The Impact of Visual Information on Reference Assignment in Sentence Production

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

Reference is the cognitive mechanism that binds real-world entities to their conceptual counterparts. Recent psycholinguistic studies using eye-tracking have shed light on the mechanisms used to establish shared referentiality across linguistic and visual modalities. It is unclear, however, whether vision plays an active role during linguistic processing. Here, we present a language production experiment that investigates how cued sentence encoding is influenced by visual properties in naturalistic scenes, such as the amount of clutter and the number of potential actors, as well as the animacy of the cue. The results show that clutter and number of actors correlate with longer response latencies in production, and with the generation of more complex structures. Cue animacy interacts with both clutter and number of actors, demonstrating a close coupling of linguistic and visual processing in reference assignment.

Keywords: language production, sentence encoding, picture description, visual information, clutter, accessibility, animacy.

Introduction

When humans comprehend or produce language, they rarely do so in isolation. Linguistic information often occurs synchronously with visual information, e.g., in everyday activities such as attending a lecture or following directions on a map. The visual context constrains the interpretation of the linguistic material, and vice versa, making processing more efficient and less ambiguous. Linguistic and visual processing have been investigated extensively in isolation, but there is little work that explicitly relates the two modalities to each other.

In this paper, we focus on a particular aspect of synchronous linguistic and visual processing, viz., the formation and maintenance of shared reference between the two modalities. Essentially, this is the problem of determining that a visually perceived entity such as CUP is referentially linked to a sequence of words such as *the cup*. On the linguistic level, the complexity of this task increases if referring expressions are embedded into larger linguistic structures (e.g., *the cup on the table*, *the spoon in the cup*). On the visual level, complexity increases if the CUP is embedded into a scene, for example an interior such as a kitchen, which can contain a large number of objects.

Insights into the mechanisms underlying the interaction between visual and linguistic processing can be obtained using the Visual World Paradigm (VWP, Tanenhaus et al. 1995). In VWP studies, participants are engaged in a synchronous visual and linguistic task while their eye-movements are recorded. For example, in a language production study conducted by Griffin & Bock (2000), participants were eye-tracked while they described pictures containing two actors (depicted alternatively as Agent or Patient of the event). The goal of the study was to investigate whether there is a sequential relation between the visual entities fixated their linguistic naming. A key observation is that the fixation of a visual entity and the production of a linguistic referent are closely timelocked; the latency between the two modalities is referred to as the *eye-voice span.*

A similar effect can be observed in VWP studies of language comprehension (e.g., Altmann & Kamide 1999). Here, participants listen to a speech stimulus while viewing a scene, and typically they fixate a visual entity shortly after a corresponding linguistic referent has been encountered in the speech. The VWP has mainly been used in psycholinguistic research, with a focus on how specific aspects of the linguistic stimuli cause certain visual objects to be selected as referents (Knoeferle et al., 2006; Snedeker & Trueswell, 2003). In these studies, visual information is merely used to provide a context for language processing, without taking into account mechanism of scene comprehension studied extensively in the visual cognition literature (Henderson et al., 2007). This is compounded by the fact that the visual stimuli used in VWP studies typically include clip-art objects arranged in arrays or pseudo-scenes. The resulting visual processing is of reduced complexity, perhaps consisting merely of responses to the linguistic input. Moreover, this visual simplicity often results in a one-to-one mapping between visual and linguistic referents. This is unrealistic compared to naturalistic scenes where the same linguistic label can often correspond to multiple visual objects.

A realistic theory of the formation and maintenance of reference across modalities has to treat visual information on a par with linguistic information. Such a theory must explain how mechanisms known to operate independently in both modalities cooperate in referent assignment. The present paper aims to contribute to such a theory. On an abstract level, the hypothesis we test is that the visual stimulus properties exert an influence on linguistic processing. Previous work has investigated the influence of low-level visual properties such as saliency (a composite of color, intensity, and orientation, Itti & Koch 2000). In a VWP study, Coco & Keller (2008) found that saliency influences the resolution of prepositional phrase (PP) attachment ambiguities in language comprehension. They found that saliency has referential effects; it is used to predict which visual objects can be encoded as post-verbal arguments in a given sentence.

However, it is important to note that low-level visual fea-

tures such as saliency are not referential per se; they are properties of image regions, not of objects (Henderson et al., 2007). It is therefore necessary to focus on higher-level visual properties, which are clearly object-driven and likely to affect the mechanisms of referent assignment in sentence encoding directly. In this paper, we present a language production experiment that investigates how scene descriptions are influenced by high-level features such as visual clutter (the density of objects in a scene) and of the number of animate referents available. We use naturalistic scenes as visual stimuli to avoid the limitations of visual arrays and clip-art images, traditionally used in the VWP literature.

Experiment

Design

In this experiment, participants had to describe a naturalistic scene, after being prompted by a single word (the description cue). As dependent variables we recorded Looking Time, i.e., the time that elapsed before the onset of the response, Description Time, i.e., the time taken to complete the response, and we also investigated the syntactic structure of the response produced. The design of the experiment manipulated both visual and linguistic referential information. We varied the total amount of visual information present in the scene in the factor Clutter (Minimal vs. Cluttered). We also manipulated the number of animate objects present in the scene in the factor Actors (One vs. Two). On the linguistic side, we varied the prompt given to participants for their description in the factor Cue, which could refer either to an animate or an inanimate object in the scene (Animate vs. Inanimate). The scenes were designed such that they always contained at least one animate object and two identical inanimate objects, so as to introduce systematic visual referential ambiguity. As an example, see Figure 1, where the clipboard is the ambiguous inanimate object. Note that the animate objects are referentially unambiguous, even in the two actor condition (man and woman in the example stimulus).

The null hypothesis for this experiment is that visual and linguistic factors do not interact in language processing. This would mean that Clutter and Actors should only influence Looking Time in a way that is compatible with behavior in standard visual search tasks: we expect longer Looking Time in the Cluttered condition, as more objects have to be searched, and longer Looking Time also in the Two Actors condition, which contains an additional object. Our experimental hypothesis is that visual information has an impact on language production, which means that we expect an interaction between the visual factors Clutter and Actor and the linguistic factor Cue (in addition to the main effects of the visual factors that may be caused by standard visual search processes).

In the following we will give a more detailed motivation for the factors included in the design and describe how they were operationalized.

Clutter A way to define reference in vision is to look for a global measure of visual information. Measures of visual information can be defined in various ways; a common approach uses the notion of *set size*. The bigger the set of objects in a visual search task, the slower the response time. Hence, the number of countable visual objects has been assumed to give a direct measure of visual information (Wolfe, 1998). However, this notion of visual information has recently been criticized (Rosenholtz et al., 2007), especially in the context of naturalistic scenes, where it can be come difficult, if not impossible, to define and count each single object in the scene.

An alternative way of quantifying visual information is *clutter* (Rosenholtz et al., 2007) (see Figure 1 for an example). Clutter is defined as the state (organization, representation) of visual information in which visual search performance starts to degrade. Clutter can be modeled statistically and quantified using the feature congestion method (for details see Rosenholtz et al. 2005). In our study, we use this Clutter measure to investigate the effect of the amount of visual information on sentence encoding.

Actors and Animacy A crucial feature that distinguishes different types of real-world entities is animacy. Animacy is known to play a role in language production; in particular, it can influence the assignment of grammatical functions and word order (Branigan et al., 2008). Animate entities are conceptually more accessible than inanimate ones (Levelt et al., 1999) and therefore privileged during syntactic encoding. This is reflected by the fact that animate entities are more likely to be encoded with the grammatical function subject, while inanimate entities occur mostly with the function object.

In this study we took a broader view of the feature animacy; animacy is not only a linguistic notion, but it is also visually encoded. We therefore manipulated animacy in both the linguistic and the visual modality. Visually, we introduce different degrees of animacy by changing the number of actors depicted in the scene. Linguistically, we either gave an animate or an inanimate noun as the cue for sentence production.

Method

The experimental design crossed three factors, each with two levels. The two visual factors were number of Actors in the scene (One or Two) and the degree of visual Clutter (Minimal or Cluttered). The linguistic factor was the Cue given to the participants to prompt their sentence production (Animate or Inanimate).

As stimuli, we created a set of 24 photo-realistic scenes using Photoshop by cutting and pasting visual objects from a set of preexisting photographs. Differences in luminosity, contrast and color balance between the different photographs were adjusted through an accurate use of layers, luminosity masks and color balancing. In order to (1) control for the se-

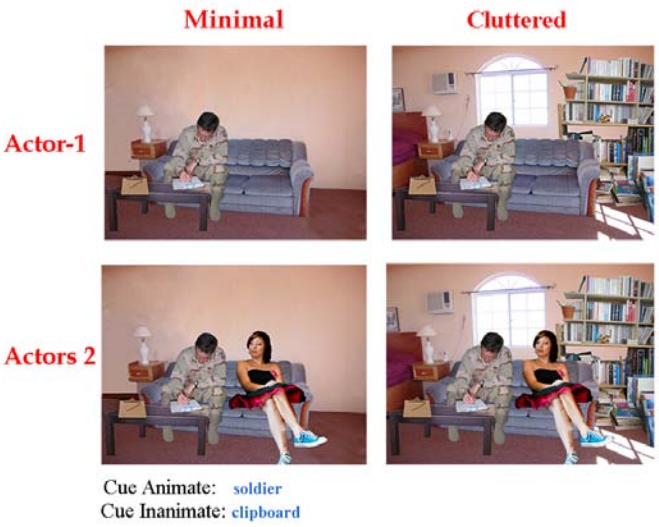

Figure 1: An example of an experimental stimulus, with the four visual variants which occur in the experiment and the two linguistic cues presented.

mantic variability across visual scenes and (2) ground language production in a restricted semantic domain, all pictures were created using six different interior environments: bathroom, bedroom, dining room, entrance, kitchen, and office. Each interior was represented by four different scenes. For each scene, we created four variants by manipulating Clutter and Actors, as illustrated in Figure 1. The scenes were designed such that the inanimate object was referentially ambiguous, i.e., each picture contained two visual instances of it, while the animate one was unambiguous, even in the Two Actors condition.

In the experiment, participants were first presented with a set of instructions explaining the task and giving examples. After a practice phase, they saw one visual stimulus at a time, together with the linguistic cue. They were instructed to provide a written description of the stimulus using the cue. The total of 192 different items were distributed over four lists using a Latin square design. Each subject saw one of the lists, i.e., 48 stimuli in total (each of the 24 scenes was presented twice, once with animate and one with inanimate cue). The stimuli were randomized for each participant, and presented without fillers. The experiment took about 15 minutes in total.

The experiment was realized using the WebExp software package for conducting psychological experiments over the web. WebExp is able to measure reaction times with accuracy comparable to that of lab-based experiments, as shown by Keller et al. (2009) for self-paced reading data.

Participation was open to both native and non-native speakers of English (this was included as a factor in the analysis). The sample included 32 participants, including 16 native speakers and 16 non-native speakers.

Results and Discussion

We analyze two response time measures. The first one is Looking Time, i.e., the time participants spent scanning the image before starting to type. It is calculated from the onset of the trial until participants pressed the first key on the keyboard. The second response time measure, Description Time, is the time participants took to type their response. It is calculated from the first key press until Enter was hit to move on to the next trial.

We also analyzed the syntactic patterns in the responses produced by participants. For this, we tagged each sentence produced using a an automatic part-of-speech tagger, viz., Ratnaparkhi's (1996) maximum entropy tagger, which performs with an accuracy of 96.6%. The tagger uses the Penn Treebank tagset to assign syntactic categories to words. We collapsed the various tags for nouns in the tagset (e.g., NNS, NNP) and verbs (e.g., VBD, VBN) to two general categories (NN, VB). For each sentence, we recorded the frequency of these two categories, as well as the occurrence of existential *there* and clause coordinator *and*. We also identified and counted the number of passive constructions (for this the full tag set was used, which marks passive verb morphology).

The statistical analyses were carried out using linear mixed-effect models (Jaeger, 2008) to determine the effect the categorical predictor variables on both reaction times and syntactic frequency. We chose mixed models for their ability to capture both fixed and random effects (Baayen et al., 2008). We included the following predictors in our analysis: Actors (One or Two), Clutter (Minimal, Cluttered), Cue (Animate, Inanimate) and Language (Native, NonNative). The baseline on which the Intercept was calculated was given by the condition Cue-Inanimate, Actor-One, Clutter-Cluttered, Language-Native. The mixed models were built and evaluated following the model selection procedure suggested by Crawley (2007). We started with a fully specified model containing all the predictors and all possible interactions and then we reduced the model iteratively by removing the highest order interaction with the highest p-value. The estimates were recomputed at each iteration.

Reaction Times Table 1 presents the coefficients and p-values of the mixed model for Looking Time (only significant predictors and interactions are included). The model intercept represents the response time in the baseline condition in milliseconds, and coefficients indicate the effect a given predictor has on Looking Time (again in milliseconds). We find that participants were significantly faster to scan the pictures in the condition Clutter-Minimal compared to Clutter-Cluttered. This finding is likely to be an effect of visual search, as the scene needs to be searched for the cued object; in the Cluttered condition more objects are present, leading to longer search time.

Significantly shorter Looking Time was also observed in the Cue-Animate condition. Again, this can be explained in terms of visual search behavior, as our stimuli contain more

Predictor	Looking Time	
	Coefficient	p
Intercept	3774.6	0.0006
Clutter-Min	-503	0.01
Cue-Anim	-1097.4	0.01
Language-NonNative	1120.3	0.04
Actors-Two:Cue-Anim	-468.4	0.05
Clutter-Min:Cue-Anim	596.2	0.01
Predictor	**Description Time**	
	Coefficient	p
Intercept	12053	0.0024
Actors-Two	987.2	0.04
Language-NonNative	2139.3	0.01
Cue-Anim	-1801.1	0.0001

Table 1: Mixed effects models of Looking Time and Description Time

Predictor	Noun	
	Coefficient	p
Intercept	2.2520	0.0002
Cue-Anim	-0.3333	0.0001
Actors-Two:Cue-Anim	0.2252	0.01
Predictor	**Verb**	
	Coefficient	p
Intercept	1.6129	0.0006
Clutter-Min	0.1968	0.004
Cue-Anim	0.2307	0.0001
Actors-Two:Cue-Anim	0.1624	0.03
Clutter-Min:Cue-Anim	-0.2267	0.002

Table 2: Mixed effects models of noun frequency and verb frequency

inanimate than animate cues, thus making it easier to discriminate animate objects, leading to reduced search time. In addition, the cue animate object was always unambiguous, while the cued inanimate object was always present twice in the scene, creating referential ambiguity, and thus increasing visual search time. There may also be an explanation in linguistic terms: As mentioned above, animate entities are conceptually more accessible than inanimate ones, which gives them a privileged status during syntactic encoding.

We also found that in the condition Language-NonNative, participants take longer to scan the picture. This can be explained by the fact that non-native speaker presumably take longer to decode the cue and to plan their utterance.

Turning to the interactions, we found that Clutter-Minimal significantly interacts with Cue-Animate: participants took longer to respond to animate prompts in the minimal clutter condition. This interaction cannot be explained purely in terms of visual search. The Clutter-Minimal, Cue-Animate condition is the one with the fewest competing objects (minimal clutter) and only one or two animate objects to consider; visual search should therefore be particularly fast, and the interaction should be absent or have a negative coefficient. The fact that we find a positive interaction indicates that a linguistic process is at work. In a visual scene with few objects it is more difficult to retrieve enough information regarding actions that a potential actor can perform. Thus, participants spend more time scanning the scene and planning their utterance before sentence encoding starts.

There is also a significant negative interaction of Actors-Two and Cue-Animate; Looking Time is reduced in this condition. Again, this cannot be explained purely in visual terms; the presence of two actors cued by the animate cue should lead to longer search times, as two objects need to be considered instead of one. Instead, we find a negative coefficient for this interaction. Presumably, the more animate entities the scene contains, the more conceptual structures are activated. The time spent on planning a conceptual structure to encode is thus shortened by the larger set of possibilities. Moreover, the unambiguous visual reference of Cue-Animate may boost the selection of those conceptual structures that are related to the actor cued, contributing on the decrease of looking time. This interpretation is supported also by our syntactic analysis (see next section) in which Actor and Cue-Animate positively correlate with the use of nouns and verbs. Participants produce longer sentence structures, often encoding both Actors.

Table 1 also presents the mixed model for Description Time (again only significant predictors and interactions are included). The results overlap with those for Looking Time. For condition Cue-Animate, participants were faster to generate a sentence compared to Cue-Inanimate. As for Looking Time, this result can be explained by the fact that animate entities are more accessible in language production, and that visual search is faster, as there is only at most one other animate object in the scene. We also find significantly increased Description Time for the Actor-Two condition. An inspection of the responses (see below) shows that participants tend to encode both actors in their descriptions of the scene, which explains why encoding takes longer in this condition, compared to the Actor-One condition, in which only one actor is encoded. Again, non-native participants show a longer response time than native ones, presumably because sentence production is slower in non-native speakers.

Syntactic Categories Table 2 presents the results for the syntactic analysis of the picture descriptions generated by the participants. We fitted separate mixed models to predict the number of nouns and the number of verbs included in the responses. Again, only significant predictors and interactions are listed in the table; the intercept represents the noun or verb frequency in the baseline condition, and the coefficients indicate how this frequency increases or decreases under the influence of the relevant predictor.

The results indicate that significantly fewer nouns are produced in Cue-Animate condition. This condition was visually unambiguous, and thus required less elaborate descriptions compared to the Cue-Inanimate condition, for which partici-

pants generated longer sentences in order to unambiguously pick out one of the two visual referents available in this condition. Moreover, the competition between the two visual objects for the inanimate cue was often resolved by encoding both visual referents within the same sentence structure. An example of a sentence produced in this condition is *The mug is beside the man, another is on top of the files, both mugs have pencils in them*. Except of the referring expression itself, all nouns are used in combination with spatial prepositions to unambiguously differentiate each visual referent.

However, when Cue-Animate was in interaction with Actor-Two, participants produced significantly more nouns. This correlates with the shorter Looking Times found for the same interaction. Participants often referentially encoded both visual actors within the same sentence structure. An example is *A man stands behind a counter in a hotel while a customer writes on a piece of paper*. Even though the cue given (here, *the man*) refers only to one actor and was visually unambiguous, the participant encoded also the second actor.

Turning now to the analysis of the number of verbs produced, we again find a significant effect of Cue-Animate, but with a positive coefficient, which means that participants generated more verbs than in the Cue-Inanimate condition. This underlines the connection between the feature Animacy and the semantics of verbs. As verbs encode actions, they are less likely to occur in descriptions of inanimate entities. The latter tend to include verbs describing static, mostly spatial, relations like *lie* or *place*, whereas animate entities can be related to a broader range of events, both static and dynamic, resulting in more verbs being generated.

An interaction between Actor-Two and Cue-Animate is also present, which is consistent with the main effect of Cue-Animate. The more animate entities are presented in the visual scene, the more verbs are used to relate them with the event that is being encoded. An example description is *A woman drinks from a cup while a man prepares a chicken to be cooked*.

The factor Clutter-Minimal also has a significantly positive coefficient, which means that more verbs are generated if the scene is uncluttered. However, there is also a significant negative interaction of Clutter-Minimal with Cue-Animate. The minimal amount of visual information available in the Clutter-Minimal scenes makes it more difficult to select and encode the actions performed by the actor, resulting in fewer verbs being generated. This result is in line with the longer Looking Time for the same interaction. We can assume that the greater number of verbs found in Clutter-Minimal can be attributed to Cue-Inanimate, in which the ambiguous visual reference leads to more elaborate descriptions. An example description that illustrates this interpretation is *An open book is sitting on the counter and there is another one sitting on the table.*

Syntactic Constructions We also selectively analyzed a number of syntactic constructions contained in the responses generated by the participants. Such construction provide information about the sentence structures employed to describe the pictures. We counted how often participants employed the existential *there* construction. The results show that this construction occurred less frequently in the Cue-Animate condition (coefficient -0.2153, $p < 0.0001$). This indicates that participants were less likely to give static spatial descriptions of animate visual referents, compared to inanimate ones.

We also find that *and* is used less frequently in the Cue-Animate condition (coefficient -0.0868, $p < 0.0001$). This result can be attributed to the ambiguous visual reference of Cue-Inanimate. The use of *and* marks a strategy of ambiguity resolution when both visual referents for Cue-Inanimate are linguistically encoded. The connection between referents is established by combining clauses through coordination.

When we analyzed the number of passive constructions, we again found a significant negative effect of Cue-Animate (coefficient -0.0436, $p < 0.0001$). This is in line with standard findings in the sentence production literature: animate entities are likely to be realized as subjects of active constructions, while inanimate tend to be realized subjects of passive constructions (assuming that the cued entity is typically realized as a subject). An example of a production that contains the use of both coordination and passive is *A teddy is being hugged by the girl sitting on the bed and another teddy is sitting on the floor at the corner of the bed.*

General Discussion

The overall goal of this study was to investigate how visual factors influence sentence encoding. The analysis focused on shared reference between the two modalities, and the mechanisms through which reference is established in language production. We assumed an interactive account of visual and linguistic processing, where informational changes in one modality are reflected in the processing of the other one. In our experimental design, we manipulated different aspects of visual reference such as visual clutter and the number of potential actors, and the animacy of the cue used for sentence production. Moreover, we systematically introduced visual referential ambiguity for the inanimate cue in order to investigate the strategies of ambiguity resolution adopted.

The analysis of Looking Time shows significant effects of the visual factors such as Clutter and Actors: the more clutter or actors, the more time the participants spend before starting to type the sentence. The Animacy of the cue was also significant: an inanimate cue resulted in longer Looking Time, mainly because of visual referential ambiguity. However, more interesting were the interactions between the visual factors and Animacy. If we assumed independence between visual and linguistic processing, we would expect response latencies typical of standard visual search tasks, based on the referential properties of the cue and influenced only by the visual properties of the stimulus. Instead, we found a

clear interaction of visual information and Animacy. A visual scene with minimal clutter means that the set of actions that can be used to relate animate actors is impoverished. Thus, longer visual search is required to integrate the animate cues with information of the scene, the opposite of what is predicted under an explanation in terms of visual search alone. On the other hand, two actors in a scene mean a larger set of conceptual structures is available to relate to the animate cue. Also, an animate actor is easier to relate to another animate actor in the same sentence with an action description, compare to an animate and an inanimate entity. This interpretation meshes with the results we obtained for the syntactic analysis of the responses produced by participants. For the Actors-Two and Cue-Animate conditions, we found that longer sentences were produced (containing more nouns and verbs), often encoding both actors. Such results can only be explained in an account in which linguistic and visual processing interact closely.

We also analyzed Description Time and the syntactic structure of the responses and found that these are mainly influenced by the animacy of the cue and the presence of visual referential ambiguity. When the cue was inanimate, participants spent more time resolving the visual ambiguity. The sentences produced in this condition contained more nouns, which were used to spatially disambiguate between the two competing visual objects. Moreover, disambiguation often occurred together with the use of conjunction *and*. In line with previous research on language production, the use of passives and existential *there* was correlated with cue animacy. An inanimate cue is more likely to be a subject of a passive and correlated with static spatial descriptions.

Our results are limited by the fact that we only had two response time measures available, Looking Time and Description Time. A more fine-grained way of investigating the time course of sentence production is desirable, e.g., using eye-tracking. In particular we have to be careful not to presuppose that Looking Time corresponds to the time spend doing visual search and planning the resulting utterance, while Description Time corresponds to the time spend generating the utterance. In reality, these two processes are likely to happen in an interleaved manner, rather than in sequence.

In future work we will further investigate integrated reference from a lexical point of view. In particular, we are interested in the relationship between contextual scene information (bedroom, kitchen, etc.) and the lexical items generated. This could be investigated using additional statistical techniques, e.g., cluster analysis. Furthermore, we are planning to conduct an eye-tracking experiment to measure the impact of visual factors on sentence encoding more directly. Such a study is likely to shed light on (1) the changes in eye-movement patters triggered by visual factors such as clutter, (2) the effects of cueing a linguistic referent which may, e.g., result in more localized visual search behavior, and (3) whether visual referential ambiguity and its linguistic resolution are correlated.

References

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

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

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

Coco, M. I., & Keller, F. (2008). Competition between visual and linguistic resources. *Presented as poster at AMLAP,14, Cambridge, UK*.

Crawley, M. (2007). *The R book*. John Wiley and Sons, Ltd.

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

Henderson, J., Brockmole, J., Castelhano, M., & Mack, M. (2007). Visual saliency does not account for eye-movements during visual search in real-world scenes. *Eye movement research: insights into mind and brain*.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489–1506.

Jaeger, T. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of memory and language*, *59*, 433–446.

Keller, F., Gunasekharan, S., Mayo, N., & Corley, M. (2009). Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods*, *41*, 1–12.

Knoeferle, P., Crocker, M., Scheepers, C., & Pickering, M. (2006). The influence of the immediate visual context on incremental thematic role-assignment:evidence from eye-movements in depicted events. *Cognition*, (pp. 481–529).

Levelt, W., Roelofs, A., & Meyer, A. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences*, (pp. 1–75).

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *In Proceedings of the conference on empirical methods in natural language processing*, (pp. 133–142).

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

Rosenholtz, R., Mansfield, J., & Jin, Z. (2005). Feature congestion, a measure of display clutter. *SIGCHI*, (pp. 761–770).

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, (pp. 103–130).

Tanenhaus, M. S.-K., M.J., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, (pp. 632–634).

Wolfe, J. (1998). Visual search. *Attention*, (pp. 13–73).