# Sentence Production in Naturalistic Scenes with Referential Ambiguity

**Moreno I. Coco (M.I.Coco@sms.ed.ac.uk)** and
**Frank Keller (keller@inf.ed.ac.uk)**
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK

## Abstract

Language production often happens in a visual context, for example when a speaker describes a picture. This raises the question whether visual factors interact with conceptual factors during linguistic encoding. To address this question, we present an eye-tracking experiment that manipulates visual clutter (density of objects in the scene) and animacy in a sentence production task using naturalistic, referentially ambiguous scenes. We found that clutter leads to more fixations on target objects before they are mentioned, contrary to results for visual search, and that this effect is modulated by animacy. We also tested the eye-voice span hypothesis (objects are fixated before they are mentioned), and found that a significantly more complex pattern obtains in naturalistic, referentially ambiguous scenes.

**Keywords:** language production; eye-tracking; naturalistic scenes; eye-voice span; referential ambiguity.

## Introduction

Language production often happens in a visual context, for example when the speaker describes a picture, gives directions on a map, or explains the function of an artifact. In these situations, the speaker needs to select which objects to talk about, and in which order. He/she also needs to disambiguate the utterance referentially. For instance, if there are multiple clipboards in the visual context, then the speaker has to encode additional visual information to pick out one of them uniquely (e.g., *the brown clipboard* or *the clipboard on the table*).

Most work in psycholinguistics has dealt with isolated sentences, but there is some existing research investigating how language is processed in a visual context. A prominent line of research employs the visual world paradigm (VWP; Tanenhaus et al. 1995; Altmann and Kamide 1999) for this purpose. In a typical VWP study, participants' eye-movements are recorded while they view a visual scene and listen to a sentence at the same time. Some VWP experiments have investigated language production; the most well-known example is Griffin and Bock's (2000) study, in which participants were asked to describe line drawings depicting two objects (e.g., a turtle and a kangaroo) performing a transitive event (e.g., splashing). The key finding of this study was that speakers fixate visual referents in the order in which they are mentioned, and they begin fixating an object about 900 ms before naming it. The span between fixating and naming a referent is known as the **eye-voice span**; other studies (e.g., Qu and Chai 2008) have reported eye-voice spans consistent with those found by Griffin and Bock (2000).

The aim of the present paper is to establish whether the simple relationship between language production and eye-movements implied by the eye-voice span extends to more realistic situations. We investigate language production in a visual context that consists of naturalistic scenes (rather than line drawings) and in which multiple objects can correspond to a given linguistic referent (in contrast to Griffin and Bock 2000). This enables us to study how scene complexity and referential ambiguity affect the eye-voice span. Furthermore, we are interested in the interaction of visual and conceptual factors during linguistic encoding. The visual factor we focus on is clutter (density of objects in the scene); **clutter** has been investigated in the visual processing literature and found to affect visual search (Henderson et al., 2009). The conceptual factor we investigate is the **animacy** of the referent; animacy has been manipulated in the psycholinguistic literature and found to affect sentence production (Branigan et al., 2008). Here, we address the question whether these two factors representing different modalities contribute independently to the formation of reference in sentence production, or whether they interact.

## Background

The recent visual cognition literature has emphasized the importance of contextual information for visual processing. For example, prior information about object categories facilitates visual search (Malcolm and Henderson, 2009; Schmidt and Zelinksy, 2009). This effect occurs if participants are asked to look for an object embedded in a scene or an object array (Brockmole and Henderson, 2006), or if categorical templates are provided which the visual system can use to determine where the target object is located (Vo and Henderson, 2010). It seems likely that similar contextual guidance effects (Torralba et al., 2006) also occur if the context is provided by another modality, e.g., by the linguistic material involved in a language production task.

In such task, speakers will often be faced with referential ambiguity, which they resolve by including disambiguating material in a sentence. For example, spatial prepositions can be used to locate an object in relation to the surrounding space, e.g., *the clipboard on the table* or adjectives can be used to contrast the intended referent with a competitor, e.g., *the brown clipboard*. Before any linguistic encoding can take place, however, the disambiguation has to happen at the visual level. When a target object is selected as a referent (because it will be mentioned in a sentence), the visual system has to retrieve scene and object information that can be used to refer to the object unambiguously. One can therefore hypothesize that if participants are faced with a linguistic task (e.g., scene description), then contextual guidance is afforded not only by visual information, but also driven by linguistic processing and the need to disambiguate.

## Experiment

In this experiment, we investigated how visual attention is influenced by contextual factors during sentence production. Participants had to describe a visual scene after being prompted with a cue word. This cue word was ambiguous, i.e., two objects in the scene could be referred to by the cue. We manipulated the animacy of the cue (e.g., *man* vs. *clipboard*), expecting an effect on both linguistic encoding and visual attention. Animate objects are associated with a larger number of conceptual structures in encoding (Branigan et al., 2008); we should therefore observe more sentences containing action information in this case (e.g., *the man is reading a letter*). At the same time, we expect visual attention to be localized on animate targets, an effect that has already been demonstrated in visual search (Fletcher-Watson et al., 2008).

The second experimental manipulation concerned a visual factor, viz., clutter, defined as the density of visual information (Rosenholtz et al., 2007). Again, this is a factor that has shown effects on the performance and accuracy of visual search: the more cluttered the scene is, the less efficient the identification of target object (Henderson et al., 2009). In a language production task, however, the effect of clutter can be expected to change, due to the disambiguation strategies required. Clutter could have a beneficial effect: the more visual information there is, the more disambiguating material can be retrieved; clutter could therefore facilitate language production.

Finally, this experiment makes it possible to investigate the effect of referential ambiguity on the eye-voice span. In previous work, the relationship between linguistic and visual referents was unambiguous: looks to the visual referent always preceded naming (Griffin and Bock, 2000) and this trend exponentially increases towards the mention (Qu and Chai, 2008). In our setting, we expect a more complex gaze-to-name relationship caused by a process of visual disambiguation that arises both before and after the intended referent is mentioned.

### Method

We used a factorial design that crossed the two factors *Clutter* (Minimal/Cluttered) and *Cue* (Animate/Inanimate). Participants' eye-movements were recorded while they described photo-realistic scenes after being prompted with a cue word, which ambiguously corresponded to two visual referents in the scene (see Figure 1).

We created 24 experimental items using photo-realistic scenes drawn from six indoor scenarios (e.g., Bathroom, Bedroom; four scenes per scenario). In each scene, we inserted two animate and two inanimate objects using Photoshop, which correspond to the two *Cue* conditions; *Clutter* was either added or removed.

Twenty-four native speakers of English, all students of the University of Edinburgh, were each paid five pounds for taking part in the experiment. They each saw 24 items randomized and distributed in a Latin square design that made sure that each participant only saw one condition per scene.


Figure 1: Example of an experimental trial, with visual region of interest considered for analysis. PRIMARY indicates that the ANIMATE and INANIMATE visual objects are spatially close and semantically connected (e.g., the MAN is doing an action using the CLIPBOARD). SECONDARY is used to indicate the remaining referent of the ambiguous pair. BACKGROUND and CLUTTER are defined in opposition: BACKGROUND is everything other than CLUTTER.

An EyeLink II head-mounted eye-tracker was used to monitor participants' eye-movements with a sampling rate of 500 Hz. Images were presented on a 21" multiscan monitor at a resolution of 1024 x 768 pixels; participants' speech was recorded with a lapel microphone. Only the dominant eye was tracked. A cue word appeared for 750 ms at the center of the screen, after which the scene followed and sound recording was activated. Drift correction was performed at the beginning and between each trial. There was no time limit for the trial duration and to pass to the next trail participants pressed a button on the response pad. The experimental task was explained using written instructions and took approximately 30 minutes to complete.

### Data Analysis

We defined regions of interest (ROIs) both for the visual and the linguistic data. The visual data was aggregated into six different regions: PRIMARY and SECONDARY ANIMATE, PRIMARY and SECONDARY INANIMATE, BACKGROUND, and CLUTTER (see Figure 1).

For the linguistic data, we made a general division between time windows *Before* and *During* production. This allows us to capture the overall trend of the two main phases of a trial. For the analysis of eye-voice span, we consider a window of 2000 ms before the referent was mentioned, similar to Qu and Chai 2008. The resolution of visual ambiguity is analyzed using a window of 1600 ms (divided into 40 time slices 40 ms each): 800 ms before and after the mention of *Cue*. This makes it possible to explore how the linguistic referent is visually located before being mentioned and just after.

In order to unambiguously analyze fixated and named referents, we aggregate eye-movements responses in four blocks (Primary, Secondary, Ambiguous and Both) by manually

checking which referent was mentioned in each sentence.[1] We introduced referential ambiguity as predictor in the inferential model described below to investigate how looks to the mentioned object differ from those to its competitor. For reason of space, we only present the analysis for the *Primary* objects mentioned. The effect of mention on eye-movements' pattern is evaluated by comparing Primary with Secondary objects.

As an initial exploration of our data, we investigate the overall trend of fixations *Before* and *During* production. Production is a task with large between-participant variability, e.g., one participant will spend 2000 ms *Before* and 1000 ms *During* production, whereas another one will show the opposite pattern. Normalizing the production data is therefore crucial, in particular as we want to interpret eye-movements in relation to phases of linguistic processing. We normalize each sequence $S^i_{old}$ of eye-movements by mapping it onto a normalized time-course of fixed length $S^i_{new}$. The length of $S^i_{new}$ is set on the basis of the shortest eye-movement sequence $\min_i[\text{length}(S^i_{old})]$ found between Before and During production, across all participants.[2] For each sequence $S^i_{old}$, we obtain the number of old time-points $k^i$ corresponding to a new time-unit $u$, as $k^i = \text{length}(S^i_{old})/\text{length}(S^i_{new})$. Proportions are then calculated over $k^i$ old time-points and subsequently mapped into the corresponding unit $u$ of the normalized time-course. In the Results section, we show plots of normalized proportions for *Primary* and *Secondary* (Animate and Inanimate) across conditions, *Before* and *During* production.

To explore the eye-voice span hypothesis, we compute the number of fixations to the mentioned object compared to the competitor. We also look at latencies, i.e., the onset of the last fixation to the referent or competitor before the mention, and gaze duration as a function of latencies, i.e., the time spent looking at the referent or competitor for the different latencies.

We also report inferential statistics for the referent region (for the time windows previously described). The dependent measure is the empirical logit (Barr, 2008), calculated as $emplog = \ln \frac{0.5+\phi}{0.5+(1-\phi)}$, where $\phi$ is the number of fixations on the region of interest. The analysis is performed using the framework of linear-mixed effect (LME) models as implemented by the R-package lme4 (Baayen et al., 2008). The predictors included were *Animacy*, *Clutter*, *Time* and *Object*. The random factors were *Participant* and *Item*. To reduce colinearity, factors were centered.

The model selection followed a conservative stepwise forward procedure that tests model fit based on a log-likelihood test comparing models each time a new parameter is included. If the fit improves, we accept the new model, otherwise we keep the old one. We include predictors, random intercepts and slopes ordered by their log-likelihood impact on model fit. We iterate until there is no more improvement on the fit; leaving us with the best model. In the result section, we show plots of the values predicted by the model for each condition.

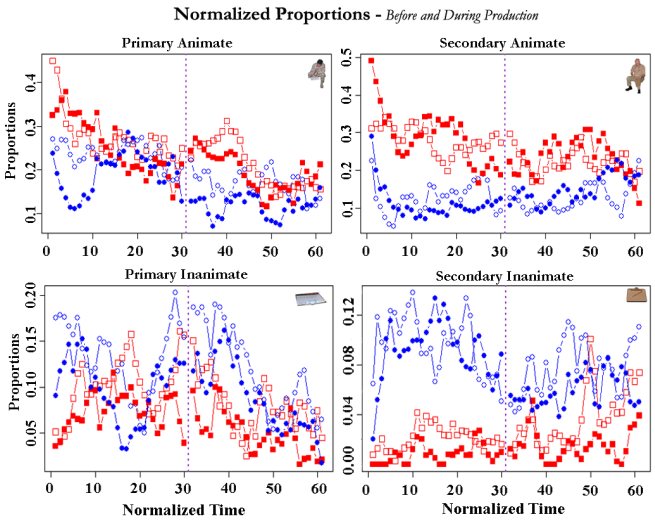



Figure 2: Normalized proportions of looks (60 bins) across the four conditions, Before and During production, for the different visual ROIs. The purple dashed vertical line indicates Before (to the left) and During (to the right) production. The four conditions are coded as following: *Animate/Cluttered*: red, full-square; *Animate/Minimal*: red, empty square); *Inanimate/Cluttered*: blue, full circle; *Inanimate/Minimal*: blue, empty circle

## Results and Discussion

**Before and During Production** We first look at how fixations are distributed when we collapse the two main phases of the experiment: *Before* and *During* production. This analysis does not distinguish whether the *Primary* or *Secondary* referent was mentioned. Figure 2 shows normalized proportions of looks on the competitor visual objects corresponding to the *Cue* (Animate/Inanimate).

The first thing to note is that for the visual ROI corresponding to the *Primary* referent, the pattern of fixations is more complex than for the ROI of the *Secondary* referent. The spatial proximity and semantic relatedness of the two *Primary* referents result in a more complex pattern of interaction. The clearest effect is found in relation with the animacy of *Cue*; we observe more fixations to the animate referent when the cue is also animate. When looking at the *Primary* ROI, the effect is seen at the beginning of both the Before and the During region. At the beginning of the trial, the visual system retrieves information about the cued objects; when production starts, the referents are fixated again, probably before being mentioned. For the *Secondary* ROIs, the relation with the *Cue* is stronger, probably reinforced by the referential competition. Moreover, the pattern of looks is much clearer than for the *Primary* ROI. This confirms that spatial proximity and

[1] PRIMARY means that the Primary Animate or Inanimate is mentioned (e.g., *The man is writing on the clipboard*). SECONDARY is used when the Secondary Animate or Inanimate is mentioned (e.g., *The man is reading a letter*). AMBIGUOUS is used when is unclear which one is referred to (e.g., *the man is sitting on the couch*). BOTH indicates that both referents are mentioned (e.g., *the man is writing on a clipboard while the other man reads a newspaper*).

[2] We remove outliers that are two standard deviation away from the mean, after having log-transformed our data. The data are not normally distributed, due to right skewness. The log-transformation helps us to reduce the skew.

Table 1: Eye-voice span statistics. *Excluding* indicates that the percentage is calculated considering only those cases in which either the referent or competitor have been fixated, *Including* takes into account also cases where both have been fixated.

| Measure | | Referent | Competitor |
|---|---|---|---|
| Percentage of looks | Including | 71.65 | 43.30 |
| | Excluding | 36.44 | 8.09 |
| Mean Latency | Including | 1032 ms | 1203 ms |
| | Excluding | 1012 ms | 1325 ms |
| Gaze Duration | Including | 489 ms | 432 ms |
| | Excluding | 568 ms | 623 ms |

semantic relatedness increase the interaction between visual referents. *Clutter* does not have a strong effect, though there is a small increase of looks when the scene is minimal and the animacy of the target matches that of the cue.

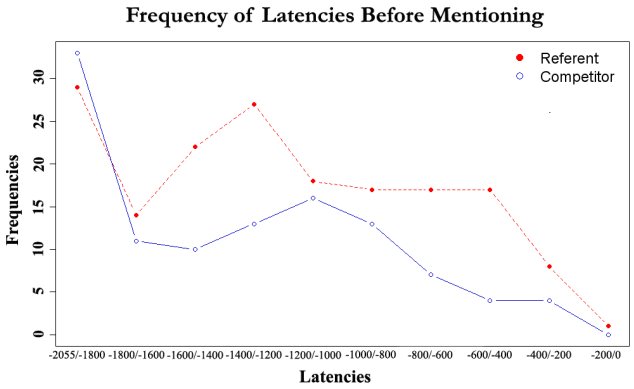
**Eye-Voice Span** We analyzed eye-voice span to investigate the gaze-to-name relation for the mentioned referent and its competitor. Table 1 shows percentages of looks to referent or competitor with mean latencies and gaze durations.[3]

There is a preference for looks to the referent over looks to the competitor, with a latency of about one second, confirming previous findings (Griffin and Bock, 2000). In a minority of cases, participants only look at the referent (36.44%); competition between the two ambiguous visual referents is the norm (71.65%). Moreover, we notice that the competitor is fixated earlier than the referent and the duration is shorter for the Including condition (which includes trials in which both referents have been fixated). This may indicate that the final decision on which referent is mentioned is made after discarding the competitor.

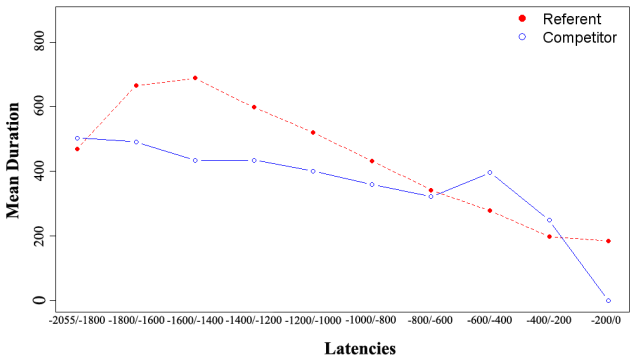
Figure 3(a) shows frequencies of *Latencies* at different temporal blocks (200 ms each) within a total window of two seconds. We find that latency frequency decreases towards the mention for both the referent and the competitor. This finding contrasts with Qu and Chai (2008) who found the opposite trend, i.e., the closer to the mention, the more gazes are associated with the referent object. Note also that this effect cannot only be due to the presence of a competitor, e.g., comparative looks before mention, as these present a similar decreasing trend.

In Figure 3(b) we show mean gaze duration as a function of the different latencies. Again, a decreasing trend is clearly visible: the closer the latency to the mention, the shorter the gaze duration. Interestingly there is a peak of gaze duration at 1600/1400 ms. The higher duration found at this latency might be an indicator of referential selection (gaze-to-name binding). We also find evidence of competition at 600/400 ms, where the competitor receives longer gazes compared to referent. A last visual check on the competitor is probably performed before referentiality is encoded linguistically.

[3]The measures are calculated only when the Primary and Secondary referent are mentioned; thus, we exclude the Both and Ambiguous cases, for which it was not possible to establish unambiguous eye-voice span relation.




(a) Frequencies of latencies at different temporal blocks (from two seconds to mention): red is the referent, blue the competitor. The latency measures the time elapsed from the beginning of the last fixation to the object (referent or competitor) until is mentioned.




(b) Mean gaze duration as a function of latency. The mean of gaze duration is calculated for the different blocks of latencies. We analyze only cases where gaze duration is shorter than latency, thus avoiding cases where fixations spill over into the region after mention.

Figure 3: Eye Voice Span statistics.

**Inferential Analysis** We now analyze the pattern of eye-movements before and after the mention of the cue word. To save space, we focus on the case where the Primary visual object is mentioned. Based on the eye-voice span analysis, we expect to find a decreasing trend of looks before the referent is mentioned, and the presence of competition should weaken the gaze-to-name relationship.

Recall that our experiment had two factors (Cue: animate/inanimate; Clutter: minimal/cluttered); we also include the object fixated (Object: primary/secondary) and Time (in 40 ms slices, see Data Analysis above) in the analysis. Figure 4 plots LME predicted values for the four conditions, Before and After mention.[4]

Beginning with the animate visual objects in Figure 4, we expect the *Primary Animate* to receive more looks than the *Secondary Animate*, and the number of looks should increase. We observe a preference for looks to Primary Animate,

[4]The intercepts for Before and During are different because they are calculated over distinct time intervals.

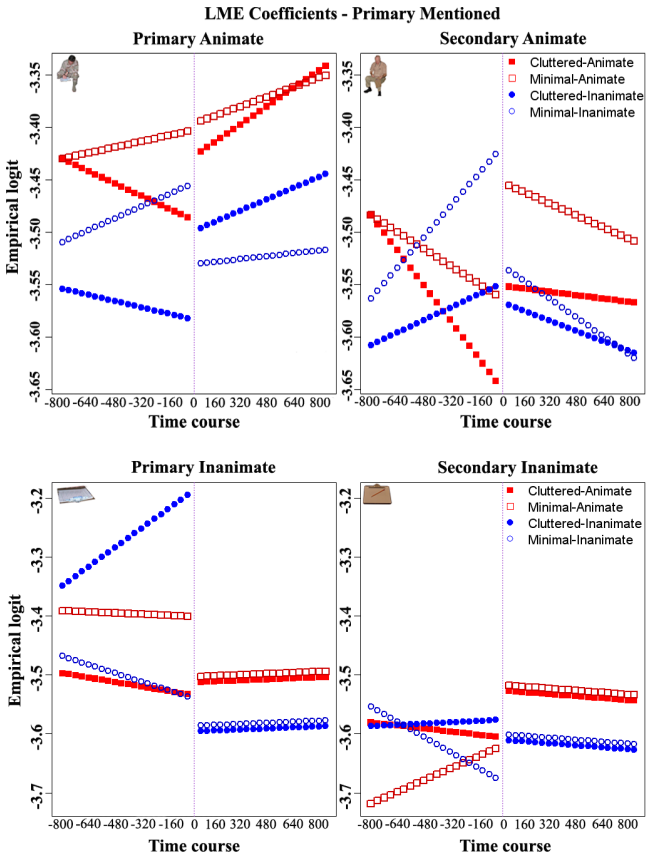



Figure 4: Linear mixed effect model: plot of predicted values (40 windows of 40 ms each) across the four conditions, Before and After referent, on the different visual ROIs. The *First* referent is mentioned and the dashed line indicates when.

but the difference is not statistically significant ($\beta_{Primary} = 0.0255; p > 0.1$). However, we find a main effect of *Cue* ($\beta_{Animate} = 0.0543; p < 0.01$): an animate cue facilitates looks to Animate visual objects. When looking at the time course, we find a general decreasing trend ($\beta_{Primary:Time} = -0.022; p < 0.01$), partly compensated by a three-way interaction of *Animacy*, *Object*, and *Time* ($\beta_{Animate:Primary:Time} = 0.049; p < 0.001$). Moreover, we observe a two-way interaction of *Clutter* and *Time* ($\beta_{Minimal:Time} = 0.024; p < 0.01$): a minimal scene makes it difficult to retrieve disambiguating information for the animate referent, forcing the visual system to look for this information on the referent itself. It is also conceivable that the minimality of the scene makes visual responses similar to those found for line drawings (Griffin and Bock, 2000); thereby explaining the increasing trend. In a cluttered environment, instead, there are more ways to relate the referent to the surrounding context, hence helping language production to disambiguate. This explains the decreasing trend of fixations on the referent in the cluttered condition.

After mention, we observe interactions of *Cue* with *Clutter* ($\beta_{Animate:Minimal} = 0.0165; p < 0.001$) and *Object* ($\beta_{Animate:Primary} = 0.017; p < 0.01$), confirming both the facilitation of the cued referent and the preference for referent information when scenes are minimal. In contrasts with previous findings, we observe increasing looks to the referent after mention ($\beta_{Primary:Time} = 0.0530; p < 0.001$). This effect could be due to referential ambiguity: the visual system is connecting disambiguating material retrieved before mention to the referent just uttered. For the *Secondary* Animate, we find an increasing trend of looks when *Cue* is Inanimate and especially for minimal scenes ($\beta_{Inanimate:Time} = 0.056, p < 0.001; \beta_{Minimal:Time} = 0.041, p < 0.01$). The minimality of the scene gives prominence to animate referents; probably the spatial and semantic proximity of one of Primary Inanimate and the Primary Animate also trigger comparative looks to Secondary Animate, i.e., participants check whether it can also be contextually related to the cue.

After the referent is mentioned (*Primary* in this case), looks to the *Secondary* Animate decrease over time in all conditions. Competition is triggered by visual ambiguity, but once the association of the visual with the linguistic referent has been established (i.e., after the mention), participants look back to the referent mentioned, presumably finalizing the choice made.

Looking at inanimate referents in Figure 4, we observe a statistically significant preference for looks to the Primary Inanimate ($\beta_{Primary} = 0.0621; p < 0.05$). This preference could be due to the spatial proximity and the semantic relation with the primary animate, which makes the primary inanimate more likely to be encoded either as a direct object or as subject of the description. As a consequence, we find an interaction with the animacy of the *Cue* ($\beta_{Animate:Primary} = 0.0155; p < 0.05$) but not a main effect ($\beta_{Inanimate} = 0.017; p > 0.1$). In contrast with standard visual search task, where performance degrades as a function of clutter, here we observe instead a positive interaction of *Clutter* and *Cue* on the target ($\beta_{Inanimate:Cluttered:Primary} = 0.028, p < 0.001$), which increase over time ($\beta_{Cluttered:Time} = 0.054, p < 0.01$). The visual system is not performing a search task, rather it is sourcing information to ground language processing. In a cluttered scene, an inanimate referent could be spatially related to many other different objects, whereas a minimal scene has fewer points to anchor the referent. The visual system therefore needs to select among the different spatial relations to find one that optimally situates the object within the contextual information.

For the secondary inanimate, there is a negative relationship between the animacy of *Cue* and the minimality of *Clutter* ($\beta_{Animate:Minimal:Secondary} = -0.0719; p < 0.001$); the proximity and relatedness of the primary inanimate and the primary animate is highlighted when visual information is minimal, which results in the secondary inanimate being fixated less.

## General Discussion

Referential ambiguity is a common phenomenon in everyday experience. In a naturalistic scene, the same object (e.g., a clipboard) can occur multiple times (e.g., on a desk or on a counter). This fact turns into linguistic ambiguity when a referent has to be selected from the set of visual competi-

tors. Typically, referential ambiguity is resolved by encoding sufficient contextual information to discriminate the intended referent from competitors (e.g., *the clipboard on the desk*). However, this process of ambiguity resolution cannot be explained by linguistic factors alone, especially given that the disambiguating material needs to be selected by the visual system prior to any encoding. We therefore hypothesized that visual factors interact with well-established conceptual factors active during language production.

We reported the results of an eye-tracking language scene description experiment that support this hypothesis. We explored how the conceptual properties of the target referent (factor *Cue*: animate/inanimate) and the density of visual information (factor *Clutter*: minimal/cluttered) interact during the resolution of referential ambiguity. The results showed that the animacy of the cue facilitates looks to animate objects, especially at the beginning of two main phases of linguistic production: before and during the mention of the referent. The data indicate that a visual search is performed to localize the objects matching the cue word (Malcolm and Henderson, 2009). Our results also contrasted interestingly with findings for visual search, where clutter decreases search performance (Henderson et al., 2009). In cases in which an animate referent is mentioned, we found that there were fewer fixations to the target object in the cluttered condition compared to the uncluttered one. In other words, clutter makes language production easier, not harder: the visual system is not just searching for the target object, but it is also retrieving visual information that can be used to linguistically anchor it (e.g., for disambiguation). The more clutter there is, the easier this process becomes, explaining the reduced number of fixations in the cluttered condition.

Turning at the relation between fixating and naming an object (the eye-voice span), previous work found that referents are fixed shortly before being mentioned (Griffin and Bock, 2000). It has also been observed that fixation probability increases with decreasing distance to the mention (Qu and Chai, 2008). In our data, we found a numerical preference for looks to the mentioned referent over looks to the competitor, but this preference was not confirmed in the inferential analysis (see Figure 4). Only if the primary inanimate was mentioned, it was fixated significantly more than the secondary inanimate. This preference is likely due to the proximity, spatial and semantic, between the primary animate and inanimate. Moreover, we found that fixation probability decreased with decreasing distance to the mention, contrary to previous results, in particular when the scene was cluttered. The competition between visual referents seems to override the standard eye-voice span effect. Interestingly, we also observed an increasing trend of fixation to the referent object *after* its mention. Once production has started, the visual system needs to retrieve contextual information to produce disambiguating linguistic material, resulting in an increase in the number of looks after mention.

Taken together, our results indicate that visual factors such as clutter interact with conceptual factors such as animacy in language production. The simple view according to which referents are fixated in the order in which they are mentioned, with a fixed eye-voice span between fixation and mention, does not seem to generalize to more realistic settings in which speakers describe naturalistic scenes that involve referential ambiguity.

## References


Altmann, G. and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Baayen, R., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59:390–412.

Barr, D. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of memory and language*, 59(4):457–474.

Branigan, H., Pickering, M., and Tanaka, M. (2008). Contribution of animacy to grammatical function assignment and word order during production. *Lingua*, 2(118):172–189.

Brockmole, J. R. and Henderson, J. M. (2006). Using real-world scenes as contextual cues for search. *Visual Cognition*, 13:99–108.

Fletcher-Watson, S., Findlay, J., Leekam, S., and Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception*, 37(4):571–583.

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

Henderson, J. M., Chanceaux, M., and Smith, T. J. (2009). The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of Vision*, 9(1)(32):1–8.

Malcolm, G. and Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: Evidence from eye movements. *Journal of Vision*, 9(11)(8):1–13.

Qu, S. and Chai, J. (2008). Incorporating temporal and semantic information with eye gaze for automatic word acquisition in multimodal conversational systems. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP). Honolulu*.

Rosenholtz, R., Li, Y., and Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, 7:1–22.

Schmidt, J. and Zelinksy, G. (2009). Search guidance is proportional to the categorical specificity of a target cue. *Quarterly Journal of Experimental Psychology*, 62(10):1904–1914.

Tanenhaus, M., Spivey-Knowlton, J., Eberhard, K., and Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, (268):632–634.

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

Vo, M. and Henderson, J. (2010). The time course of initial scene processing for eye movement guidance in natural scene search. *Journal of Vision*, 10(3):1–13.