Article

Language *and* Speech

Visual Salience Modulates Structure Choice in Relative Clause Production

Language *and* Speech
2014, Vol. 57(2) 163–180

DOI: 10.1177/0023830913495656
las.sagepub.com
SAGE

Jessica L Montag
University of Wisconsin–Madison, USA

Maryellen C MacDonald
University of Wisconsin–Madison, USA

Abstract

The role of visual salience on utterance form was investigated in a picture description study. Participants heard spoken questions about animate or inanimate entities in a picture and produced a relative clause in response. Visual properties of the scenes affected production choices such that less salient inanimate entities tended to yield longer initiation latencies and to be described with passive relative clauses more than visually salient inanimates. We suggest that the participants' question-answering task can change as a function of visual salience of entities in the picture. Less salient entities require a longer visual search of the scene, which causes the speaker to notice or attend more to the non-target competitors in the picture. As a result, it becomes more important in answering the question for the speaker to contrast the target item with a salient competitor. This effect is different from other effects of visual salience, which tend to find that more salient entities take more prominent grammatical roles in the sentence. We interpret this discrepancy as evidence that visual salience does not have a single effect on sentence production, but rather its effect is modulated by task and linguistic context.

Keywords

Sentence production, relative clause, visual salience

Introduction

Speakers who want to convey an idea generally have many alternative utterance forms to accomplish their goal. For example, an English speaker may describe a single scene in a number of different ways: *The dog is chasing the cat, the cat is being chased by the dog, the cat is fleeing from the dog*, and many other variations. The availability of multiple syntactic and lexical alternatives offers

Corresponding author:
Maryellen C MacDonald, Department of Psychology, University of Wisconsin–Madison, 1202 W. Johnson Street, Madison, WI 53706, USA.
Email: mcmacdonald@wisc.edu

speakers a great deal of flexibility, but also a challenge in that successful communication requires converging on only one utterance plan, and not a nonsensical blend of alternative forms, such as *the chasing cat the dog*. The mechanisms by which speakers settle on one utterance form are therefore an important component of research in language production, particularly in grammatical encoding, the early stage in which lexical and grammatical choices are made.

One component of this process of converging on an utterance plan appears to derive from the *accessibility* of concepts and their related words, where accessibility can be defined as the ease with which the entity can fit into a developing utterance plan. Thus, accessibility can vary with factors such as conceptual prominence or frequency, and more accessible elements are thought to enter into planning earlier. A number of studies have suggested that speakers tend to place more salient entities, such animate (Bock, Loebell, & Morey, 1992; Ferreira, 1994; McDonald, Bock, & Kelly, 1993; Tanaka, Branigan, McLean, & Pickering, 2011), prototypical (Onishi, Murphy, & Bock, 2008), or imageable nouns (Bock & Warren, 1985), as well as more easily articulated (Bock, 1987) words and phrases earlier in an utterance plan and/or in a prominent syntactic position (such as grammatical subject) during utterance planning. Immediate conversational context may also modulate the relative accessibility of sentence elements. For example, *given* information, which has previously been mentioned in the discourse, is more accessible than new information (presumably via a mechanism such as lexical/conceptual priming), and so gets placed earlier in the sentence than new information (Bock & Irwin, 1980; Ferreira & Yoshita, 2003; MacWhinney & Bates, 1978; Prat-Sala & Branigan, 2000).

Other external factors, such as qualities of the visual environment, can also affect production choices when speakers are referring to visible entities and events, as in picture description tasks. Obviously the visual scene must play a major role in the sentence used to describe the scene; if the scene consisted of a dog chasing a squirrel rather than a cat, that aspect would be reflected in the utterance. However, more subtle aspects of the visual scene also affect production choices. For example, Bock, Irwin, Davidson, and Levelt (2003) examined Dutch and English speakers' preferences for producing absolute reports of time (e.g., *It's ten fifteen*) versus relative reports (*It's a quarter after ten*) as a function of the visual environment – whether the speakers were looking at a digital or analog clock. The digital display increased the rate of absolute time reports, but the effect of the display varied as a function of the language spoken, as Dutch and English differ in the relative frequency of absolute versus relative time usages. These results suggest that visual and linguistic influences interact during the formation of an utterance plan.

MacDonald (2013) suggested that accessibility effects are emergent from the gating functions of attention, which prioritize easily-retrieved elements from long term memory during the development of the utterance plan. If so, variation in visual salience, the extent to which elements of a visual scene can attract a speaker's attention, should affect the accessibility of visual elements and thereby affect utterance form. For example, when visual cues are provided that deliberately (Tomlin, 1995, 1997) or implicitly (Gleitman, January, Nappa, & Trueswell, 2007) draw speakers' attention to one pictured element, that entity tends to be mentioned early in subsequent utterances, affecting the structure of utterances. For example, for a picture of a dog chasing a mailman, when Gleitman et al. (2007) used a subliminal flash of light near the dog, implicitly drawing attention to this element, descriptions tended to be in the active voice, with the dog as the sentence subject (*The dog is chasing the mailman*), but when attention was drawn to the mailman, descriptions tended to be passive, with the mailman as subject, as in *The mailman is getting chased by the dog*. Gleitman et al. (2007) suggested that the attention-directing cue affected the order of object recognition processes, so that the cued element was recognized first, which in turn affected the accessibility of the linguistic labels for entering an utterance plan. These visual cuing effects on utterance form vary in

strength as a function of the language spoken; they appear to be weaker in languages in which the passive voice is a less viable alternative than in English (Gennari, Mirković, & MacDonald, 2012; Kaiser & Vihman, 2006; Myachykov & Tomlin, 2008). Cross-linguistic differences in cuing effects add more evidence for an interaction between visual, conceptual and linguistic influences during utterance formation.

Although the visual cuing affecting the order of elements in an utterance can be interpreted as the effects of attentional capture (via a flash of light or other visual cue), the effect of visual cuing on structure choice can also be interpreted as altering the message of the utterance. When an object is cued, it becomes the topic or focus of the sentence, and that shift in topic is reflected in the structure choice. Thus, visual salience can interact with the linguistic context of a message, and the nature of the visual salience effect can vary depending on other aspects of the task or context in which the utterance is produced.

Taken together, these results suggest that the visual salience of elements in view clearly affects people's descriptions of those elements, but the effect is not a simple linkage between visually salient elements and early sentence positions. Instead, the effect of visual salience appears to be task dependent, so that visual information may be viewed or interpreted differently depending on the goals of the task. For example, in another study in which speakers reported the time while looking at clocks, Kuchinsky, Bock, and Irwin (2011) found that speakers' eyegaze patterns differed when they were asked to read the time normally (small clock hand for the hour, big hand for the minutes) or reversed (big hand minutes, small hand hours). Their results suggest that top-down, goal-directed factors affect visual gaze patterns in seeking information from the visual scene, so that the effects of visual salience on production choices can be modulated by the speaker's task and goals, and by the language that is being spoken.

The studies described above show how the form of utterances is shaped by the complex interactions between visual salience, task goals, and language spoken, but they do not address one additional potential complication, linguistic context. That is, most language production in the context of a visual environment is also conducted in a linguistic environment, such as a conversation about the visible objects and events, which affect the speaker's goals and intended message in several ways. First, interaction with an interlocutor (such as a question or comment about the visual scene) can affect the speaker's communicative goals, with consequences for the utterance plan (Christianson & Ferreira, 2005; Prat-Sala & Brannigan, 2000). Second, linguistic context of this sort has powerful effects on visual attention and eye movements (e.g., Altmann & Kamide, 2007), and so it is reasonable to expect that hearing a question or other comment about a scene could influence the speaker's visual search of the scene and thus the relative salience of scene elements. We see from previous results that the effect of visual salience varies based on aspects of the task goals and language, and that linguistic context has a profound effect on gaze and direction of attention to visual information. Thus, if we are to understand visual salience as one of many factors that affect production choices, we must investigate the role of visual salience in various task contexts. In line with these goals, our study investigates the joint effects of linguistic context and visual salience in the context of language production, which has not been extensively investigated. Here we present a study that begins to address these issues, with a task that investigates the role of visual salience on production choices in a linguistic context that is unlike that of previous studies.

In our study, native English speakers described pictures in answer to a spoken question such as "what is red?" Such questions do not directly name elements in the visual environment but provide a linguistic context, directing speakers' attention and search of the scene. To formulate their answer, speakers had to focus on an element in the picture and describe its relationship to some other pictured participants. In some trials, the spoken questions lead to descriptions of a visually and

Table 1. Sample utterances in the Gennari et al. (2012) relative clause production task from which our task is adapted.

	Inanimate target: book	Animate target: baby
Active object relative	The book (that) the woman carried	The baby (that) the woman carried
Passive relative	The book (that was) carried by the woman	The baby (that was) carried by the woman

conceptually salient animate entity, and in other trials, the description target was an inanimate entity, which varied in visual salience across pictures. We investigated the kinds of structures speakers produced in their descriptions, and we also collected measures of production difficulty, such as initiation latencies.

We used a relative clause picture description task developed by Gennari et al. (2012). In this task, participants viewed pictures and responded to spoken questions about animate or inanimate entities that were the patient/theme of some pictured action. These questions served to focus the entity in question so that this animate or inanimate element was the head of (i.e., was modified by) a relative clause. Several types of relative clauses can be used to describe this relationship. Speakers can say either sentences containing active object relative clauses or passive relative clauses when describing these animate or inanimate target nouns (Table 1). Previous studies show that speakers' structure preferences are strongly affected by the animacy of the entity being described by the relative clause, among other factors (Gennari & MacDonald, 2009; Gennari et al., 2012; Roland, Dick, & Elman, 2007). Relative clauses headed by an action's animate patient tend to appear primarily in a passive form in English, in which the patient of the action (*baby*) is the subject of the relative clause, and very rarely as active object relatives, or center-embedded relatives, in which the head of the relative clause (*book*) is the direct object of the relative clause verb. Inanimate-headed relatives routinely appear in both passive forms and in active forms. By eliciting relative clause descriptions of animate and inanimate entities such as *baby* and *book*, we can explore the effect of conceptual factors (animate vs. inanimate entities) and visual salience on utterance planning.

Relative clauses are a good choice for investigating these phenomena, for two reasons. First, there are well-known patterns of structure biases (in English, passive bias for animate heads and a fairly even passive/active split for inanimate heads) from which to evaluate the effects of salience (Gennari & MacDonald, 2009; Gennari, et al., 2012; Roland et al., 2007). Second, the relative clause structure unconfounds surface word order and grammatical role assignments to nouns in the sentence. In English main clauses, the grammatical subject appears first in the sentence, and when a passive form is uttered (e.g., *The baby was carried by the woman*), it is unclear whether the high accessibility of *baby* yielded a passive because the assignment of the subject role to *baby* or through a process that placed *baby* early in the linear order, forcing a passive (see Tanaka et al., 2011 for discussion). In both the passive and active object relative forms in relative clauses, however, the head noun described by the relative clause remains in the same early position in the utterance (e.g., *baby* is in the same location in *The baby that was carried by the woman* and *The baby that the woman carried*). Thus, the higher rate of passive relatives describing animate nouns than for inanimate nouns cannot be attributed to early placement of this noun in the linear sequence of words. Instead, the effect appears to owe more to the tendency to assign animate nouns to prominent grammatical roles such as subject (Bock & Levelt, 1994; Bock & Warren, 1985; Ferreira, 1994).

Figure 1 illustrates this hypothesized process of grammatical role assignment in the case of an animate (on the left) or inanimate (on the right) object of the action. When a speaker is charged

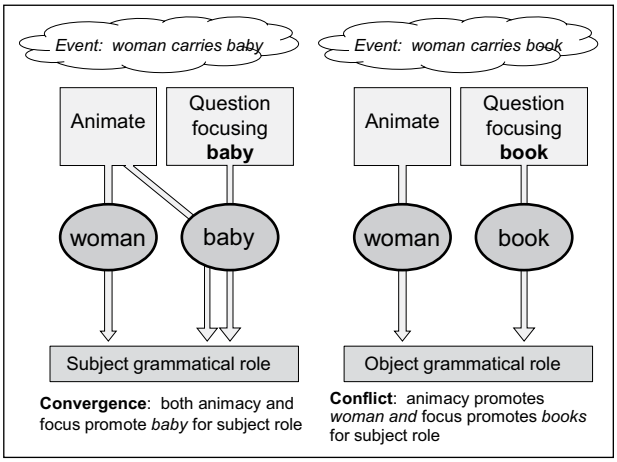

Figure 1. Illustration of converging and conflicting cues to grammatical role assignment during picture description. On the left, animacy of both *woman* and *baby* promote their assignment of the subject grammatical role, but *baby* is also focused by the task question. With *baby* receiving the subject grammatical role, a passive relative clause ensues. On the right, animacy of *woman* supports subject grammatical role assignment while focus to *book* supports subject assignment to this entity, leading to conflict and variability in passive relative versus active object relative structure choice.

with answering a question about an animate entity such as *baby*, this noun is both animate and the focus of the utterance being formulated, and both of these factors contribute to the prominence of the noun. As a result, *baby* is assigned to the most prominent grammatical role, subject, yielding a passive relative clause utterance. In a condition in which speakers are requested to answer a question about an inanimate target item, which is shown on the right of Figure 1, speakers must balance two conflicting grammatical role assignments to place the focal entity (*book*) or the animate agent (*woman*) in subject position. If there is inconsistency in how this conflict is resolved, across speakers and/or across particular pictures/events as a function of the salience of the elements or other factors, then inanimate focused entities should yield a combination of active object and passive relative clause utterances. This conflict might be expected to increase the difficulty of settling on an utterance form, but measures of planning difficulty, such as initiation latency, have not been collected in prior work.

Given a question such as "what is red?" and thus causing participants to search a scene in order to formulate a response, some pictured entities will be found more rapidly than others. For example, Dobel, Gumnior, Bölte, and Zwitserlood (2007) found that in picture description tasks, the animate pictured entities are identified extremely rapidly. This effect likely reflects some combination of their inherent visual salience in a picture (perhaps owing to size, location in the picture, amount of other non-target "clutter" in the scene, and other visual features) and also the attentional focus naturally devoted to searching for animate elements in a scene, especially given a task of describing an action, for which animate entities are likely to play a large role. Thus, visual salience itself is complex and likely to be a combination of picture features, the degree to which the linguistic context (e.g., *what is red?*) directs visual search efficiently in a particular picture, and other task demands influencing visual search.

These considerations concerning the multi-dimensionality of visual salience and its interaction with linguistic context guided our decisions on how to investigate the effects of these factors on

Figure 2. Test pictures for the verbs "throw" (left) and "hug" (right).

utterance form. A traditional factorial design would manipulate the salience of the items at high and low levels, but the generality of any such manipulation is questionable, given the myriad factors that contribute to salience. Recent developments in the use of mixed-effects models have increased the viability of non-factorial designs in which factors such as visual salience can vary continuously (Baayen, Davidson, & Bates, 2008; Jaeger, 2008; Quene & van den Burgh, 2008). We therefore treated visual salience as a continuous and multi-dimensional factor in our study. Specifically, we made no special efforts to manipulate salience of elements in the scenes in our picture description task but instead collected three different measures of how visually salient the critical elements were in each visual scene. We then used these salience measures in multi-level analyses to assess how salience variation affected utterance form.

2 Method

2.1. Participants

One hundred twenty-two undergraduates at the University of Wisconsin–Madison participated in exchange for pay or for extra credit in an introductory psychology course. Of these, 68 participated in the production experiment and 54 participated in one of three visual salience rating tasks. All were native speakers of American English.

2.2. Materials

Twenty verbs that can each take both an animate and inanimate grammatical object were selected. A color cartoon picture was created for each verb or adapted from pictures used by Gennari et al. (2012). Each picture contained two depictions of events named by the verb, one acting upon an animate grammatical object and once acting upon an inanimate grammatical object. For example, the pictures for the verbs *throw* and *hug* are shown in Figure 2. Each picture shows an animate entity – a man – as the object of the action, and also an inanimate direct object – a ball in the case of *throw* and a toy for the *hug* picture. The animate and inanimate objects of the action were the target items in the experiment. Each picture also contained other elements, always including one or more additional inanimate and animate elements matching the target items. Thus the *throw* picture in Figure 2 includes a second ball and several other men in addition to the man throwing/ball being thrown, and the *hug* picture contains another man and another toy in addition to the man/ball being

thrown and man/toy being hugged. These extra elements increased the specificity of speakers' descriptions of the target elements in order to distinguish them from other similar elements in the pictures.

To elicit speech, spoken questions were recorded for each picture, and the participants' task was to answer the question presented with the picture. Questions for experimental trials asked participants to describe a particular target person or object in the picture. For example, questions corresponding to Figure 2 would be "Who is wearing orange?" to elicit a description of the animate "man" target wearing an orange jacket and "What is red?" for the inanimate "ball" target being thrown by a man. There are multiple men and balls in the picture, so participants had to further describe the target item in order to identify it.

Forty-three filler pictures were included to reduce strategic effects and structural priming (the repetition of utterance sentence structure from one trial to the next) (Bock, 1986). For filler trials, participants were asked to describe what a particular person was doing or identify a particular object; these items were designed to elicit simple sentences without relative clauses.

2.3. Production experiment procedure

Participants in the production experiment first completed a pre-exposure task in order to familiarize participants with the drawing style of the pictures, to encourage uniform verb usage (e.g., to use *carry* rather than *hold* for the picture showing carrying) and to normalize lexical retrieval times across verbs. In the pre-exposure, participants viewed a small portion of a picture, depicting a single action (for example, the portion showing the man throwing the ball in the *throw* picture of Figure 2). These smaller pictures contained only the target animate or inanimate entity and the agent acting upon that entity. After two seconds, a verb describing the action appeared underneath the picture, in the case of experimental picture segments, or a noun appeared for filler picture segments. Participants were instructed to read aloud the word underneath the picture. Participants viewed, in random order, two picture segments for experimental pictures (one segment showing action on an animate entity and one acting on inanimate object) and one picture segment for filler pictures. Fillers were included so that all pictures in the main experiment would have had some pre-exposure.

After the pre-exposure phase, participants performed the main experiment, using a variant of the task developed by Gennari et al. (2012). Participants were told that they would view pictures and answer questions about them. They were provided with a cover story in the instructions that promoted the use of action descriptions with relative clauses in their responses. Participants were told that their picture descriptions (their responses to the questions) would be shown to a later group of participants who would try to guess which picture elements their responses described. They were told that superficial changes would be made in the pictures when they were shown to the new participants, and so to be clear in their descriptions, participants should describe the actions in which the pictured people and objects were taking part. The combination of encouraging action descriptions in the instructions and the presence of multiple entities in the picture (encouraging modification of the target noun) elicited a high rate of relative clause responses without any explicit instruction to produce them.

At the start of each trial, a picture appeared on the screen and remained onscreen throughout a trial, until the participant pressed a key to go on to the next trial. Three seconds after the picture appeared, participants heard a recorded question asking about a target person or object in the scene, such as "What is red?" Participants answered by speaking into a microphone; initiation latencies and all responses were digitally recorded for later analysis. These methodological choices formed

the primary differences from those in Gennari et al. (2012), in which no initiation latencies were collected and most studies presented written rather than auditory questions and collected written productions. Using a methodology that allows for the collection of initiation latencies should allow for additional investigation of production processes and sources of production difficulty.

Animate and inanimate questions for experimental items were counterbalanced across participants so that each participant saw each picture only once and received 10 trials with a question about an animate patient (e.g., the man being thrown in Figure 2) and 10 trials with an inanimate theme question (e.g., the ball being thrown in Figure 2). Test and filler trials were pseudo-randomized such that there were always at least two fillers between any two test trials.

2.4. Visual salience procedure

Because visual salience is a complex notion and not simply the effect of a single factor such as size, we collected three separate measures of the visual salience of target entities. For example in Figure 2, a number of features could contribute to greater salience of the bear being hugged than the ball being thrown. These include the relative size of these entities (the bear is bigger than the ball), their location in the picture (the bear is in the foreground of its picture, the ball in the background), the salience of the agent acting on the object (the hugging girl is in the foreground and larger than the man throwing the ball), the color, and other features. Our salience measures were aimed at capturing variation of this sort.

The first salience measure was a visual salience rating task of the target items, in which 17 participants who did not take part in the main experiment explicitly rated the visual salience of the target entities (e.g., the ball). In this task, participants first passively viewed all 20 pictures one at a time, pacing themselves through the entire set, in order to become familiar with the drawing style and events depicted in the pictures. This pre-exposure was followed by the rating task. On each trial, participants saw one of the 20 experimental pictures. They read a description of a target item (e.g., *The ball that is red*), and indicated how visually salient or "easy to locate" the named element was in the picture, using a 1–7 scale, with lower numbers indicating greater visual salience.

In the second salience task, a different group of 17 participants rated the visual salience of the animate agents acting on the inanimate target items (e.g., the man throwing the ball). We reasoned that especially for inanimate objects, identification of the agent of the action could be crucial to identifying the target entity itself. For example, identifying the man and his throwing action may help identify the item near his hand as a ball. The procedure for the agent salience task was identical to the first rating task, with the only difference being that the participants received a statement referring to an agent (e.g., *the man wearing red*) and rated on a 1–7 scale how easy that entity was to locate in the scene.

The third task was designed to be an implicit measure of visual salience of the target items. An additional 20 participants who did not participate in either the other rating tasks or main task viewed each of the 20 experimental pictures once, and the inanimate target entity was probed with a spoken question, such as "What is red?" The participants were instructed to locate the inanimate target item[1] as quickly as possible and name it as soon as they located it, always using the frame *the noun* (e.g., *the ball*), with no other words. The dependent measure was thus not an explicit rating but instead the latency to name the target entity. While these latencies were likely affected by lexical retrieval times of the target object nouns in addition to their visual salience, we were mindful of prior research suggesting that visual salience is influenced by task demands (Kuchinsky

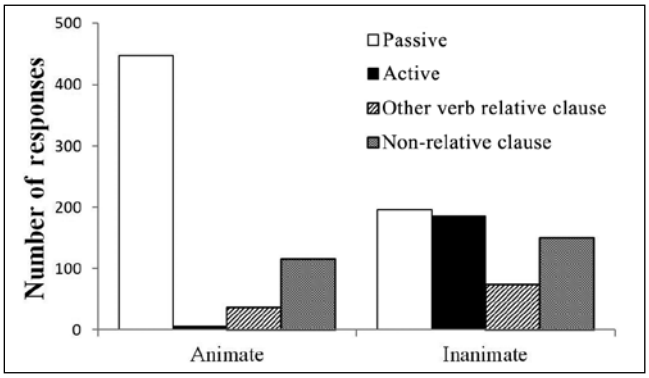

Figure 3. Number of passive, active (object relative), relative clauses using a different verb, and non-relative clause ("other") responses. These counts exclude the verb *spray*, which was removed from all analyses.

et al., 2011). We therefore sought to include a salience measure that had similar task demands to the picture descriptions in the main experimental task.

Ratings and initiation latencies from the three rating tasks were then *z*-scored by subject. The target salience rating task (Task 1) correlated with both explicit agent salience ratings (Task 2), $r(15) = 0.44$, $p < 0.06$, and with object naming times (Task 3), $r(15) = 0.65$, $p < 0.01$, such that objects given a high salience rating tended to be acted on by agents with a high salience rating and tended to be named more quickly in the object naming task. The correlation between agent salience ratings and object naming times was in the same direction but was not reliable, $r(15) = 0.28$, $p < 0.3$. Given that the object salience ratings correlated with the other two measures, and given the view that salience is composed of many subcomponents, we summed these three *z*-scored measures into a single composite measure of visual salience.

3 Results

3.1. Production data

Before analysis of relative clause choice and production difficulty measures, several types of irrelevant trials were removed, including utterances that were not relative clauses (e.g., *the ball*, or *the ball on the right*), affecting 17.6% of animate trials and 25.8% of inanimate trials. We suspect the higher exclusion rate among inanimate trials is due to the fact that participants occasionally failed to locate some inanimate competitors in the pictures and subsequently produced a simple noun phrase (e.g., *the ball*) that did not distinguish the target from similar elements in the scene. Participants almost never failed to notice an animate competitor, with the result that animates were more often modified. Relative clauses in which the participant produced a different verb than the one provided in pre-training were also excluded from analysis, affecting 8.8% of animate and 12.3% of inanimate trials. Inclusion of these different-verb trials in the analyses reported below did not change any of the results; we took the more conservative path of excluding these trials so that animate and inanimate trials would not vary in their rate of conformity to experiment instructions. As most responses for the verb *spray* contained a different verb, all trials for this verb were removed.

One participant was excluded for producing almost no action descriptions and three were excluded for having mean initiation latencies more than two standard deviations from the mean of

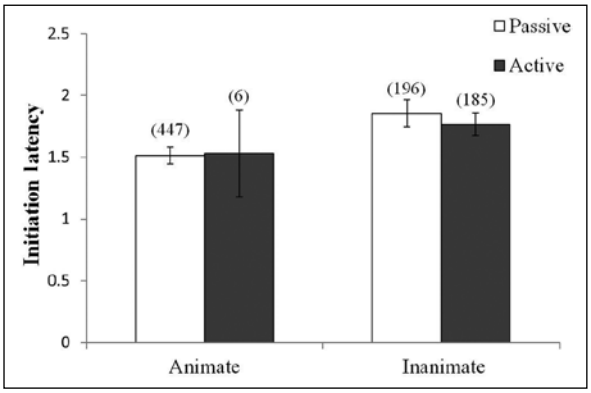

Figure 4. Mean initiation latency (by subject) as a function of target animacy and utterance structure. Numbers in parentheses refer to the number of observations contained in each bar.

all participants. Utterances with disfluencies or initiation latencies longer than six seconds were removed (n = 25). A total of 834 relative clause utterances from 64 participants were analyzed (65% of total utterances). This rate is similar to that in Gennari et al. (2012) and not unexpected given the inherent difficulty in depicting actions with high name agreement and the fact that participants were not explicitly instructed to produce relative clauses.

Participants' responses were coded either as active or passive relative clauses. The rates of these utterances are shown in Figure 3. As expected, relative clauses with animate target head nouns were overwhelmingly passive (98.7% passives, *SD* 4.1) while inanimate targets were nearly evenly split between active and passive relative clauses, with 47.3% (*SD* 38.6) passives. This result replicates previous experimental and corpus results (Gennari & MacDonald, 2009; Gennari et al., 2012; Roland et al., 2007). Interestingly, there were substantial individual differences in the proportions of active versus passive utterances for inanimate targets. About a third of participants produced almost exclusively active object relatives, about one third produced almost exclusively passive relatives and the remaining third produced nearly equal numbers of passives and actives. This study was not designed to examine individual differences, and the small number of observations on which these patterns are based (10 inanimate trials per participant) make it difficult to draw any conclusions from these results. However, future research on individual differences should further investigate individual differences in structure choice.

Two features of the utterances were coded to assess production difficulty across conditions: initiation latencies and relative pronoun use. These measures of production difficulty are known to correlate with difficulty (e.g., length) of upcoming phrases (*initiation latency*: Ferreira, 1991; Ferreira, 1996; *relative pronoun use*: Ferreira & Firato, 2002; Jaeger, 2005; Race & MacDonald, 2003). Speech initiation latencies, defined as the time from the offset of the spoken question to the onset of the response, were measured from the digitized recording of the utterance. Initiation latencies are contained in Figure 4. Relative pronoun use was defined as the presence or absence of an optional relative pronoun (*that, who, whom*). We investigated the effect of these measures of production difficulty on the animacy of the target (head) noun, the response structure and the number of words in the head of the relative clause (e.g., *the ball*, ranging from 1–8 words) and the number of words in the relative clause itself (range = 2–18 words), excluding any relative pronoun.

Table 2. Results of mixed-effects model predicting initiation latency from head-noun animacy. The model contained intercepts and by-subjects (*s*) and by-items (*i*) slopes.

	Effect of animacy on initiation latency			
	Coefficient	SE	*t*	Random slope
Intercept	1.69	0.09	18.54*	
Animacy	–0.33	0.13	–2.58*	*s*, *i*

*A coefficient is considered a significant predictor if $|t|>2$ (Baayen, 2008).

We conducted several analyses on these initiation latencies to investigate whether this complex picture description task yielded similar effects to those previously reported with other generally simpler pictures and tasks. First, a linear mixed effects regression (lmer) analysis (Baayen et al., 2008) using the lme4 package in R (Bates, Maechler, & Bolker, 2011) was conducted, with both participants and items (pictures) as random effects and animacy (animate or inanimate target) included as a fixed effect. In all models, random slopes were included only if they significantly improved model fit (Baayen et al., 2008). We found a significant effect of animacy on initiation latencies, such that latencies to animate entities tended to be shorter than those to inanimate entities targets (Table 2). Model fit did not improve when response type (active or passive) was included in the model, suggesting that the animacy of the target, not the utterance type, predicted initiation latencies.

Consistent with previous findings (Ferreira & Firato, 2002; Jaeger, 2005; Race & MacDonald, 2003), number of words in the relative clause was a significant predictor of relative pronoun use (the optional words *that, who* or *which* at the start of the relative clause), as shown in Table 3. Longer relative clauses were more likely to be preceded by relative pronouns, and there was a relative clause length by response type interaction. Active utterances more often contained relative pronouns, but only with longer relative clauses (approximately greater than five words). It is unclear why this length effect would be particularly strong in the active utterances. Perhaps the overall rarity of the active object relative clauses contributed to increased planning difficulty. Previous research has also suggested that the length of the head noun (e.g., *the ball*) influences relative pronoun use (Jaeger, 2005), but as the head nouns in this experiment were almost all two words long, it was impossible to examine the effects of head noun length in this study.

3.2. Visual salience

Given this evidence that the picture description task yields familiar, interpretable data for utterance choices, initiation latencies, and relative pronoun use, we next examined the relationship between visual salience and production measures for inanimate trials (recall that animates had essentially no variation in structure choice and were all highly accessible and salient, similar to results in other studies, e.g., Dobel et al., 2007). First, as shown in Table 4, visual salience (the composite of the three visual salience measures) was a reliable predictor of time to begin speaking in inanimate target trials, such that responses to more visually salient inanimate entities tended to have shorter initiation latencies than responses to less salient inanimates. This result is what would be expected if participants did in fact take more time to locate the less-salient objects in the pictures.

More interestingly, visual salience also affected the structure of the utterance produced in inanimate trials; analyses are shown in Table 5. The direction of the effect, shown in Figure 5, was that

Table 3. Results of mixed-effects logistic model (Jaeger, 2008) predicting relative pronoun usage from number of upcoming words in the relative clause and response structure (active or passive).

	Relative pronoun usage			*p*	Random slope
	Coefficient	*SE*	*z*		
Intercept	–4.85	0.94	–5.17*	$p < 0.001$	
Words in RC	0.95	0.18	5.39*	$p < 0.001$	
Response	–1.03	1.85	–0.56	$p = 0.54$	*s*
Words in RC × response	0.94	0.35	2.67*	$p < 0.01$	

RC: Relative clause; s: by-subjects.
*Significant values.

Table 4. Results of linear mixed-effects model predicting initiation latency of inanimate target trials from visual salience of the target item. The analysis included only inanimate trials because all animate target entities were highly visually salient human entities.

	Initiation latency			
	Coefficient	*SE*	*t*	Random slope
Intercept	1.93	0.12	16.48*	
Salience	0.8	0.09	4.43*	*s, i*

s: by-subjects; i: by-items.
*Significant values.

Table 5. Results of mixed-effects logistic model predicting active (versus passive) produced sentence structure on inanimate target trials from visual salience of the target item.

	Sentence structure				
	Coefficient	SE	*z*	*p*	Random slope
Intercept	–0.41	0.67	–0.062	$p = 0.95$	
Salience	1.77	0.79	2.23	$p = 0.026$*	*s, i*

s: by-subjects; i: by-items.
*Significant values.

the more visually salient inanimate targets were more likely to be described with active object relative constructions such as *The toy that the girl is hugging* than the less visually salient inanimate items, which tended to yield more passive relative clauses such as *the ball that's being thrown by the man*. This finding is notable because animate entities, which are overall highly salient, were almost exclusively described with passive relatives, and thus highly salient animate beings and low-salient inanimate objects patterned together in their tendency to be described with passive relative clauses, while an intermediate case, salient inanimate entities, tended to be described with active object relative clauses. This is an unexpected finding if visual salience alone is driving speakers' structure choices. Minimally, these results suggest that in a task in which people describe pictured events, visual salience appears to have different effects on animate entities, who are viable potential agents of actions and inanimate entities, who are not likely to be event agents. We consider these visual salience effects further in the discussion.

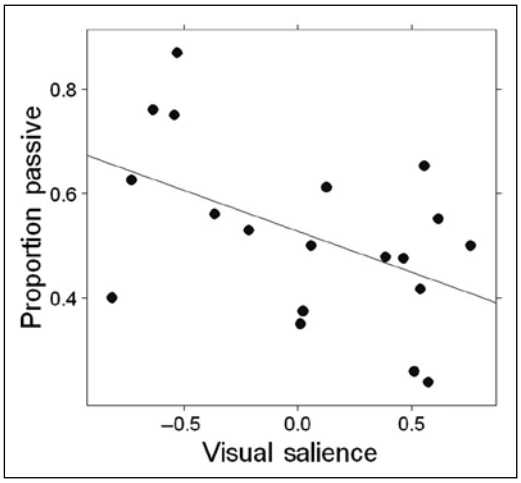

Figure 5. Proportion of relative clauses describing inanimate entities produced as passive structures as a function of visual salience of the entity described. (r = 0.49).

We suspect that a better understanding of the complex visual salience effects requires attention to the task demands that are inherent in situations in which speakers are jointly guided by both visual and linguistic context. First, in our task speakers described actions in their replies to questions, and so it makes sense that animates (potential agents of actions) yielded different production behavior than inanimate entities, which are typically not the source of actions. A more subtle second issue is the fact that task or communicative demands can change over the course of a trial. That is, search times varied with visual salience, and the longer a participant searched the scene for the target object that was necessary to answer the question, the more the participant will have inspected non-target aspects of the scene during the visual search. We hypothesize that variation in the amount of visual search actually changes the task and communicative goals for the speaker: the longer search times to locate the less-visually salient entities make the participants more likely to notice other items in the scene, including the competitor objects in the picture (e.g., the non-target ball). Consistent with this view, error trials for inanimate items were numerically more likely to contain an error related to failing to distinguish the target item from competitors (e.g., saying only *the ball* in the presence of a competitor ball) when the target entity was more visually salient, r = .39, p = 0.1, though this pattern failed to reach significance due to the small number of errors of this type (77 across all inanimate trials) and the lack of statistical power stemming from having only 19 items. Nonetheless, this suggests that participants really were less likely to notice or focus on competitors when the target inanimate item was highly visually salient.

These results suggest that speakers' additional focus on competitor objects, in conjunction with the linguistic context asking a question about a particular entity, encourages the speaker to be more explicitly contrastive – to distinguish the target item from the other similar item in the picture, rather than to just describe one of many items in a scene. This goal is best achieved by using a passive relative, in which the speaker focuses on the target entity and that entity becomes the subject of the relative clause. In particular, a passive relative allows the agent of the action to be dropped (e.g., *The ball being thrown* as opposed to *The ball being thrown by the man*), allowing the speaker to focus on the target entity and avoid explicit mention of any other pictured entity. We investigated this hypothesis by contrasting the rate of agentless passives in participants' utterances. Previous

work on agent-dropping has linked it to semantic interference; Gennari et al. (2012) found higher rates of agentless passives when the agent and patient were more semantically similar. This is consistent with our data, where we found a higher rate of agent dropping for animate targets (Animate: 54.6% of passives were agentless, Inanimate: 35.7% of passives were agentless, $t(18) = 4.08$, $p < 0.001$). However, a closer look at agent-dropping within the inanimate targets suggests that semantic similarity cannot be the whole story, because there was a correlation between agent-dropping and visual salience, such that agentless passives were produced often for less salient inanimate targets, $r = 0.48$, $p < 0.05$. Thus, once again the low-salience inanimate entities patterned with the highly salient animate targets. These data suggest that a number of factors may affect agent omission, but within the inanimate entities, the higher rate of agent omission for low salience entities is consistent with the suggestion that agent dropping is one element of the passive, in addition to making the target the sentence subject, that speakers can use to focus the target element.

This hypothesis suggests that subtle changes in the properties of the visual scene can affect structure choice through changing the intended message of the utterance. Describing an item in a scene is a subtly different task than describing that item with the goal of distinguishing it from a competitor. If the increased search time to less salient objects makes participants more likely to notice competitor objects and thus give responses with a more explicitly contrastive element, this would be an additional example of an interaction between visual and linguistic contexts, which in this situation could be driving our visual salience effect.

An alternative hypothesis, suggested by a reviewer, is that the salience effects observed here are not being driven by the salience of the target per se but by the salience of the agent acting on the target. Indeed, these two factors (both assessed in our salience measures) are positively correlated, as reported above. As a result, the present study cannot tease apart the effects of agent-salience and target-salience. In our visual salience ratings, the salience of the agent acting on the inanimate targets was correlated with the salience of the target itself, so we treated the salience of the agent acting on the inanimate target entity as contributing to the salience of that target, as a more salient agent should make the segment of the picture depicting the action more salient overall. It is possible pictures designed to contrast salience in this way could better distinguish these factors. Because we observed almost no object relative utterances describing animate target entities, which have visually salient agents acting upon them, we do not anticipate large independent effects of agent salience (that is, beyond the fact that salient agents may speed the identification of the target and identification of the action being done on the target). This suggestion awaits further testing.

4 Discussion

This study investigated the nature of visual salience and its potential interactions with linguistic context to influence speakers' utterance choices during language production. Although more work remains to be done to elucidate these effects, this study made progress in several domains. First, we established that visual salience in a complex scene can be assessed with a composite of several different measures, and that this composite can reliably assess visual search and utterance formulation processes, as indexed by speech initiation latencies. Second, we confirmed that animacy of the entity to be described, which is strongly tied to visual and conceptual salience, has a robust effect on speakers' utterance forms, with animate entities routinely described with passive relatives, while inanimates were described with a mix of passive relatives and active object relatives. Third, we showed that the link between visual salience and utterance form is not straightforward, and it likely interacts with linguistic context in complex ways.

The important visual salience result in this study was that visual salience alone cannot predict structure choice in language production. While the (always salient) animate entities yielded passive relatives, inanimate entities were more likely to yield passive descriptions when they were less salient in the visual scene. These results stand in contrast to ones obtained in other tasks that have manipulated visual salience (Gleitman et al., 2007; Tomlin, 1997), which generally have found a much simpler relationship between salience and structure, in that increases in visual salience yielded increases in making the target the sentence subject. These tasks did not contain animacy manipulations, a linguistic context, visual competitors, or, crucially, a contrastive element, unlike our question-answering task and more complex scenes. The more complex contexts of our task reveal interactions among factors that can arise when both linguistic and visual context are present.

This complex relationship between task, linguistic context and visual salience is consistent with the findings of Kuchinsky et al. (2011), who found that task is a better predictor of gaze patterns than visual scene properties, suggesting that a speaker's communicative goals can drive both the pattern of scanning a scene and the speaker's utterance plan, and is also consistent with other work that describes a more task-dependent relationship between visual properties of scenes and behavior (Brown-Schmidt, Byron, & Tanenhaus, 2005; Kaiser, Runner, Sussman, & Tanenhaus, 2009). The effect of visual salience is moderated by the goals of the speaker, so we expect the role of visual salience to differ based on the context of the utterance. Thus, we expect that visual salience may not play the same role in a task in which speakers describe a simple scene containing a salient element (Gleitman et al., 2007; Tomlin, 1997), versus when the visual salience of scene elements stay constant as the task changes (Kuchinsky et al., 2011), versus when items that vary in visual salience must be contrasted with competitors (as in the present study). These data are not necessarily in conflict with each other, but rather reflect the task-dependence of visual salience effects. There is also no evidence that any of these tasks are more or less natural than the others – "natural" language production spans many linguistic and visual environments, and all of these tasks share elements of language production as it happens outside the laboratory. It is thus not surprising that the effect of visual salience could interact with task, because production itself occurs in varied situations and speaker goals, only some of which include referring to elements in the visual environment.

A potentially related interpretation of our results is that they are an example of audience design (Brennan & Clark, 1996), in which speakers tailor their utterances to respond to perceived listener needs. For example, when a speaker has spent a relatively long time finding and identifying a low-salience inanimate pictured entity, he/she senses that a listener will have similar difficulty when viewing the scene. To ameliorate this inferred difficulty, the speaker could assign the low-salience entity to the prominent subject grammatical role and thereby describe the entity with a passive, topicalizing it and focusing it for the listener, thereby presumably aiding identification in the visual scene. This view ascribes very direct attention to listener needs on the part of the speaker (Brennan & Clark, 1996), in that the speaker's own difficulty in visual search leads to inferences about the listener's state and needs, which in turn shape the speaker's choice of utterance form. One concern about this interpretation is that there is increasing evidence that speakers may not in fact be formulating sophisticated assessments of listener needs during sentence production (Barr, Gann & Pierce, 2011; Keysar, Barr, & Lin, 2003; Wardlow Lane & Ferreira, 2008), nor with work suggesting that production behavior that could be interpreted as audience design actually has a speaker-internal origin (e.g., Bard et al., 2000; Horton & Keysar, 1996). These concerns lead us to prefer the first alternative, which requires fewer inferences on the part of the speaker about the state of the listener. On that view, long search times lead to a complex visual

scene representation for the speaker, which then influences the speaker's pragmatic interpretation of the question, such as "What is red?" We suggest that a felicitous answer to that question varies as a function of scene complexity, so that speakers are more explicitly contrastive (using a passive construction) on those occasions when they have become most aware of competitor objects in the scene. This contrastive focus of non-salient objects is necessarily for the audience (the speaker is answering a question, after all), but it does not assume complex inferencing about the listener's cognitive state. This view is therefore more consistent with approaches in which speakers use their own cognitive state to guide utterance choices instead of generating inferences about the listener (Horton & Keysar, 1996).

These two perspectives are not mutually exclusive, in that adjustments to utterance form can have both a speaker-internal and a more inference-based audience design origin (MacDonald, 2013). Additional work is required to distinguish these two accounts and identify potential interaction between these influences. These approaches share an important feature that also merits more research, namely the claim that the act of visual search itself (its duration and its outcome) can change task demands in a way that can shape utterance form. This view and the current results are valuable in promoting the joint study of linguistic and visual context and suggest that considering the linguistic context of the utterance in conjunction with the visual salience of entities in a scene can provide a more coherent account of structure choices in language production than when either of these two factors are considered individually.

Funding

This research was supported by the National Institutes of Child Health and Human Development [Grants T32 HD049899 and R01 HD047425]; the National Science Foundation [Grant number 1123788]; and the Wisconsin Alumni Research Fund.

Note

1. We do not report latencies to locate and name animate entities as we found it prohibitively difficult to get participants to produce only a noun phrase (e.g., the man) when there were multiple men in the picture. Participants almost always produced an additional description of the animate noun. This tendency to provide an additional description rarely occurred with inanimate nouns.

References

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*, *54*, 502–518.

Baayen, R. H. (2008) *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge, UK: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, *42*, 1–22.

Barr, D. J., Gann, T. M., & Pierce, R. S. (2011). Anticipatory baseline effects and information integration in visual world studies. Acta psychologica, 137, 201–207.

Bates, D. M., Maechler, M., & Bolker, B. (2011). Lme4: Linear mixed-effects models using S4 classes. R package version 0.999375–40.

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

Bock, J. K., & Irwin, D. E. (1980). Syntactic effects of information availability in sentence production. *Journal of Verbal Memory and verbal Behavior*, *19*, 467–484.

Bock, J. K., & Levelt, W. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of Psycholinguistics* (pp. 945–983). San Diego, CA: Academic Press.

Bock, J. K., Loebell, H., & Morey, R. (1992). From conceptual roles to structural relations: Bridging the syntactic cleft. *Psychological Review*, *99*, 150–171.

Bock, J. K., & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formation. *Cognition*, *21*, 47–67.

Bock, K. (1987). An effect of the accessibility of word forms on sentence structures. *Journal of Memory and Language*, *26*, 119–137.

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

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1482–1493.

Brown-Schmidt, S., Byron, D. K., & Tanenhaus, M. K. (2005). Beyond salience: Interpretation of personal and demonstrative pronouns. *Journal of Memory and Language*, *53*, 292–313.

Christianson, K., & Ferreira, F. (2005). Conceptual accessibility and sentence production in a free word order language (Odawa). *Cognition*, *98*, 105–135.

Dobel, C., Gumnior, H., Bölte, J., & Zwitserlood, P. (2007). Describing scenes hardly seen. *Acta Psychologica*, *125*, 129–143.

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, *30*, 210–233.

Ferreira, F. (1994). Choice of passive voice is affected by verb type and animacy. *Journal of Memory and Language*, *33*, 715–736.

Ferreira, V. S. (1996). Is it better to give than to donate? Syntactic flexibility in language production. *Journal of Memory and Language*, 35, 724–755.

Ferreira, V. S., & Firato, C. E. (2002). Proactive interference effects on sentence production. *Psychonomic Bulletin and Review*, *9*, 795–800.

Ferreira, V. S., & Yoshita, H. (2003). Given-new ordering effects on the production of scrambled sentences in Japanese. *Journal of Psycholinguistic Research*, *32*, 669–692.

Gennari, S. P., & MacDonald, M. C. (2009). Linking production and comprehension processes: The case of relative clauses. *Cognition*, *111*, 1–23.

Gennari, S. P., Mirković, J., & MacDonald, M. C. (2012). Animacy and competition in relative clause production: A cross-linguistic investigation. *Cognitive Psychology*, *65*, 141–176.

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*, 544–569.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? Cognition, 59, 91–117.

Jaeger, T. F. (2005). Optional *that* indicates production difficulty: Evidence from disfluencies. *Proceedings of DiSS'05, disfluency in spontaneous speech workshop*, 10–12 September 2005, Aix-en-Provence, France, 103–109.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Kaiser, E., Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2009). Structural and semantic constraints on the resolution of pronouns and reflexives. *Cognition*, *112*, 55–80.

Kaiser, E., & Vihman, V. (2006). Invisible arguments: Effects of demotion in Estonian and Finnish. In B. Lyngfelt & T. Solstad (Eds.), *Demoting the agent: Passive and other voice-related phenomena* (pp. 111–141). Amsterdam, the Netherlands: John Benjamins.

Keysar, B., Barr, D. J., & Lin, S. (2003) Limits on theory of mind use in adults. *Cognition*, *89*, 25–41.

Kuchinksy, S. E., Bock, K., & Irwin, D. E. (2011). Reversing the hands of time: Changing the mapping from seeing to saying. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37*, 748–756.

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

MacWhinney, B., & Bates, E. (1978). Sentential devices for conveying givenness and newness: A cross-cultural developmental study. *Journal of Verbal Learning and Verbal Behavior*, *17*, 539–558.

McDonald, J. L., Bock, K., & Kelly, M. H. (1993). Word and world order: Semantic, phonological and metrical determinants of serial order. *Cognitive Psychology*, *25*, 188–230.

Myachykov, A., & Tomlin, R. S. (2008). Perceptual priming and structural choice in Russian sentence production. *Journal of Cognitive Science*, *6*, 31–48.

Onishi, K. H., Murphy, G. L, & Bock, J. K. (2008). Prototypicality in sentence production. *Cognitive Psychology*, *56*, 103–141.

Prat-Sala, M., & Branigan, H. P. (2000). Discourse constraints on syntactic processing in language production: A cross-linguistic study in English and Spanish. *Journal of Memory and Language*, *42*, 168–182.

Quene, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*, 413–425.

Race, D. S., & MacDonald, M. C. (2003). The use of "that" in the production and comprehension of object relative clauses. In R. Alterman and D. Kirsh (Eds) *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, pp. 946–951. Boston, MA: Cognitive Science Society.

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, *57*, 348–379.

Tanaka, M. N., Branigan, H. P., McLean, J. F., & Pickering, M. J. (2011). Conceptual influences in word order and voice in sentence production: Evidence from Japanese. *Journal of Memory and Language*, *65*, 318–330.

Tomlin, R. S. (1995). Focal attention, voice, and word order. In P. Downing & M. Noonan (Eds.), *Word order in discourse* (pp. 517–554). Amsterdam, the Netherlands: John Benjamins.

Tomlin, R. S. (1997). Mapping conceptual representations into linguistic representations: The role of attention in grammar. In J. Nuyts & E. Pederson (Eds.), *Language and conceptualization* (pp. 162–189). Cambridge, UK: Cambridge University Press.

Wardlow Lane, L., & Ferreira, V. S. (2008). Speaker-external versus speaker-internal forces on utterance form: Do cognitive demands override threats to referential success? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *34*, 1466–1481.