth of *Speaking* was something else entirely. The book covered the history of the field as well as the various approaches to studying it, and it took on essentially the entire system, starting with conversation and the processes that support thinking-for-speaking, and proceeding all the way to articulation and speech monitoring.

Following the publication of the book, the field of language production began to flourish. The number of scientific studies conducted to evaluate an idea proposed in that single book is exceptional; one can only get a sense of the numbers by noting that *Speaking* has been cited almost 11,000 times. Topics that have been extensively investigated in the post-*Speaking* era include coordination in dialogue, lemma selection and word-form retrieval, syntactic formulation, prosodic planning, production of disfluencies and repairs, and speech monitoring. Architectural issues have also been extensively explored, including how different levels of the production system communicate and over what kind of domain planning proceeds. One important idea that has emerged from this research is that speakers are strategic about the domain over which they plan, sometimes preparing a very large chunk of speech and sometimes planning over domains no larger than single words or even individual syllables (Ferreira & Swets, 2002; van de Velde & Meyer, 2014). Incremental production implies that often the words a speaker utters are the result of opportunistic cognitive processes that take advantage of the state of the language and cognitive systems, allowing the speaker essentially to multitask planning and execution.

At the same time, not all the topics received the attention they deserve, and one understudied domain is what Pim Levelt called *linearisation*. The idea is that thoughts must ultimately be mapped onto individual words spoken one at a time. This requirement to push complex thought through a sequential channel means that speakers must confront the linearisation problem: They must decide in what order to output their thoughts

CONTACT Fernanda Ferreira fferreira@ucdavis.edu

and ideas. A simple (and much discussed) example is that any transitive event must be mapped onto a sequence that specifies whether the agent or patient of the action goes first – that is, the speaker must choose between an active and a passive construction (*The dog chased the cat* versus *The cat was chased by the dog*). This type of linearisation is one in which the speaker must decide how to map thematic roles onto syntactic positions in a hierarchical structure, and these decisions are influenced by factors such as the activation levels of concepts – more activated concepts go early, for example. Less appreciated, however, is the speaker's need to linearise sentences or utterances. If you are to describe a car accident you witnessed on the way to work or the layout of an airport you frequently visit, you must decide where to start and where to end, and you must make use of some type of bookkeeping system to keep track of what's been described and what hasn't. This is the topic of this piece, written to communicate this new line of work from our lab which we believe to be novel and exciting, and also to pay tribute to a great intellectual and scientist: the one and only Pim Levelt.

Linearisation in language production

The requirement to linearise is obvious when it comes to descriptions of complex visual scenes. Consider Figure 1A, which we will use throughout this article: To describe the scene in reasonable detail to another person, multiple expressions must be uttered, and those expressions must be sequenced. Based on scene research, we know that the gist of the scene is extracted in less than 100 ms (Castelhano & Henderson, 2007, 2008), and a single fixation likely yields information about most of the large objects and surfaces in that scene (Biederman, 1981; Potter, Wyble, Hagmann, & McCourt, 2014). But viewers must also make additional fixations on individual objects to know what else is present, and those fixations must be ordered. In other words, there is a linearisation problem both in scene processing and in language production. Because linearisation is necessary in both cases, we can ask whether the solution to one problem is the solution to the other: that is, does the viewer's plan for ordering fixations constitute the speaker's linearisation plan as well? This possibility is an intriguing implication of Levelt's work on linearisation, both in *Speaking* and in an article published in 1981 on the same topic (Levelt, Page, & Longuet-Higgins, 1981). Levelt argues that the requirement to linearise is not due simply to the fact that we can't make multiple speech sounds simultaneously with our articulators, or that we can't output one message orally and an entirely different one with our hands. The demand to linearise, Levelt et al. (1981) argued, is attentional, and it arises because neither the production nor the comprehension system can process two or more semantically unrelated streams of information simultaneously. Levelt suggests further that the allocating and shifting of attention which take place during planning and production reflect the speaker's linearisation scheme, leading to the fundamental assumption behind the line of research we summarise here: that eye movements reflect the speaker's linearisation of a multidimensional visual stimulus.

Surprisingly few studies have examined linearisation directly. What we do know suggests that speakers attempt to linearise complex visual stimuli by adopting a strategy of minimising the number and size of jumps needed to cover the relevant visual territory. In work investigating the description of networks of coloured circles organised into branches of different lengths and complexities, Levelt et al. (1981) observed that speakers prefer to describe short branches before long ones, and they choose to describe simple branches before ones that are complex (i.e. that contain an additional choice point). This "Minimal Load" or "Easy First" strategy (MacDonald, 2013) reduces the burden on memory by minimising the number of circles over which the choice point must be stored and later retrieved, and by eliminating the requirement to maintain more than one choice point in memory at any one time (see also Ferreira & Henderson, 1998). Speakers thus prefer to begin descriptions of visual displays with something that is easy, and they explore that area of the display before moving away to more distant regions, reducing the need to jump back and forth. Room descriptions were the subject of another study from the same era (Shanon, 1984), although unfortunately subjects in that experiment did not describe scenes in front of them or even scenes recently viewed, but instead relied on their memory for a scene well known to them, and their descriptions were written down rather than spoken. Shanon reported that writers tend to proceed from the general to the specific, usually starting with a scene label (e.g. "it's a kitchen") and then proceeding to objects that are difficult to move (e.g. countertops and large appliances). Small objects are mentioned next, and typically in relation to larger ones (e.g. "there's a kettle on the stove"). On the basis of these results, Shanon argued that speakers generate descriptions by following a cognitive schema that structures the visual world hierarchically.

This small literature suggests that speakers generate multi-utterance chunks meant to capture some aspect of the visual world by generating a hierarchical speaking plan that is organised around a scene schema. To

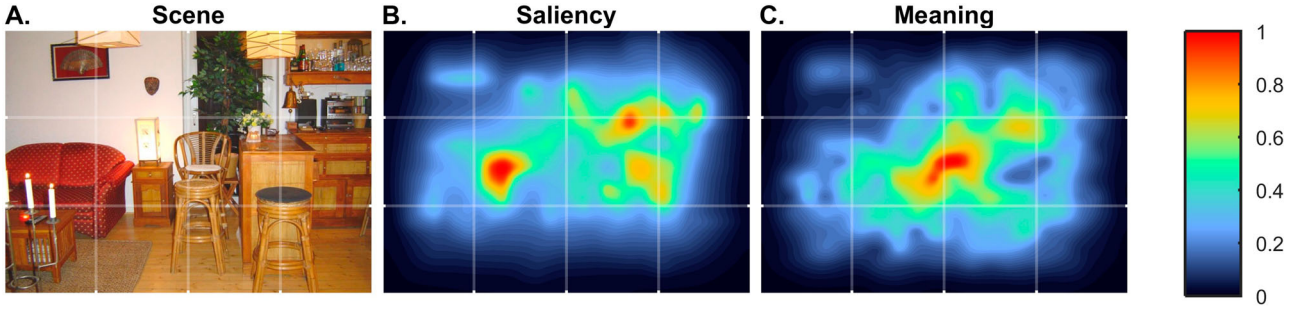

Figure 1. Representative scene image (A) used to elicit scene descriptions, along with its corresponding saliency map (B) and meaning map (C).

encourage incremental production, they tend to start with something that is easy, and they keep talking about the things in that area until most of the relevant objects are mentioned. Then they move on to other semantically coherent regions. At least, this is what we infer from the verbal protocols themselves. But what can we learn from monitoring attention during speaking tasks by recording speakers' eye movements as they generate descriptions of the visual world?

Previous work on vision-language production using eye movements

A fair bit of research has been conducted using eye movements to examine how people describe visual objects that are often described as scenes, although the stimuli typically lack many key properties of true scenes, including background, complexity, occlusion, and so on (see Henderson & Ferreira, 2004, for review). Most so-called scenes are essentially a few clip-art images pasted into a simple or nonexistent background, and typically they show events such as one (animate) entity chasing another. As mentioned above, because the visual display is usually describable as a single event and in a single utterance ("A dog is chasing a man"), the linearisation challenge for the speaker is somewhat simpler than for multi-sentence sequences, or at least it's clearly different, involving mainly the need to map thematic roles on to syntactic positions in a single clausal structure.

A well-known factor that affects a speaker's linearisation scheme is recent experience with a particular structure, better known as structural priming (Bock, 1986; Bock & Ferreira, 2014; Chang, Dell, Bock, & Griffin, 2000; Melinger, Branigan, & Pickering, 2014). Specifically, the word order and syntax a speaker uses to describe a picture can be primed by that of a previously read or spoken sentence. For example, a speaker may use a dative sentence to describe a picture after saying a dative prime on the previous trial (Bock & Loebell, 1990). Beyond priming a particular lexical item or syntactic structure, the conceptual structure of a prime sentence affects the information speakers choose to include in their descriptions of events (Bunger, Papafragou, & Trueswell, 2013). Bunger et al. (2013) primed subjects with written event descriptions, after which subjects were asked to describe an animated motion event. The primes either (1) overlapped with targets conceptually (describing the same general type of motion event) and lexically (the prime verb could be applied to the target), (2) overlapped with the target event conceptually, but the verb could not be reused, or (3) had no overlap with the target. Primes that overlapped with the target influenced the linearisation of speakers' descriptions of the event such that information (e.g. path vs. manner of motion) was included in roughly the same order. As a whole, this body of work indicates that recent experience shapes linearisation strategies. It is important to note that the priming literature is responsible for the majority of what is known about linearisation strategies in language production. Next, we discuss what has been revealed thus far from the few studies that have been done investigating how speakers describe events when recent experience is less relevant.

Griffin and Bock (2000) conducted a key study using eye movements to investigate formulation processes during production. From the finding that speakers required about 1600 ms to begin labelling the event in the picture and the finding that eye movements during that "apprehension phase" did not reflect the ultimate order of mention for the objects in the sentence, Griffin and Bock concluded that, prior to articulation, speakers conceptualise the entire event, and phrasal order is driven by speakers' perspective on that event rather than by what Griffin and Bock call simple salience. As argued in a later paper, "When speakers produce fluent utterances to describe events, the eye is sent not to the most salient element in a scene, but to an element already established as a suitable starting point" (Bock, Irwin, Davidson, & Levelt, 2003; see Bock, Irwin, &

Davidson, 2004 for a similar argument). Gleitman, January, Nappa, and Trueswell (2007) took issue with this interpretation, suggesting that perhaps the results Griffin and Bock (2000) observed were driven mainly by the subjects' desire to avoid generating the dispreferred passive. In addition, Gleitman et al. suggested that Griffin and Bock's scenes might have been less interesting and engaging to their subjects, and so they made use of visual stimuli with colour and some other visual details, which they argued allowed participants to apprehend the events shown in their displays more rapidly and to begin speaking as soon as they identified a single entity that could serve as the subject of the sentence. Nevertheless, in the Gleitman et al. study, the latency to begin speaking was even longer than in Griffin and Bock, at least for active/passive pairs (which did not differ significantly from each other): around 2000 ms (versus 1600 ms in Griffin & Bock). Thus, whether speakers planned their event descriptions more incrementally, as argued by Gleitman et al., or more holistically, as suggested by Griffin and Bock, it does appear that speakers wait before starting to talk – indeed, the delay is equivalent to the time required to make five to 10 fixations. This result will be discussed further when we consider our own studies of eye movements and linearisation.

Further evidence for incrementality in language production comes from a paradigm in which two interlocutors instructed one another to click on a single target in a grid-like display using coordinated noun phrases (Brown-Schmidt & Konopka, 2015; Brown-Schmidt & Tanenhaus, 2006). A competitor that differed from the target object in one dimension (e.g. size) was present in the display, which required disambiguation. Brown-Schmidt and Konopka (2015) contrasted an initial preparation account of language production with the continuous incrementality account. The critical difference between the two accounts as outlined by the authors is that the latter allows for updating of the speech plan on the fly (e.g. adding a modifier to a noun phrase) without the need to make an appropriateness repair in the second noun phrase (e.g. "the star and the dog, uh small dog"). The authors found that speakers were able to fluently modify the second noun phrase even when they had not fixated on the contrasting object prior to speaking, supporting the continuous incrementality account. However, it is important to note that the task employed conversational speech that required only very simple syntactic structures to communicate relevant information, both of which may favour incrementality over planning. In addition, speakers would not be able to extract scene gist from displays of the sort used in this paradigm. The availability of gist information in more natural scenes could encourage less incremental strategies.

A recent study by Elsner, Clarke, and Rohde (2018) employed a similar paradigm to investigate language production, in which speakers described more complex image arrays. The authors focused more directly on the extent to which scene properties might influence the content of descriptions as well as the time to initiate them. One goal of their study was to reconcile incremental production with the fact that speakers cannot understand the entire content of a scene without examining it closely (i.e. fixating on individual objects), which could complicate the job of speaking if a referential expression needs modifiers to disambiguate the mentioned object from others of the same type not yet identified in the scene. Their study made use of "scenes," which were collections of coloured circles and squares of different sizes that were meant to elicit expressions such as *the only circle*. The average latency to begin to speak in this study was 2100 ms, but importantly, this initiation time increased with the presence of competitors that forced or at least encouraged the speaker to include disambiguating terms such as *green* and *only*. Moreover, if the subject began to speak before taking in enough of the scene to appreciate the presence of a competitor requiring linguistic disambiguation, the result was insufficiently elaborated descriptions that had to be further linguistically refined during articulation, which indicates that there are speed-accuracy tradeoffs in utterance planning. This pattern shows a cost, then, for incremental production, and the findings also highlight the tradeoff describers of a visual world must constantly navigate between the need to fully and accurately describe a scene and the desire to maintain fluency and minimise memory and attentional load during language formulation. Thus, taking time to plan, as demonstrated in Griffin and Bock (2000) as well as Gleitman et al. (2007), is computationally costly, but it may pay off in the form of clearer, more accurate, and more fluent expressions.

These and other studies on single-sentence descriptions of visual displays have been important and influential, but many questions remain to be answered, and critically, a different paradigm for studying vision-language interactions is required to address them. First, as mentioned, the stimuli used in these experiments are not true scenes – they are not images like the one shown in Figure 1A, for example, and this limits the conclusions that can be drawn from these studies, for reasons discussed in Henderson and Ferreira (2004). In addition, because most previous work is meant to address the (clearly important) question of how speakers assign thematic roles to syntactic positions, the paradigms have been designed to elicit single sentences,

and thus the linearisation challenge speakers must face is qualitatively different from the one presented by the need to describe a scene such as the one shown in Figure 1A.

Some disagreement remains over what processes occur before the speaker begins to describe a scene. Griffin and Bock (2000) suggested that the delay between stimulus onset and speech onset reflected an apprehension stage during which speakers interpreted the scene. Gleitman et al. (2007) argued instead that the delay reflected both event apprehension and a preliminary planning stage before speech. To determine how much processes of utterance planning vs. event apprehension contribute to the delay before speaking, it is important to consider how quickly we can extract the relevant information from a scene when no language planning is necessary. Scene processing research has revealed that scene gist can be extracted from presentation intervals shorter than 100 ms (Castelhano & Henderson, 2007, 2008) and from a single fixation (Biederman, 1981; Potter et al., 2014). Hafri, Papafragou, and Trueswell (2013) investigated whether event gist – namely, understanding what action was shown and assigning thematic roles to event participants – could be extracted rapidly, like scene gist. Event gist was successfully extracted even at presentation intervals as short as 37 ms. Similarly, speakers were faster to name actions when primed by an image portraying either the same action or a related action, even when the prime was only shown for 50 ms (Zwitserlood, Bölte, Hofmann, Meier, & Dobel, 2018). This suggests that event apprehension alone is not sufficient to explain the delay that occurs before speakers describe the scene, as those delays are far longer than the time required for gist extraction.

Finally, few studies have measured the visual properties of the displays in a systematic way that is informed by theories of scene meaning and scene processing. The use of the term "salience" is a good example of this shortcoming. In most psycholinguistic studies, "salience" is used interchangeably with words such as "importance" and "interestingness," so that a cat shown in a uniform background would be described as visually salient. But in the scene literature, salience (or "saliency") has a technical definition, referring specifically to visual features such as luminance, colour, and edge orientation (Henderson & Hayes, 2017; Itti & Koch, 2001; Itti, Koch, & Niebur, 1998; Torralba, Oliva, Castelhano, & Henderson, 2006) that vary continuously over the entire scene. Visual salience ignores the semantics of the scene; instead, visually salient regions are assumed to pull attention and eye movements, allowing the visual system to identify objects and assign meaning to them post-fixation. The degree to which an image region stands out from its surroundings with respect to these visual features – luminance, colour, edges – can be automatically quantified using readily available tools (e.g. Graph-Based Visual Saliency: Harel, Koch, & Perona, 2007), which has made it practical for researchers to investigate the influence of salience on visual cognitive processes. Scene meaning is a different story, as there is currently no fully automated way to quantify it. Historically, the question of how to quantify meaning within scenes has been considered a non-trivial problem (Chun, 2000), and it is only recently that researchers have begun to take on the challenge (e.g. Henderson & Hayes, 2017; Koehler & Eckstein, 2017a, 2017b; Rehrig, Cheng, McMahan, & Shome, in prep), thanks in part to the advent of crowdsourcing data collection platforms. In most psycholinguistic studies of vision-language interactions, scene meaning is also treated in a similarly informal way, typically referring simply to the presence of objects that are likely to be of interest (e.g. humans, animals, points of interaction among agents). Without quantitative measures of meaningfulness, it is difficult to predict which objects are likely to draw attention on semantic grounds, and researchers tend to fall back on their intuitions (e.g. animate things are more semantically prominent than are inanimate things and therefore more likely to be mentioned first, etc.). As a group, psycholinguists have already accepted that intuitions alone are insufficient to gauge the properties of linguistic stimuli, as evidenced by the now widespread practice of norming experimental stimuli. But, the same care and attention has not been extended to the visual stimuli used to study vision-language interactions. Thus, what the field has lacked thus far are continuous, high-resolution, high-density measures of both visual salience and meaning. This shortcoming of the literature is also addressed in our current work.

A detour through scene literature: does visual salience or meaning drive attention in scenes?

In this section we briefly describe work from visual cognition that informs the current psycholinguistic project. The fundamental debate in the scene literature has been about whether eye movements in scenes are driven by visual properties or by meaning. As mentioned, visual salience can be measured precisely, and algorithms such as the Graph-Based Visual Saliency (GBVS) Toolbox (Harel et al., 2007) yield saliency maps like the one shown in Figure 1, Panel B. Eye movements are predicted to be drawn to the brighter regions in that map. To quantity meaning in similar detail and on the same scale, Henderson and Hayes (2017) developed what

they term "meaning maps," an example of which is shown in Figure 1, Panel C. The meaning map represents the spatial distribution of meaning across the scene and is generated via crowdsourcing. In Henderson and Hayes, 165 naïve subjects viewed patches of 40 scenes at two spatial scales (3 degrees and 7 degrees), rating the individual patches from 1 to 7 (very low to very high meaning). These ratings are then combined to form meaning maps like the one shown in Figure 1C, which, like the saliency map, essentially constitutes a set of predictions about where naïve viewers are likely to fixate in the scene.

The next step is to pit visual salience against meaning by observing where people in fact look when they view scenes, and to compare fixation "hot spots" to places with high visual salience or high meaning. One important challenge, which the reader might have already appreciated by comparing Figure 1B and C, is that visual salience and meaning are not unrelated. Intuitively, this correlation makes sense: after all, nothing interesting is likely to be found on a uniformly blank wall; in contrast, people are drawn to objects in scenes which are meaningful and tend to be bounded by visually salient edges. Thus, to truly disentangle the contributions of visual salience and meaning to attentional control during scene viewing, it is important to take this correlation into account. This has now been done in a number of studies from the Henderson lab (Henderson & Hayes, 2017, 2018; Henderson, Hayes, Rehrig, & Ferreira, 2018; Peacock, Hayes, & Henderson, 2018), and in all of them the conclusions are the same. Overall, and for virtually every single scene, meaning maps do a better job of predicting where people look when they look at scenes once the intercorrelation between the two is partialled out; indeed, saliency contributes nothing once the effect of meaning is taken into account. This is true whether the subjects' task is to view the scene in anticipation of a later memory test, to assess the aesthetic qualities of the scene, or even to find patches in the scene that are visually salient (Peacock et al., 2018). The conclusion that emerges from this work is that meaning guides attention in scenes. Visual salience may generate a set of potential fixation targets, but it is a flat target landscape. Meaning is what assigns priorities to those targets, making some objects more likely to be fixated than others.

Based on this conclusion, we can now ask three basic questions about the linguistic descriptions of scenes: First, when people generate a linearisation scheme to talk about the visual world in front of them, what controls attention: visual salience or meaning? In previous studies of vision-language interaction, some researchers have suggested that visual salience influences word order and word choice (Griffin & Bock, 2000; Gleitman et al., 2007; Myachykov, Thompson, Scheepers, & Garrod, 2011; Vogels, Krahmer, & Maes, 2013), though the salience of their stimuli was not quantified, and so they could not empirically test this claim. Second, do the relative contributions of salience versus meaning differ depending on the nature of the speaking task, or depending on whether people are in a planning or a speaking phase of production? And third, we have seen that speakers wait about one to two seconds before saying even one-sentence descriptions. How long are the latencies to speak in our task, and can we learn more about what those latencies reflect by looking at task differences and eye movement patterns?

Spoken descriptions of complex scenes: visual salience or meaning?

To answer these three fundamental questions, we will summarise the results of three scene description experiments that we have completed thus far as part of this project (Henderson et al., 2018). In each one, 30 people were tested on 30 real world scenes such as the scene shown in Figure 1A. All trials began with subjects fixating in the centre of the screen. In the first experiment, the scene was presented and subjects had 30 seconds in which to describe it. They began as soon as the scene appeared and were encouraged to talk for the entire 30 seconds (via practice items). For the second experiment, to encourage the generation of sentences with verbs other than just the copula, subjects were asked to describe what they or someone else might do in that particular scene. As in the first experiment, subjects began speaking as soon as they could once the scene appeared, and they stopped after 30 seconds. And for the third experiment, subjects viewed the scene silently for 30 seconds and then the scene was removed, at which point their task was to describe it from memory for 30 seconds. We designed the first and third experiments to parallel the extemporaneous and prepared speech tasks that Griffin and Bock (2000) used, respectively. To determine whether meaning or saliency best accounted for visual attention during scene viewing, we computed attention maps empirically from viewer fixations. We then correlated attention maps with the corresponding saliency and meaning maps, and subsequently analysed the variance in attention that the two maps captured.

The results of all the experiments can be described very straightforwardly: In each of the three, meaning maps accounted for more of the unique variance than did visual salience. In Experiment 1, meaning accounted

for 46% of the variance and visual salience for 34%, a highly significant difference. To take into account the correlation between meaning and visual salience, we also computed semi-partial correlations, and those revealed that meaning accounted for 17% of additional variance once salience was accounted for, but salience accounted for a nonsignificant 4% of the variance once meaning was taken into account. Moreover, this pattern held for almost every one of the 30 scenes (no significant reversals were found for any scene). In addition, in this experiment we examined whether the ability of meaning versus visual salience to predict fixations changed over the course of the 30 second viewing period. The logic here is that perhaps salience is more predictive early on given that subjects may not yet have identified all the objects in the scene and might be using salience to establish fixation targets, and then meaning might come into play. But contrary to this idea, we actually observed the opposite pattern: The superiority of meaning over salience was greatest at the start of the 30 second interval; over time, meaning and salience eventually converged but to values hovering near zero, indicating that about halfway through the 30 second interval for describing the scenes, neither meaning nor salience predicted where participants were looking.

The next possibility we explored was to assess whether, in this same experiment, the relative dominance of meaning over salience might differ depending on whether people were still engaged in speech planning versus actively articulating. Beginning with latencies to start describing the scene, we observed that people waited on average 1678 ms before starting to talk. It is interesting to note, as an aside, that this time is the same as for the single-sentence descriptions elicited from subjects in Griffin and Bock (2000) and shorter than those reported by Gleitman et al. (2007). We will return to this point later. Given that a typical fixation during scene viewing lasts around 300 ms, we can assume that 1600 is enough time for a viewer to make five to eight fixations, and during this period, we found that meaning predicted fixation locations better than did visual salience. Similarly, when people paused (defined as a filled or silent hesitation lasting at least 500 ms) during the 30 seconds in which they were supposed to be producing their scene descriptions, we found that meaning predicted what people were fixating on better than salience did for the period including 1000 ms before the pause onset and during the pause itself.

Finally, we also pulled out the cases in which speakers spontaneously generated a colour term and analysed the location of fixations 1000 ms before and during the production of that colour expression. Our logic was that colour is a component of visual salience and is used to generate saliency maps, so perhaps visual salience would finally dominate over meaning before and during production of colour terms. However, our intuition was wrong – the correlations between fixation locations on the one hand and meaning maps and saliency maps on the other were the same and all near zero, suggesting that whatever drives fixation locations before and during the production of a colour word, it is neither of those two scene properties. To rule out the possibility that the analysis window was too broad, we repeated the analysis using (1) fixations that took place 300 ms before and during the production of a colour term (the first fixation preceding the term, and fixations during the term itself), and (2) on only the first fixation after the onset of the colour term. The results of the follow-up analyses were the same as the original, in that meaning and saliency were essentially the same, and both were very poor predictors of attention. We interpret these findings with caution because there were few fixations overall in these analyses, given that colour terms comprised a small fraction of the productions we recorded.

Thus, for this first experiment in which subjects produced open-ended descriptions of the scenes, meaning maps clearly dominated over visual salience. We found this result in the other two experiments as well – that is, we observed that meaning predicted fixation locations better than did salience whether people provided action descriptions for the scenes or described the scenes from memory after those scenes had disappeared. In some ways, this set of results across the three experiments is not surprising, as any type of speaking task certainly requires viewers to access scene meaning (as opposed to, say, an aesthetic judgment task which could arguably be performed even on a meaningless stimulus). Nonetheless, it is important to remember one of our central observations, and that is that the term "salience" is used in psycholinguistics quite loosely, and often carries the connotation that purely visual features are what drive attention during language production (or comprehension) tasks. Our results show clearly that the terminology used in psycholinguistics needs to be far more precise, because if what researchers mean by "salience" is interestingness, then the concept at issue is meaning, not visual salience. In addition, many current models of vision-language interaction continue to assume that visual salience plays a central role, particularly during the earliest stages of interpretation or formulation (for example, see Cohn, Coderre, O'Donnell, Osterby, & Loschky, 2018). Again, if our conclusions are correct, it is not visual salience that controls the earliest stages of language planning, but rather meaning extracted almost from the moment the scene is shown.

In summary, then, scene meaning drives attention when people describe complex scenes. This pattern holds across language tasks, throughout the entire viewing period, and for online descriptions it holds during pauses and during speech preparation. We can conclude, then, that the linearisation scheme for describing visual stimuli like the one shown in Figure 1A is based on meaning-based representations, not on visual properties.

Another interesting aspect of the linearisation process required for scene descriptions is that it takes speakers several fixations and some seconds to be ready to start talking. That this scene apprehension stage reflects the difficulty of the speaking task can be seen from the differences in time to speak across the three tasks: Apprehension times were longest in the experiment in which speakers had to imagine and describe actions that could be performed in the scene (about 2500 ms); shortest when they described the scene from memory (1300 ms), presumably because the apprehension time excluded any processes related to interpreting the scene; and in-between for the open-ended descriptions (almost 1700 ms). Interestingly, and as mentioned previously, these times are in the range of those seen in experiments in which the subjects' task is to describe simple scenes that can be captured in a single sentence. Given that the productions we elicited were several sentences in length, it seems plausible to assume that speakers were not planning the entire protocol before starting to talk, consistent with findings reported by Ferreira and Henderson (1998) for network descriptions. But even the shortest apprehension times we and others have observed are far longer than standard estimates of the time required to extract the gist of a scene (Castelhano & Henderson, 2007, 2008), and also longer than recent estimates of how long people need to apprehend actions in scenes (Hafri et al., 2013; Zwitserlood et al., 2018). We argue, then, that apprehension times in scene description tasks reflect three processes: a mandatory macro-planning stage during which the speaker formulates a linearisation scheme; an optional gist extraction stage (absent if, for example, the speaker generates the description from memory or already has been looking at the scene and knows what it is); and an optional action retrieval stage that is mandatory for scenes depicting events and optional for scenes that represent states (e.g. a room with objects but no animate entities).

Linearisation in language production: theoretical developments

We have argued that, to describe something like a scene which lacks inherent order, speakers need a linearisation strategy. We have argued further that eye movements reflect that strategy. At this point, we want to turn to a more detailed discussion of what eye movements can tell us about the specifics of speakers' linearizations. To start, let's consider a strategy that ignores the scene features and content, which might be one in which speakers simply organise their descriptions by proceeding from top to bottom, left to right. Our data show this is clearly not what speakers do, as can be seen from any of the attentional maps generated from our experiments, and as revealed by the significant correlations between eye movements and meaningful scene areas.

Strategies for describing complex scenes typically are based on scene properties, then, and our approach is to distinguish between two types: visual salience and meaning. If visual salience had predicted attention in scenes during language production, then we would have observed the eye being pulled to visually salient regions of the scene, and then the language production system would prepare a verbal description of the item or area on which the eye landed. But as we clearly observed, visual salience does not predict eye movements once its correlation with meaning is taken into account; instead, it is meaning that predicts where people look. Thus, it appears that the scene content that is relevant is the spatial distribution of meaning as reflected in meaning maps. We can say, then, that the cognitive system pushes the eye to meaningful areas of the scene (identified based on scene gist, the extraction of information about large surfaces and objects, and individual fixations made prior to speaking) and then speakers use the language production system to describe the information conveyed to them visually.

A related possibility is that meaning drives the eye to look at certain objects, which then get described. This is compatible with a recent proposal suggesting that, fundamentally, gaze control is prediction (Henderson, 2017). As Henderson argued, a fixation to a target can be thought of as a prediction about what is likely to be in a location. That is, based on the scene gist and what has been fixated already, the viewer might anticipate other things likely to be in the scene and their likely locations. An anecdote might help to illustrate the idea. One day the first author was cycling to campus along a bike path, and coming towards her was another cyclist wearing a backpack containing a set of crutches. This led her to make a direct eye movement from the backpack containing the crutches to the cyclist's pedals, to verify her prediction that somehow this person was able to cycle with an injured foot, something that struck her as remarkable. The point of this anecdote is that the eye movement from the torso of the cyclist to his feet was triggered by a series of inferences about

the dynamic scene unfolding in front of her, including the inferences that someone carrying crutches is likely injured, and that riding a bike with such an injury is interesting. Thus, we see gaze control as a response to the predictions the cognitive system actively makes, and not merely as a passive reflection of scene meaning.

We add to this idea of gaze control as prediction the suggestion that, in production tasks involving descriptions or elaborations of visual stimuli, not only do people predict a likely object, they may also begin to generate a speech plan for describing that item and its constituent components (e.g. determiners, modifiers, and other parts of speech). For example, consider Figure 1A once again. The standard way psycholinguists think about the relationship between eye movements to visual displays and language production is that they assume the speaker lands on an object – let's say the fan in the frame above the couch – and then they say something like "and there's a fan inside the frame." Next, based on peripheral information and possibly also previous fixations made already, the speaker's visual attention may move from the fan to, for example, the radio. An eye movement would then follow the attentional movement, and the speaker additionally utters something like "and there's a radio." But let's consider an approach that allows some flexibility in the ordering of these processes. Imagine the speaker's attention shifts now from the radio up to the alcoholic beverages on the shelf above. If we think about the speaker as having predicted the presence of those beverages based either on previous fixations or on schema knowledge (or both), we can think of the shift of attention and the eye movement as a prediction – in this case, the prediction that interesting beverages are in that location, and we might also assume that the speaker could begin to describe those bottles before actually landing on them. Thus, the speaker might plan to say "and above the radio there are some liquor bottles" while preparing the saccade to those objects in the scene, well before actually landing on them.

This kind of approach allows visual, attentional, and production processes to be interleaved throughout the scene description task, along the lines of what Gleitman et al. (2007) proposed. In a similar way, Griffin (2004) suggested that fixations to objects precede their mention in an utterance so that the speaker can retain those objects in memory before naming them. But unlike the Gleitman et al. model in which language production processes always follow shifts of attention and eye movements, or the Griffin model in which fixations reinforce memory, in our new proposal we allow language formulation processes to precede the eye movement, based on the speaker's prediction that the object about to be named is in that spot in the visual display. Specifically, we suggest that the cognitive mechanism that linearises fixations through a scene can also linearise productions. This flexible interleaving of attentional, visual, and linguistic processes facilitates incremental production because it discourages delays that would be incurred if we assumed the mouth must always wait for the eye. Exploiting predictions about what might be where, rather than waiting for visual confirmation, would allow for faster planning and execution of utterances. This strategy would likely be effective in most cases, and would only be problematic when predictions are not confirmed. Speakers in Elsner et al.'s (2018) task who underspecified may have done so because they could not use information that would be available in real-world scenes (e.g. scene gist) to generate reasonable predictions, and planning an utterance based on inaccurate predictions increased the likelihood of inaccurate descriptions. With a more flexible architecture like the one we propose here, production can proceed incrementally, and the speaker can avoid pauses and delays that might occur when the mouth has to wait for the eye movement system to deliver something interesting to describe. Of course, pauses do occur because sometimes even the most incremental and efficient system will find itself unable to multitask planning and articulation, but the point is that an interleaved, incremental architecture reduces the likelihood that silent or filled pauses will be required to allow visual and language processes to realign. In addition, this architecture allows for the possibility that what is meaningful (which, as we saw, is the feature of the scene the language production system cares about) may also be flexible. That is, perhaps the meaning map for a scene differs depending on the speaker's task, because the task will affect what is deemed to be relevant and interesting. If the speaker takes her task to be simply to itemise the objects in the scene, the places attention is drawn to might differ from the places prioritised in a situation in which the speaker is supposed to ask thoughtful questions about the scene, state the dollar value of the objects in the scene, or perform some other type of judgment or assessment.

Future directions

This research programme on linearisation strategies in language production, using natural scenes assessed for their meaning and visual salience properties, has confirmed Levelt's argument that the need to linearise is due to limitations on human attention. The attentional system seeks out interesting and meaningful targets to describe, and those target characteristics can be precisely assessed using meaning maps, developed entirely

independently to capture scene semantics to further research in visual cognition. This empirical support for Pim Levelt's original proposals concerning the nature of linearisation in language production is only the beginning of a research programme we hope will provide more detailed information about how attentional, visual, and linguistic processes interact during language production.

In our future work we hope to examine whether meaning maps predict not only where speakers look when describing scenes, but also in what order objects are described. Because meaning maps yield a semantically prioritised landscape of possible fixation targets, fixation sequences likely reflect the hierarchy of object meaningfulness that can be derived from the meaning maps. In addition, we know that scene descriptions begin with a relatively long apprehension stage during which the speaker presumably gathers information about the scene and begins to develop a plan for describing it. We also have observed in our studies that, as described by Shanon (1984), speakers tend to begin their descriptions with a high-level label for the scene, and then they proceed to mention large surfaces and objects. Another central question, then, is how these apprehension phase processes, scene gist extraction, and linearisation plans emerge and get coordinated. Another research strand will be expanding the concept of meaning maps in different ways, allows us to explore to what extent the properties of meaning maps change depending on what aspects of a scene are relevant given the speaker's task. Expansion of our visual stimuli to scenes containing people and other agents is also an important future direction.

Conclusions

Thanks in large part to Pim Levelt's work on language production and especially to his book *Speaking*, production and comprehension are now nearly equal partners in the psycholinguistic enterprise. However, the topic of linearisation in language production has not received much attention over the last 25 or 30 years, and we believe this constitutes a major gap in our understanding of how speaking works. This project is an attempt to open up this conversation, and we hope it will inspire more research on complex, multi-utterance production, both in direct response to visual stimuli and in other communicative situations.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Biederman, I. (1981). On the semantics of a glance at a scene. In *Perceptual organization* (pp. 213–253). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387.

Bock, K. (1987). Exploring levels of processing in sentence production. In *Natural language generation* (pp. 351–363). Dordrecht: Springer.

Bock, K., & Ferreira, V. (2014). Syntactically speaking. In *The Oxford handbook of language production*. Oxford: Oxford University Press.

Bock, K., Irwin, D., & Davidson, D. J. (2004). Putting first things first. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action* (pp. 224–250). New York, NY: Psychology Press.

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

Bock, K., & Loebell, H. (1990). Framing sentences. *Cognition*, *35*, 1–39.

Brown-Schmidt, S., & Konopka, A. E. (2015). Processes of incremental message planning during conversation. *Psychonomic Bulletin & Review*, *22*(3), 833–843.

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*, 592–609.

Bunger, A., Papafragou, A., & Trueswell, J. C. (2013). Event structure influences language production: Evidence from structural priming in motion event description. *Journal of Memory and Language*, *69*(3), 299–323.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 753–763.

Castelhano, M. S., & Henderson, J. M. (2008). The influence of color on the perception of scene gist. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(3), 660–675.

Chang, F., Dell, G. S., Bock, K., & Griffin, Z. M. (2000). Structural priming as implicit learning: A comparison of models of sentence production. *Journal of Psycholinguistic Research*, *29*(2), 217–230.

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, *4*, 170–178.

Cohn, N., Coderre, E., O'Donnell, E., Osterby, A., & Loschky, L. (2018, July 28). The cognitive systems of visual and multimodal narratives. *The 40th Annual Cognitive Science Society Meeting*, Monona Terrace Convention Center, Madison, WI.

Elsner, M., Clarke, A., & Rohde, H. (2018). Visual complexity and its effects on referring expression generation. *Cognitive Science*, *42*(4), 940–973.

Ferreira, F., & Henderson, J. M. (1998). Linearization strategies during language production. *Memory and Cognition*, *26*(1), 88–96.

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

Fromkin, V. (1973). *Speech errors as linguistic evidence*. The Hague: Mouton.

Garrett, M. F. (1988). Processes in language production. *Linguistics: The Cambridge Survey*, *3*, 69–96.

Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, *57* (4), 544–569.

Griffin, Z. M. (2004). Why look? Reasons for speech-related eye movements. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action* (pp. 192–222). New York, NY: Psychology Press.

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

Hafri, A., Papafragou, A., & Trueswell, J. C. (2013). Getting the gist of events: Recognition of two-participant actions from brief displays. *Journal of Experimental Psychology: General*, *142*(3), 880–905.

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.

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

Henderson, J. M., & Ferreira, F. (2004). Scene perception for psycholinguists. In J. M. Henderson, & F. Ferreira (Eds.), *The interface of language, vision, and action* (pp. 1–58). New York, NY: Psychology Press.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, *1*(10), 743–747.

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

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Koehler, K., & Eckstein, M. P. (2017a). Beyond scene gist: Objects guide search more than scene background. *Journal of Experimental Psychology: Human Perception and Performance*, *43*, 1–17.

Koehler, K., & Eckstein, M. P. (2017b). Temporal and peripheral extraction of contextual cues from scenes during visual search. *Journal of Vision*, *17*(2), 1–32.

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

Levelt, W. J. M., Page, R. B. L., & Longuet-Higgins, H. C. (1981). The speaker's linearization problem. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *295*, 305–315.

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

Melinger, A., Branigan, H. P., & Pickering, M. J. (2014). Parallel processing in language production. *Language, Cognition and Neuroscience*, *29*(6), 663–683.

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.

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2018). Meaning guides attention during scene viewing, even when it is irrelevant. *Attention, Perception, & Psychophysics*, *81*(1), 20–34.

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics*, *76*(2), 270–279.

Rehrig, G., Cheng, M., McMahan, B. C., & Shome, R. (in preparation). Why are the batteries in the microwave?: Use of semantic information under uncertainty in a search task. Manuscript in preparation.

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

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786.

van de Velde, M., & Meyer, A. S. (2014). Syntactic flexibility and planning scope: The effect of verb bias on advance planning during sentence recall. *Frontiers in Psychology*, *5*, 1174.

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.

Zwitserlood, P., Bölte, J., Hofmann, R., Meier, C. C., & Dobel, C. (2018). Seeing for speaking: Semantic and lexical information provided by briefly presented, naturalistic action scenes. *PLOS ONE*, *13*(4), 1–22.