The Interaction of Visual and Linguistic Saliency during Syntactic Ambiguity Resolution

Moreno I. Coco and Frank Keller
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
Phone: +44 131 650 4407, Fax: +44 131 650 4587
mcoco@staffmail.ed.ac.uk, keller@inf.ed.ac.uk

Abstract

Psycholinguistic research using the visual world paradigm has shown that the processing of sentences is constrained by the visual context in which they occur. Recently, there has been growing interest on the interactions observed when both language and vision provide relevant information during sentence processing. In three visual world experiments on syntactic ambiguity resolution, we investigate how visual and linguistic information influence the interpretation of ambiguous sentences. We hypothesize that (1) visual and linguistic information both constrain which interpretation is pursued by the sentence processor, and (2) the two types of information act upon the interpretation of the sentence at different points during processing. In Experiment 1, we show that visual saliency is utilized to anticipate the upcoming arguments of a verb. In Experiment 2, we operationalize linguistic saliency using intonational breaks and demonstrate that these give prominence to linguistic referents. These results confirm prediction (1). In Experiment 3, we manipulate visual and linguistic saliency together and find that both types of information are used, but at different points in the sentence, to incrementally update its current interpretation. This finding is consistent with prediction (2). Overall, our results suggest an adaptive processing architecture in which different types of information are used when they become available, optimizing different aspects of situated language processing.

Keywords: Situated language comprehension; visual saliency; intonational breaks; adaptive constraint-based architecture; ambiguity resolution.

Introduction

When linguistic input occurs in a visual context, the cognitive system can draw on information from both modalities to perform sentence understanding. This means that the linguistic and visual processing streams need to be coordinated to perform standard sentence comprehension tasks such as syntactic ambiguity resolution.

Nearly 20 years of psycholinguistic research have provided insights into cross-modal processing by investigating language comprehension in a visual context, mostly utilizing the visual world paradigm (VWP; Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). This work has demonstrated that the allocation of visual attention is constrained by a wide range of linguistic factors: from perceptual properties of speech, such as phonetics or prosody (Allopenna, Magnuson, & Tanenhaus, 1998; Snedeker & Trueswell, 2003; Snedeker & Yuan, 2008) to structural and conceptual properties, such as syntactic representations, thematic roles, and event semantics (Arai, van Gompel, & Scheepers, 2007; Knoeferle & Crocker, 2006; Altmann & Kamide, 1999). Visual attention also indicates that the sentence processor can anticipate upcoming linguistic material, i.e., it is able to predict parts of the input that have not been encountered yet during incremental interpretation (e.g., Kamide, Altmann, & Haywood, 2003; Altmann & Mirkovic, 2009; Crocker, Knoeferle, & Mayberry, 2010; Kukona, Fang, Aichera, Chen, & Magnuson, 2011; Arai & Keller, 2013). A direct consequence of this finding is that the visual properties of the context can be expected to mediate the way linguistic information is processed.

A range of VWP studies have confirmed this claim: objects sharing visual features, such as shape (e.g., ROPE and SNAKE), or color (e.g., PEAS and SPINACH), and even abstract features (e.g., PIANO and TRUMPET) compete for visual attention when the linguistic item is presented (Dahan & Tanenhaus, 2005; Huettig & Altmann, 2005, 2007, 2011).[1] Further evidence for the impact of visual information on sentence processing comes from research on language production, which has shown that visual features mediate the lexical and syntactic choices during sentence production (e.g., Griffin & Bock, 2000; Gleitman, January, Nappa, & Trueswell, 2007; Coco & Keller, 2009). Gleitman et al. (2007), for example, show that if a low-level visual cue (a light flash) occurs prior to scene onset, then the encoding of the referent depicted at that location is preferred.

Previous research has therefore demonstrated that spoken linguistic processing and (non-linguistic) visual processing interact and can mutually constrain each other. However, the exact nature of this interaction has not yet been fully uncovered. To address this question, we present three visual world experiments on syntactic ambiguity resolution, which investigate how visual and linguistic information influence which interpretation the sentence processor pursues for an ambiguous sentence. Based on an adaptive view of language processing (to be detailed below), we predict that (1) visual and linguistic information both constrain which interpretation is pursued by the sentence processor, and (2) the two types of information act upon the interpretation of the sentence at different points during processing.

We investigate these predictions by focusing on the notion of *saliency*. Saliency has several definitions, often referring to different phenomena. For the purpose of this paper, we conceptualize saliency as a set of low-level features of the input which gives prominence to specific aspects of the visual or linguistic information being processed. Thus, in our definition, saliency does not carry information directly related to a specific aspect of a comprehension task, but rather enhances the likelihood that the information made salient is used during comprehension.

In the visual cognition literature, saliency has a precise definition, referring to a composite measure of low-level visual information (color, intensity, and orientation), computed across different spatial scales (Itti & Koch, 2000). Visual saliency is known to guide visual attention during free viewing tasks; in which, in the absence of a specific target object, attention is attracted by highly salient regions, especially shortly after scene onset (e.g., Parkhurst, Law, & Niebur, 2002, see how-

[1] In this paper, we will use SMALL CAPS to indicate visual objects, and *italics* to refer to words and phrases.

ever Tatler, Baddeley, & Gilchrist, 2005 for contrasting results, and Einhäuser, Rutishauser, & Koch, 2008 for evidence of top-down control of attention at scene onset). Moreover, the attractiveness of visually salient locations also correlates with human judgments (Borji, Sihite, & Itti, 2013). In goal-oriented tasks, in which a target object is specified (e.g., visual search), saliency effects are weaker, if not neutralized. Attention is mostly directed to regions that are contextually related to the search target (e.g., looks to a TABLE when searching for a MUG, Henderson, Brockmole, Castelhano, & Mack, 2007). More recent studies, however, indicate that visual saliency can interact with goal-oriented, top-down processes. Coco, Malcolm, and Keller (2014), for example, demonstrate that visually salient objects are attended and also mentioned earlier than non-salient objects in an object naming task. Such effects are, nevertheless, modulated by other conceptual properties of the object, such as their contextual congruency.

In the psycholinguistic literature, saliency is often regarded as a feature associated with referents.[2] Saliency affects the prominence of a referent in a given linguistic (Ariel, 1990) or visual context (Fukumura, van Gompel, & Pickering, 2010). A referent that is more salient (prominent) is more easily accessible in the subsequent discourse, although for some cases the saliency of a referent might depend on its form (see Kaiser, Runner, Sussman, & Tanenhaus, 2009). In the present study, we approach linguistic saliency from a more perceptual perspective and focus on prosodic prominence; in particular, we are interested in the effect of intonational breaks, i.e., the pauses separating constituents.

An intonational break constitutes low-level linguistic information, as it does not contribute directly to the meaning of a referent. Rather, it affects the grouping of referents into larger phrases.[3] This definition of intonational breaks is used by previous studies of syntactic ambiguity resolution in communicative tasks situated in a visual world. In this context, intonational breaks are used by speakers to contrast the intended referent with a referential competitor present in the same visual context (Snedeker & Trueswell, 2003; Snedeker & Yuan, 2008; Ito & Speer, 2008).

As an example, take Snedeker and Trueswell's (2003) study, which presented ambiguous sentences such as *put the frog with the flower in the box*) in a visual context containing a single FROG, a FROG WITH A FLOWER, and a FLOWER. The results show that speakers either combine the ambiguous prepositional phrase *with the flower* with the direct object into a single intonational phrase, i.e., *the frog with the flower [BREAK] in the box* to obtain a modifier interpretation, or they set the break after the direct object, i.e., *the frog [BREAK] with the flower in the box* to obtain an instrument interpretation. This effect was later confirmed using eye-movement responses, where an instrument break triggered more looks to the FLOWER during the ambiguous region *with the flower*, compared to a modifier break (Snedeker & Yuan, 2008). Furthermore, a weaker effect on eye-movements was also found by Bailey & Ferreira, 2007, who used filled pauses instead of intonational breaks, using stimuli such as *put the apple [uh uh] on the towel in the box.* It is important to stress that an intonational break does not manifest itself merely as a pause between words, but also correlates with the duration of the word preceding the break, and can also result in pitch accent changes (Ferreira, 1993).

In a recent study, Vogels, Krahmer, and Maes (2012) investigated the interaction between the saliency of visual and linguistic information during language production (using a sentence comple-

[2] In the context of foreign accents, saliency is sometimes also regarded as an acoustic feature (Bradlow, Clopper, Smiljanic, & Walter, 2010).

[3] Though intonational information can also have pragmatic effects, or affect the truth-conditional meaning of a sentence (Jackendoff, 2002).

tion task). Visual saliency was manipulated as the size of the referent object (either in the *foreground* or *background* of the scene), whereas linguistic saliency was manipulated by changing contextual information about the referent object. A context sentence was given prior to the production task, which either did, or did not, refer to the referent object. The authors found that visual saliency boosts the likelihood of producing referring expressions about the target object, but does not impact on the form used (e.g., full NP, pronoun). On the other hand, linguistic saliency had a negative impact on the likelihood of referring to the target object. Crucially, there was no significant interaction between the two types of saliency. This result is consistent with the assumption that visual and linguistic information are used to optimize different aspects of language production.

Moreover, recent work by Ferreira, Foucart, and Engelhardt (2013) on situated understanding of syntactically ambiguous sentences found that the resolution strategies observed relate to the visual complexity of the context, as well as to the amount of preview given prior to comprehension. The authors suggest that in a situation where a preview of the visual scene is available (or where the scene is simple), participants pre-compute the affordances of the objects depicted, and then use these objects to incrementally fill a syntactic template as the speech unfolds. In contrast, when no preview is available (or the scene is complex), both visual object recognition and linguistic processing have to happen incrementally, creating a situation in which comprehenders use a more superficial processing strategy. More broadly, Ferreira et al.'s (2013) results indicate that the human language processor, rather than using a fixed processing strategy, is adaptive, and can use visual and linguistic information as and when it becomes available, tailoring its strategy to the current task and the processing stages within that task.

In the current study, we bring this line of investigation one step forward by investigating how visual and linguistic saliency are accessed during situated language comprehension when they are, or are not, concurrently available. The critical issue examined is that visual and linguistic information are available at different points during comprehension. Information about salient parts of the visual scene is available throughout the trial: prior, during, and after a sentence has been presented, while linguistic saliency is local to the relevant points in the speech signal.

However, even though visual saliency is always available, it might be particularly useful during certain phases of the language comprehension task. For example, visual saliency could have an effect at the start of trial, when attention is drawn towards salient regions of a scene (Itti & Koch, 2000). But saliency could also be useful later on: a sentence unfolds incrementally and the sentence processor anticipates upcoming linguistic material given the current linguistic and visual context (Altmann & Kamide, 1999). Such predictions happen, for example, during the processing of the verb phrase in a transitive structure (e.g., *the girl will put . . .*); here, a salient object might be anticipated as the argument of the verb. In contrast to visual saliency, linguistic saliency is available only locally as the speech stream unfolds: for instance, predictions based on intonational breaks can only happen once the referent delineated by the intonational breaks has been fully perceived.

These theoretical considerations, together with the results of Vogels et al. (2012) and Ferreira et al. (2013), therefore suggest an adaptive architecture of cross-modal processing, in which the language processor recruits visual and linguistic information for different aspects of the comprehension task. The two types of information are complementary and brought to bear at different points in time during language processing. An example for a model that is compatible with this view is the constraint satisfaction model proposed by MacDonald, Pearlmutter, and Seidenberg (1994), in which listeners utilize lexical constraints to incrementally resolve temporal ambiguity in the speech input, and pursue the interpretation that optimally satisfies these constraints at the current point in

time. It is natural to assume that visual constraints in addition to lexical constraints are evaluated in MacDonald et al.'s framework.

In the present article, we test this view of the language/vision interaction using a classic VWP task which requires the resolution of syntactic attachment ambiguities (e.g., Tanenhaus et al., 1995). Such ambiguities occur when phrases can be combined in different syntactic configurations, licensing distinct semantic interpretations of the sentence. A well-known instance is the prepositional phrase (PP) attachment ambiguity. An example is *put the apple on the towel in the box*. In one interpretation, the PP *on the towel* is interpreted as a modifier of *apple*, i.e., the referent is the apple that is on the towel. In the other interpretation, *on the towel* is a goal location, i.e., the apple is put on the towel that is in the box. The presence of a visual context acts upon the resolution of syntactic ambiguity by constraining which interpretations are available. Depending on the visual information present, certain resolutions will be more probable than others.

Tanenhaus et al.'s (1995) study, for instance, demonstrates that the syntactically ambiguous phrase *on the towel* is resolved differently when the object APPLE is depicted in the context of another APPLE on a NAPKIN or with a DISTRACTOR (1 Referent vs. 2 Referent condition). The precondition for observing the different resolutions is that in both contexts an EMPTY TOWEL is depicted. In the one referent context, the EMPTY TOWEL is ambiguously interpreted as the goal location for the putting action (i.e., move the APPLE from the one TOWEL to the other TOWEL); whereas in the two referent context, such ambiguity is resolved by the presence of the visual competitor APPLE ON NAPKIN. This clearly shows that the interpretation of a sentence depends on the plausibility of the referential choices made available by the visual context. The visual availability of the two depicted APPLEs promotes the interpretation in which the single APPLE is moved onto the EMPTY TOWEL.

The visual information conveyed by objects is not only referential, but also perceptual (see Huettig & Altmann, 2011, for a study on the interaction between perceptual and conceptual properties of referents). Different APPLEs could be visually dissimilar (a red vs. a green apple) even if they can both be referred to by the word *apple*. Moreover, the linguistic information conveyed by a sentence can go beyond the information pertaining to the referents. As shown by Snedeker and Yuan (2008), linguistic saliency, in the form of intonational breaks, plays a crucial role in ambiguity resolution; changes of prosodic prominence can modulate the way linguistic information is interpreted.

In the present study, we first investigate the effects of visual saliency and intonational breaks in a mono-modal setting by manipulating them separately, before looking at a cross-modal setting, in which both types of information are manipulated at the same time. Importantly, we are not claiming that the two types of manipulation have an equivalent impact on the eye-movement pattern during syntactic ambiguity resolution, as visual and linguistic information are known to trigger distinct effects. Instead, we want to test how the two types of information are used when they are concurrently available during situated language understanding, i.e., for which subtasks (e.g., prediction or disambiguation) they are utilized, and at which point in the incremental processing of the sentence.

Specifically, Experiment 1 investigates the impact of visual saliency on PP attachment ambiguity resolution. During sentence understanding, referential information builds up incrementally as the linguistic input unfolds; therefore, the less linguistic information has been processed, the more uncertainty there is about the referents that will be mentioned in the sentence. Thus, we expect visual saliency to be used by the sentence processor to predict upcoming arguments, especially early on in the sentence, when argument referents are not yet available. If the effect of visual saliency is

related to anticipatory processes of sentence understanding, then fixation to the salient object (compared to non-salient objects) should increase as the verb unfolds. For our example *put the orange on the tray in the bowl* corresponding to the visual context shown in Figure 1, we would expect this anticipation process to be measurable at the offset of *put*, with increased looks to the TRAY IN BOWL if this object is visually salient (indicating a modifier interpretation), and more looks to BOWL if this object is salient (indicating a goal location interpretation).

Experiment 2 uses a very similar design as Experiment 1 to test the effect of intonational breaks on PP attachment ambiguity resolution. Instead of making referents prominent through visual saliency, we make them prominent by grouping words together using intonational breaks. In our running example, we would expect participants to prefer the modifier interpretation if there is an intonational break between *the tray* and *in the bowl*, while they should prefer the goal location interpretation if the break is between *the orange* and *on the tray*. Such a finding would be consistent with the previous literature, where it has been shown that intonational breaks focus attention on the visual object referenced by the intonational phrase (e.g., Snedeker & Yuan, 2008). In this context it is important to recall a principled difference between visual and linguistic salience: the former is available throughout the whole sentence, while the latter is inherent to the speech stream and therefore available for ambiguity resolution only locally.

Experiment 3 investigates the interaction of both forms of saliency. We combine the manipulations of Experiments 1 and 2 so that visual saliency and intonational breaks are either consistent or inconsistent with respect to the resolution of the syntactic ambiguity. In the consistent condition, both forms of saliency agree, i.e., they both give prominence to the same interpretation. In the inconsistent condition, they give prominence to different interpretations. The inconsistent condition allows us to establish whether visual and linguistic information is utilized at the same time and for the same sub-task of sentence comprehension (in which case we should see a conflict). If however, as Ferreira et al.'s (2013) account predicts, the sentence processor is adaptive, then we expect it to use the two types of information at different stages in the incremental understanding of the sentence, and for different sub-task of sentence comprehension. In this case there should be no conflict, and expect to observe the same eye-movement patterns in Experiment 3 as in Experiments 1 and 2, in which linguistic and visual saliency are tested in isolation.

Throughout this paper, we investigate the impact of visual saliency and linguistic prominence on eye-movement responses on the two target objects with which our saliency manipulations are directly associated (i.e., Single Location and Compound Location, explained below). We limit our analyses to two temporal region of interest: the direct object region (ROI:NP, *the orange* in our example) and the first prepositional phrase (ROI:1PP, *on the tray*). During the direct object region, we specifically examine whether the visual saliency of the object is utilized to predict upcoming arguments of the sentence. As argued before, visual saliency exerts guidance on visual attention especially when there are no specific targets to look for. Thus, at this region, we test whether anticipatory eye-movement driven by the visual saliency of the objects develops during the verb and manifest itself at the onset of the direct object. During the first prepositional phrase, in contrast, we investigate ambiguity resolution rather than anticipation, as this is the region in which ambiguity is first encountered. Here, we expect linguistic prominence to play an important role. In line with previous literature, an intonational break is expected to give prominence to the object enclosed by it. In the context of syntactic ambiguity resolution, this should lead to the sentence processor interpreting the object enclosed by the intonational break as the goal location of the direct object for the *putting* action.

As our study comprises of three experiments, each rich both in possible target objects and linguistic regions of interest, we only report results on the objects and regions discussed in the previous paragraph, as these are directly relevant to the hypotheses investigated. Results on additional objects and regions, which provide corroborating evidence, are reported in the Supplemental Material. A paragraph summarizing the results reported in the Supplemental Material is made available in the present manuscript.

Experiment 1: Visual Saliency in Syntactic Ambiguity Resolution

This experiment investigates the impact of visual saliency on syntactic ambiguity resolution. We test the hypothesis that the sentence processor takes the saliency of visual objects into account when incrementally assigning an interpretation of the linguistic material to the objects of the visual context.

Participant in a situated language processing task pursue a sequence of goals. At the beginning of the trial, before speech begins, participants are free-viewing the scene, thus image-based information is mostly expected to guide visual attention at this stage. During speech, visual attention is expected to be driven by linguistic information. However, there are intermediate phases while speech is unfolding, e.g., while the verb is being processed, where linguistic information is not yet sufficient to predict all upcoming arguments. Thus, at this point of the speech stream, we expect visual saliency to be utilized by the sentence processor to predict upcoming arguments of the verb.

In particular, we expect to observe a stimulus-driven shift of visual attention at scene onset to the visually salient object in the display (see Supplementary Material: Decay of Visual Saliency Effect During Preview). After this initially strong shift, effects of visual saliency decay and remain at a baseline level until the arguments of the verb are spelled out. During this phase, visual saliency is utilized as a predictive proxy for these arguments. A rise in fixations to the salient target should therefore be interpreted as the sentence processor utilizing saliency, rather than as a language-independent effect of saliency. This is especially the case if the anticipatory effect of visual saliency dissipates after the verb arguments have been spelled out. In this case we expect to observe the same amount of fixations to salient and non-salient objects by the end of the linguistic region of interest, i.e., the direct object *the orange*.

Method

In a visual world eye-tracking experiment, participants viewed a visual scene and concurrently listened to sentences containing a PP attachment ambiguity. Look-and-listen experiments use a simpler setting than goal-oriented tasks (e.g., a display screen instead of a real array of objects) and aim to trigger language comprehension effects purely based on sentences interpretation, rather than based on physical actions. Thus, instead of imperative sentences (*put* ...), we used declarative sentences (*the girl will put* ...) and depicted the subject of the action in the visual scene. Such look-and-listen experiments have been used extensively in the VWP literature (Novick, Thompson-Schill, & Trueswell, 2008), even though there are important theoretical consequences that follow from using an experimental task which does not require a goal-directed action (see Salverda, Brown, & Tanenhaus, 2011). An example for a sentence stimulus used in our experiment is the following:

(1) The girl will put the orange on the tray in the bowl.

The experiment used a 2×3 design, crossing the factors *Number of referents* (*1 Referent* or *2 Referents*) and *Saliency* of objects (*Single Location*, *Compound Location*, or *No Saliency*). The experimental manipulation is illustrated in Figure 1.

Number of Referents refers to the number of visual objects associated to the direct object, *the orange*. In the *1 Referent* condition, only one ORANGE is depicted ON A TRAY, whereas in the *2 Referent* condition, the context contains also a single ORANGE instead of the distractor, i.e., the BOTTLE in Figure 1. This experimental condition was introduced to replicate the classic effect of referential competition, i.e., the more visual objects share the same referent, the more attention will be distributed among them (Tanenhaus et al., 1995). Moreover, this manipulation can also help us establish whether having more resolutions (in the two referents condition) results in the sentence processor resorting more or less strongly to saliency. *Saliency* refers to the visual object carrying the saliency manipulation, where *Single Location* is the object associated with the goal location (i.e., BOWL for the prepositional phrase *in the bowl*), *Compound Location* is the visual compound object associated with the prepositional modifier (i.e., TRAY IN BOWL, for the interpretation *on the tray in the bowl*) and *No Saliency* is the baseline condition, in which saliency is not manipulated. The Single Location object is visually simpler and semantically more plausible than the Compound Location. Plausibility is known to have a direct impact on fixation distribution (Chambers, Tanenhaus, & Magnuson, 2004), and we will discuss the implication of plausibility and complexity on fixation distribution when presenting relevant results.

Similar to Bailey and Ferreira (2007), we use fully ambiguous visual contexts. In the example shown in Figure 1, the object TRAY associated with the ambiguous modifier *on the tray* is depicted with an ORANGE (ORANGE ON A TRAY), and with a BOWL (TRAY IN BOWL). Such a visual context is compatible with both local and global ambiguities. Local ambiguities arise on parsing the ambiguously attached PP, e.g., *on the tray*, and are progressively resolved through competition between the visual referents associated with the different interpretations (e.g., single ORANGE, ORANGE ON TRAY, and TRAY IN BOWL). Global ambiguity refers to the fact that even after parsing the whole sentence, the interpretation remains ambiguous, i.e., linguistic and visual referents can still be related to each other in two ways: the ORANGE can either finish in the BOWL or on the TRAY IN THE BOWL.

We allow both local and global ambiguity in our materials in order to increase referential competition, and to make the manipulation of intonational breaks, presented in Experiment 2 of this paper, comparable with the work of Bailey and Ferreira (2007) and Snedeker and Yuan (2008). Setting up our materials this way makes it possible to map each referring expression to a precise depicted referent when enclosed by the relevant intonational breaks. Note that there would be no target object for the phrase *on the tray in the bowl* for the PP modifier condition if we did not allow global ambiguity in the visual context, i.e., if the target TRAY IN BOWL was not depicted.

The experiment also included a between-participant condition that manipulated scene preview. In the *Preview* condition, participants had 1000 ms of visual preview before the onset of the speech stimuli; in the *No Preview* condition, speech and visual stimuli started simultaneously at beginning of the trial. In visual search tasks using naturalistic scenes, change of preview-time has a direct impact on search performance, with longer preview boosting identification of target objects (Vo & Henderson, 2010). Preview time has also been found to significantly influence visual attention in object arrays with large number of targets (Ferreira et al., 2013). Here, we test whether such preview effect carries over to sentence understanding, situated in an object array containing a relatively small number of objects. Also, we wanted to investigate whether preview time significantly inter-

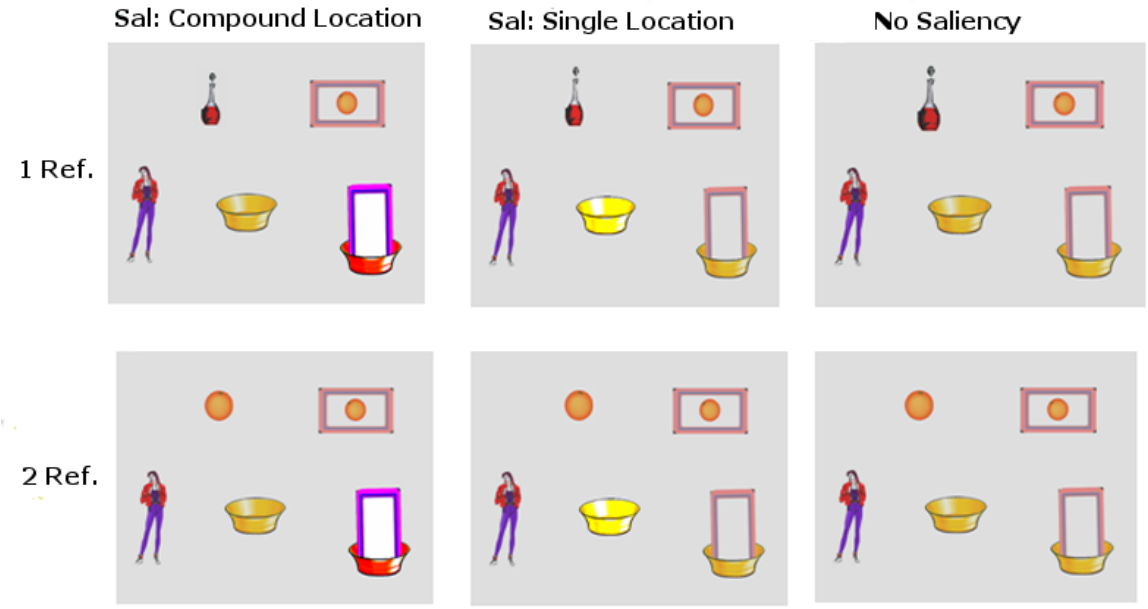

Figure 1. Example stimuli for Experiment 1. The design crosses Number of referents (1 Referent, 2 Referents) with Saliency (Single Location, Compound Location, No Saliency).

acts with visual saliency while an ambiguous sentence is processed. We expected less fixation to the mentioned target, e.g., fixation to the ORANGE when the direct object *the orange* is spoken, in the no-preview condition compared to the preview condition. Presumably, in the no-preview condition, visual attention is still engaged in recognizing all objects of the visual context when the linguistic material is concurrently presented.

Materials. For each condition, we created a set of 36 pairs of experimental items (sentence–scene pairs). In the scenes, we used Photoshop to manipulate the saliency of the target object by altering the object's luminosity (+50 percent), contrast (+50 percent), and RGB color balance to reinforce the color dominance of the target object. We also manipulated the black and white input/output curves of the objects in cases where the edges of the target object were too prominent. To validate the saliency manipulation, we used the Saliency Toolbox (Walther & Koch, 2006) to confirm that the target object had a higher saliency value than all the other objects in the visual display.

Compared to the sentences in the original Spivey-Knowlton, Tanenhaus, Eberhard, and Sedivy (2002) study, we increased the lexical variability by including synonyms of the verb *put* (*move*, *place*, and *lay*). Each of the four verbs was used in nine sentences.

The sentences were recorded with two different intonations by a female native speaker of English at a normal speech rate (refer to Experiment 2 for details). However, since the current experiment focuses on effects driven by visual saliency, we neutralized the effect of intonation by cross-splicing the sentences. We used Adobe Audition to split and merge the two different recordings of each sentence. We took the direct object (e.g., *the orange*) and the second prepositional phrase (e.g., *in the bowl*) from sentences produced with a PP modifier intonation, where a break is placed between the direct object and the first preposition. We took the first PP (e.g., *on the tray*) from an NP modifier intonation, where the break is placed between the first and the second preposition.

We then merged the fragments as follows:

(2) the orange [$_{\text{PP}}$ modifier] + on the tray [$_{\text{NP}}$ modifier] + in the bowl [$_{\text{PP}}$ modifier]

Between each phrase, a 50 ms pause was added to yield a more realistic prosody. The resulting materials sounded natural, and participants were not able to detect that they were cross-spliced when queried during the post-experimental debriefing. We decided to do cross-splicing instead of re-recording the materials in order to prevent the speaker from introducing intonational cues involuntary.

In addition to the 36 experimental items, the experiment included 48 fillers, which were object arrays with the same structure (5 objects depicted) paired with unambiguous sentences. We counterbalanced visual saliency in one third of the fillers to obtain a balanced set (i.e., half of the images with, and half without salient objects), and to prevent participants from associating visual saliency with sentence ambiguity. To counter viewing bias effects, we also rotated the visual objects in eight different configurations, clockwise and along the diagonals. Materials were distributed across six lists in a Latin Square design. Individual randomizations were created for each participant, making sure that between two experimental items there was always at least one filler.

Procedure. An SR 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 $\times$ 768 pixels. A test of eye dominance was performed at the beginning of each session and only the dominant eye was tracked. Participants were asked to wear swimming caps in order to prevent the eye-tracker from sliding during the experiment.

The experiment was explained to participants using written instructions. At the beginning of every session, participants were given four practice trials to familiarize them with the experiment. After each image was presented for 0 or 1000 ms (depending on the preview condition), the corresponding recorded sentence was played. Once every six trials, participants had to respond to a question about the content of either the sentence or the scene. This was done to ensure participants engaged with the task. Calibration was carried out at the beginning of the session and repeated again approximately halfway through the session. Some participants required more than two calibrations. We performed drift correction between trials. The entire experiment was approximately 30 minutes long.

Participants. Thirty students at the University of Edinburgh, all native speakers of English, participated in the experiments. The participants gave informed consent and were paid 5 pounds each.

Analysis. The analysis focuses on the two linguistic temporal regions of interest (ROIs): the direct object region (ROI:NP, *the orange* in our example) and the first prepositional phrase (ROI:1PP, *on the tray*). At ROI:NP, we investigate anticipatory phenomena, and examine whether visually salient objects are activated during the processing of the verb to predict upcoming verbal arguments. At ROI:1PP we investigate syntactic ambiguity, and examine whether visual saliency is utilized to guide its resolution. In this region, more looks to the salient object would indicate its selection as possible goal location for the direct object. Note that we focus on specific linguistic temporal regions, rather than on the whole sentence, because we are interested in showing *when* in the sentence crucial effects emerge. Individual sentences vary in length, and in the absolute position of linguistic regions; it is therefore important to focus on fixation data that is temporally aligned to the linguistic regions at which the experimental manipulations are expected to produce effects. Without such alignment, we could not statistically assess whether any fixation patterns observed are trig-

gered by a specific linguistic ROI. (For readers interested in descriptive patterns of fixations across the whole sentence, we provide relevant plots in the Supplementary Material, Time-course Plots across the Whole Sentence for Single-Location, Compound-Location and Other objects, focusing on Experiment 3, in which both visual saliency and linguistic prominence were manipulated.)

Each of our analyses will focus on a specific target object and compare the effect of the experimental manipulations (number of referents, visual saliency) on the fixations on this object. In the main text, we consider only two objects: Single Location (BOWL) and Compound Location (TRAY IN BOWL), which are the objects more directly affected by our experimental manipulations. In the Supplementary Material, we will also report results for the objects ORANGE/DISTRACTOR and ORANGE ON TRAY. Analyzing fixation probabilities for specific target objects (rather than on all objects) makes it possible to directly observe the effect of our experimental manipulations. Also, a comparison of fixations across all objects for the four conditions in three different linguistic ROIs would drastically increases the number of plots reported, making the result section very hard to follow.

For each target object, we align fixations with the onset of the different temporal linguistic region in the speech (e.g., the phrase *the orange*) and consider a window of 600 ms for the ROI:NP (duration = 597.98 ± 152.38) and 700 ms for the ROI:1PP (duration = 725.26 ± 133.53) from 100 ms after their onset to account for the oculo-motor delay, i.e., the time needed to launch a saccade in response to the auditory input. A fixation is counted from the onset of the saccade leading into it, and we consider a basic unit of two milliseconds, because it corresponds to the sampling rate (500 Hz) of the SR Eyelink II.

As the onset of the linguistic ROI varies across sentences, we align fixations on an item-by-item basis, i.e., we use the timestamps for the linguistic ROI onset of each sentence to decide which fixations to include. Moreover, fixations crossing the end of the ROI (i.e., its offset) are excluded from the analysis, again on an item-by-item basis. This step ensures that fixations that are potentially contaminated by the next word (e.g., the ROI:1PP when the region analyzed is the ROI:NP) are excluded, which accounted for 2% of the data.[4] In our plots, we do not mark the beginning or end of the ROI, as we are only considering fixation strictly inside of the ROI.

For both visualization and data analysis, we use fixation probabilities on the target object, aggregated over windows of 100 ms (50 points per window) to reduce the correlation between consecutive time bins, as they cannot be considered independent samples. The aggregation into larger windows helps us to minimize Type 1 error (D. Barr, 2008). Moreover, an aggregation over 100 ms is large enough to overcome the dichotomous nature of eye-movement data, i.e., a participant might be fixating two different objects within the same time window, even when fixations are kept at the level of individual trials (rather than aggregated by participants or items). We decided to follow this procedure because, we did not want to treat fixations as a binomial (sampled every 2 ms), and have to apply arbitrary decision to aggregate them into larger window (e.g., 25 fixation points into a 50 data points window).

We use linear mixed effect (LME, Pinheiro & Bates, 2000) models to statistically analyze our fixation data. The LME analysis uses the fixation probability as the dependent variable (6 or 7 time bins of 100 ms each) and the following centered predictors: *Saliency* (Single Location, Compound Location, No Saliency), which is contrast coded using No Saliency as reference level, *Number of Referents* (1 Referent, -0.5; 2 Referents, 0.5) and *Time* represented as orthogonal polynomial of

[4] We also tried to only exclude fixations launched after the end of the ROI, and we obtain exactly the same pattern of results.

order two (Time^1, Time^2). The polynomial representation of Time, originally proposed by Mirman, Dixon, and Magnuson (2008), gives us a better way of capturing the temporal dynamics of fixations, and returns a more accurate estimate of the model fit. The main reason for using this approach is that fixations almost never distribute linearly in time; fitting a spline allows us to estimate non-linear changes in fixation probability across time. We use a polynomial of order two. The linear term of the polynomial has exactly the same interpretation as a linear regression of fixations over time. The quadratic term can be used to identify sudden changes in the linear trend, e.g., a decrease followed by an increase. Higher order terms improve the model fit but are difficult to interpret (Mirman & Magnuson, 2009). Note that we significantly depart from Mirman et al. (2008), who aggregates fixation data by participants; as outlined above, we only aggregate over temporal points within the individual trials. This enables us to include randoms intercepts and slopes for both participants and items, rather than having to have two separate analyses for participants and items, which can potentially yield inconsistent results. By including both random effects for participant and item, we have a more conservative model in that the true probability of incorrectly rejecting the null hypothesis (i.e., the Type 1 error) is not greater than the nominal level. The inclusion of random efect structures with both participants and items is advocated by Baayen, Davidson, and Bates (2008) and, more recently, by D. J. Barr, Levy, Scheepers, and Tily (2013).

We utilize a best-path forward selection procedure to obtain our final model, which is shown by D. J. Barr et al. (2013) to give a rate of Type-1 error comparable to a model with a maximal random structure (which does not converge on our data). We start with an empty model and add random intercepts, fixed predictors, and uncorrelated random slopes on predictors,[5] step by step based on log-likelihood improvement of the model fit. To assess whether the improvement is significant, we compare nested models, i.e., the model with and without an additional parameter, using a χ^2 test. Note that we diverge from the best-path forward selection approach described by D. J. Barr et al. (2013) in that we perform the selection of random slopes in tandem with the selection of fixed effects.

In the results section, we visualize the observed fixation probability for the different experimental conditions as shaded bands indicating the standard error around the mean, and overlay the model fits as lines. In the results tables, we report the coefficients and standard errors of the LME models, and derive p-values from the t-values (also reported) for each of the factors in the model. The t-distribution converges to a z-distribution when there are enough observations, and hence we can use a normal approximation to calculate p-values. Moreover, for completeness, in the table, we also report the formula of model selected, using the R's `lme4` syntax.

Results

Before examining the effects of our within-participants conditions (i.e., *Saliency* and *Number of Referents*), we briefly discuss our between-participants preview manipulation (*No Preview* vs. *Preview*). In the no-preview condition, we found a trend of fewer fixation to the target object when the direct object was mentioned compared to a preview condition. However, this trend was not statistically significant. The lack of significance was consistent across the different temporal linguistic region considered.

Preview plays an important role in naturalistic scenes because these are visually complex, i.e., they contain many different objects in a coherent configuration (Vo & Henderson, 2010). A visual

[5] We also tested models with correlated random slopes, and found the same pattern of results.

array, in contrast, contains only a few disconnect objects, which presumably can be processed very rapidly, explaining the weak effect of preview in our study (see however Ferreira et al.'s (2013) study, which finds effects of preview in a comparable setup but with more visual objects present in the display). In all following analyses we will therefore aggregate the data from the short and long preview conditions.

The results section focuses on effects on the time course of fixations at the objects whose saliency was manipulated. In particular, we report results for the Single Location (e.g., BOWL) during the direct object region (ROI:NP) (e.g., *the orange*). We hypothesized that visual saliency can be used by the sentence processor to predict upcoming arguments. This means that at the beginning of the direct object, the sentence processor anticipates a suitable filler for this linguistic role, and uses saliency to select the relevant object. Note that this claim is based on the assumption that visual saliency does not have an effect overall (e.g., because more salient regions attract non-language related shifts of attention across the whole sentence). Rather, we expect a selective use of visual saliency, restricted to the phases of the spoken stream where there is uncertainty about the upcoming arguments, i.e., the beginning of the direct object region. Subsequently, as the direct object unfolds, visual attention becomes more strictly constrained by the linguistic information and the salient object should have an equal likelihood of being looked at as the object in the non-salient condition by the end of that region.

The classic referentiality effect, i.e., more looks to the single ORANGE in the 2 Referent condition, compared to looks to the DISTRACTOR in the 1 Referent condition, originally shown by Tanenhaus et al. (1995) is replicated in our study, and presented in Supplementary Material (Experiment 1: Referentiality Effect at ROI:NP *the orange* on ORANGE). Note that visual saliency did not have an effect on the regions where ambiguity emerges, i.e., ROI:1PP and we therefore omit this analysis. As the sentence unfolds, linguistic processing gradually becomes the main driver of visual attention, and purely visual information sources, such visual saliency, are less likely to be utilized.

In Figure 2, we show the probability of fixations on BOWL at the ROI:NP direct object *the orange* across the different experimental conditions. We observe a main effect of Saliency, in that for Single Location, there are significantly more anticipatory looks, compared to both No Saliency and Compound Location. This appears as a main effect of Single Location in the mixed effects model, as shown in Table 1.

The anticipatory effect of saliency, however, decays as the direct object unfolds (interaction SaliencySingleLoc:Time^1). Furthermore, when two referent are depicted, we observe a higher anticipation of looks to the visually salient Single Location. Possibly, the presence of a single target for the direct object ORANGE makes the Single Location a more likely goal location, hence increasing the likelihood to be anticipated as possible argument. We also observe also a significant main effect of SaliencyCompoundLoc, indicating that looks to the Single Location are reduced when visual saliency is on the Compound Location.

This is indeed the case. In Figure 3 we show the fixation curve for TRAY IN BOWL at the direct object *the orange*. We observe a main effect of saliency for Compound Location, i.e., the compound object TRAY IN BOWL receives more looks when it is visually salient, compared to the Single Location condition in which the BOWL is salient (refer to Table 2 for the estimates). The fixation curves show an increase followed by a decrease (interaction SaliencyCompoundLoc:Time^2): as information about the identity of the direct object accumulates, the effect of saliency diminishes. Visual saliency is utilized to predict upcoming linguistic information.

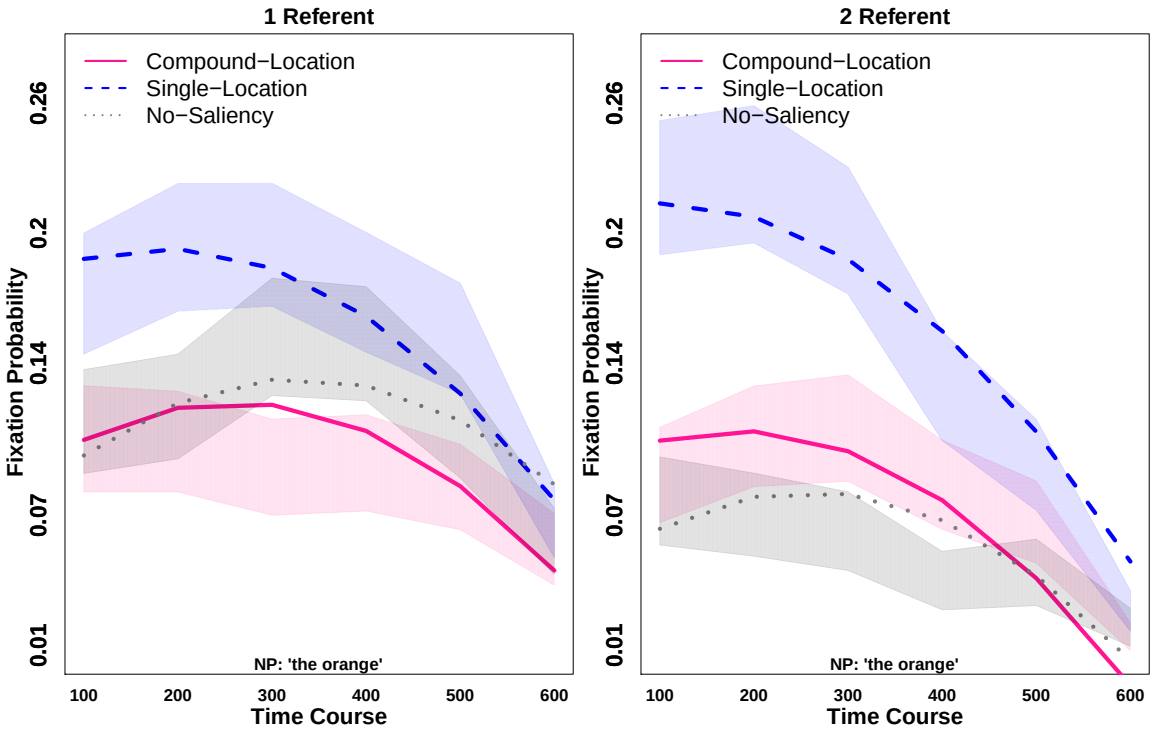

Figure 2. Experiment 1. Time course plot of fixation probability for the object BOWL (corresponding to the Single Location) from 100 ms to 600 ms at ROI:NP *the orange*. Left panel: 1 Referent condition, right panel: 2 Referent condition. The saliency conditions (No Saliency, Single Location and Compound Location) are marked through line types and colors. The shaded bands indicate the standard error around the observed mean. The lines represent the predicted values of the LME model reported in Table 1. Note that the offset of the region of analysis varied by items, but fixations crossing the offset were excluded, see Analysis section for details.

When linguistic information is not sufficient to generate a prediction about upcoming arguments, sentence understanding relies on image-based visual information to make such predictions. However, as soon as the linguistic referent *the orange* is spelled out, looks to the salient object decrease over time. Note that we can claim that the effect of visual saliency really results from *linguistic* anticipation, rather than from non-language related shifts of attention: we observe a decreasing trend of fixations while the direct object unfolds. If the effect of visual saliency was due to non-language related shifts of attention, then the salient object should instead have a higher likelihood across the whole region compared to the non-salient condition. Moreover, in order to further strengthen the point that a visually salient object does not per se have a higher likelihood of being fixated across the trial, we carried out an additional analysis, reported in the Supplementary Material (Decay of Visual Saliency Effect During Preview). Here, we analyzed the time-course of fixations during preview, and monitored whether purely visual saliency effects decay prior to the onset of the sentence. We demonstrate that the effect of saliency per se is restricted to the beginning of the trial (i.e., non-language related shifts of attention) and that it decays as the preview unfolds. We conclude that the effect we observe at the current region can therefore be attributed to the interaction of visual saliency with the verb and the onset of the direct object (i.e., we are dealing with linguistically

Table 1
Experiment 1. Mixed model analysis of the fixation probability to the object BOWL *(corresponding to the Single Location) at ROI:NP* the orange. Saliency *is contrast coded with* No Saliency *as reference level,* Number of Referents *is coded as* -0.5 *for 1 Referent, 0.5 for 2 Referents. Time (100 to 600 ms, in 100 ms intervals) is represented as an orthogonal polynomial of order two (Time^1, Time^2).*

Predictor	β	*SE*	*t*	*p*
Intercept	0.106	0.011	8.966	<1e-04
Time^1	-0.07	0.017	-4.067	<1e-04
SaliencySingleLoc	0.094	0.022	4.191	<1e-04
Time^2	-0.041	0.008	-4.751	<1e-04
Referent	-0.026	0.017	-1.501	0.1
SaliencyCompoundLoc	-0.051	0.018	-2.786	0.005
SaliencySingleLoc:Time^1	-0.089	0.021	-4.260	<1e-04
SaliencySingleLoc:Referent	0.053	0.017	3.092	.002
Referent:Time^1	-0.044	0.017	-2.548	.01

Formula: (1 | item) + (1 | participant) + Time^1 + (0 + Time^1 | item) + SaliencySingleLoc + (0 + SaliencySingleLoc | item) + (0 + SaliencySingleLoc | participant) + Time^2 + Referent + (0 + Referent | participant) + (0 + Referent | item) + SaliencyCompoundLoc + (0 + SaliencyCompoundLoc | item) + (0 + SaliencyCompoundLoc | participant) + Time^1:SaliencySingleLoc + SaliencySingleLoc:Referent + Time^1:Referent

driven anticipation).

We also analyzed looks during the linguistic regions in which ambiguity is expected to occur, i.e. 1PP, across the different target objects but failed to find any significant effect. This negative result supports the idea that visual saliency is active when the sentence processor is uncertain about upcoming arguments. However, as uncertainty decreases, visual attention becomes more and more constrained by the linguistic input that has been unfolding, and the role of visual information diminishes.

Discussion

In this experiment, we tested the hypothesis that low-level visual information such as visual saliency (Itti & Koch, 2000) plays a role during situated language comprehension. We found evidence for anticipatory looks to salient objects at the beginning of the direct object region (*the orange* in our example). The effect was modulated by the number of visual referents sharing the same linguistic referent, which we will discuss in a moment.

These results corroborate our hypothesis that situated understanding is a task that includes two phases: a free-viewing preview phase and a task-based language processing phase. Visual saliency is mostly active at the onset of the scene, then decays before the sentence onset (see Supplementary Material, Decay of Visual Saliency Effect During Preview), and is re-activated as speech begins, especially during the processing of the verb, which is a point of the speech stream where it is not yet possible to generate a full prediction of upcoming linguistic material. (Unless thematic dependencies can be extracted from pre-verbal arguments, as shown by Kamide et al. (2003) for Japanese, but this is not the case in this study, as the agent does not convey any predictive information about the direct object.) This finding is underscored by the fact that the verbs utilized in this

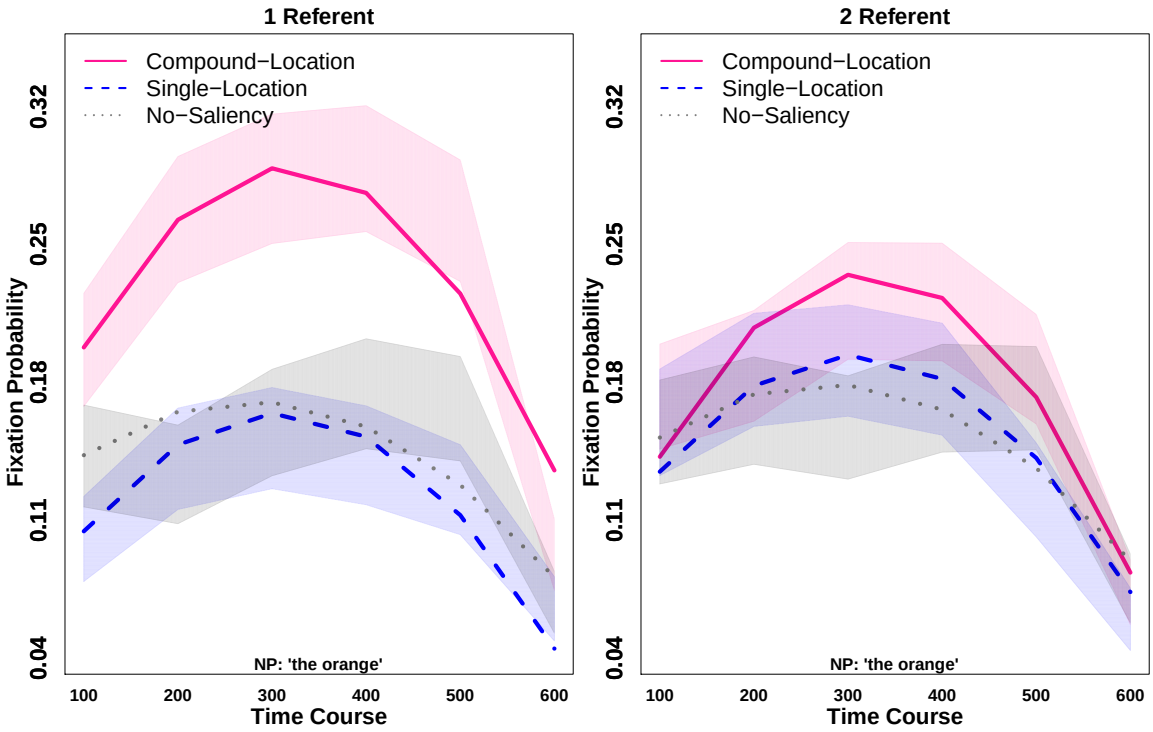

Figure 3. Experiment 1. Time course plot of fixation probability for the object TRAY IN BOWL (corresponding to the Compound Location) from 100 ms to 600 ms at ROI:NP *the orange*. Left panel: 1 Referent condition, right panel: 2 Referent condition. The Saliency conditions (No Saliency, Single Location and Compound Location) are marked through line types and colors. The shaded bands indicate the standard error around the observed mean. The lines represent the predicted values of the LME model reported in Table 1. Note that the offset of the region of analysis varied by items, but fixations crossing the offset were excluded, see Analysis section for details.

study, e.g., *put*, do not favor any particular object in the visual array (unlike in studies of prediction that manipulate verb selections restrictions such as Altmann & Kamide, 1999). However, once the targets have been linguistically identified, visual saliency effects cease to guide attention. Note that our results generalize Huettig and Altmann's (2011) study about the effect of color on shifts of attention by showing that visual saliency, and not just color, is predictively utilized during situated language understanding.

On a more general level, while the results of this experiment demonstrate an interaction between low-level mechanisms of visual attention and sentence processing, they do not clarify the pattern of interaction between the two modalities. In particular, this experiment used materials in which the information provided by visual saliency is *complementary* to the linguistic information. However, the crucial case is when the information in the two modalities is inconsistent, i.e., one modality suggests one interpretation, while the other modality suggests another. Investigating this case would allow us to test the key hypothesis discussed in the introduction, viz., that the sentence processor is adaptive. Under this view, we expect that the two types of information are used at different stages in the incremental understanding of the sentence, and for different sub-task of sentence comprehension.

Table 2

Experiment 1. Mixed model analysis of the fixation probability to the object TRAY IN BOWL *(corresponding to the Compound Location) at ROI:NP* the orange. Saliency *is contrast coded with* No Saliency *as reference level,* Number of Referents *is coded as* -0.5 *for 1 Referent, 0.5 for 2 Referents. Time (100 to 600 ms, in 100 ms intervals) is represented as an orthogonal polynomial of order two (Time^1, Time^2).*

Predictor	β	*SE*	*t*	*p*
Intercept	0.161	0.015	10.649	<1e-04
Time^2	-0.085	0.009	-8.692	<1e-04
SaliencyCompoundLoc	0.084	0.028	2.992	0.002
Time^1	-0.050	0.021	-2.379	0.01
SaliencySingleLoc	-0.055	0.025	-2.159	0.03
Referent	-0.002	0.019	-0.150	0.8
SaliencyCompoundLoc:Referent	-0.097	0.022	-4.314	<1e-04
Time^2:SaliencyCompoundLoc	-0.065	0.023	-2.731	0.006
SaliencySingleLoc:Referent	0.058	0.022	2.552	0.01

Formula: (1 | item) + (1 | participant) + Time^2 + SaliencyCompoundLoc + (0 + SaliencyCompoundLoc | participant) + (0 + SaliencyCompoundLoc | item) + Time^1 + (0 + Time^1 | item) + (0 + Time^1 | participant) + SaliencySingleLoc + (0 + SaliencySingleLoc | participant) + (0 + SaliencySingleLoc | item) + Referent + (0 + Referent | participant) + (0 + Referent | item) + SaliencyCompoundLoc:Referent + Time^2:SaliencyCompoundLoc + SaliencySingleLoc:Referent

In order to lay the ground for an investigation of the interaction of linguistic and visual information, Experiment 2 studies the effect of linguistic saliency in the form of intonational breaks, which give prominence to the linguistic material they enclosed. The effect of intonational breaks can be investigated with the same materials, but separately from visual saliency. Once we have established the effects of visual saliency and linguistic prominence in Experiments 1 and 2, respectively, Experiment 3 can investigate the interaction of the two types of prominence.

Experiment 2: Linguistic Saliency in Syntactic Ambiguity Resolution

This experiment investigates the effect of intonational breaks on the processing of sentences containing PP-attachment ambiguities, using the same design as Experiment 1. Intonational breaks are not directly linked to semantic information, but can be used to give prominence to referring expressions in a sentence. This means that intonational breaks are a source of information that is comparable to visual saliency, which acts on the low-level prominence of visual objects. As stressed before, we do not claim that the two saliency manipulations are equivalent. Rather, our aim is to compare two ways of increasing the prominence of a referent object, in the visual and linguistic domain, and to investigate how they interact with each other during spoken language understanding.

In line with the previous literature, we expect the visual objects associated with a referring expression bounded by appropriate intonational breaks to receive more looks compared to referring expressions which are not intonationally prominent (Snedeker & Yuan, 2008). For example, a break introduced before the second preposition should increase the likelihood that the referring expression that precedes it (in our example, *the orange on the tray*) is associated with the corresponding visual object (here, ORANGE ON TRAY).

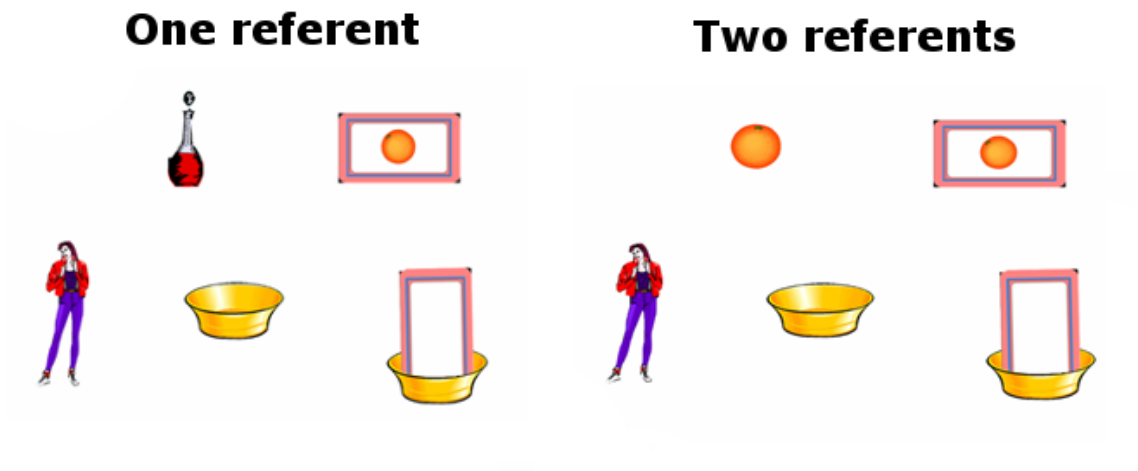

Figure 4. Example stimuli for Experiment 2. The design crosses Number of referents (1 Referent, 2 Referents) with Intonational Breaks (NP modifier, PP modifier).

Method

The experimental design crossed the factors *Number of Referents* (1 Referent or 2 Referents) and *Intonational Break* (NP modifier or PP modifier). As in Experiment 1, participants listened to sentences with ambiguous PP attachments such as (1) while concurrently viewing a visual context that supported both interpretations. We manipulate the intonational breaks similarly to Snedeker and Yuan (2008), considering two cases:

(3) *NP modifier:* intonational break after the first prepositional phrase (1PP)
[The girl will put] [the orange on the tray] [in the bowl]

(4) *PP modifier:* intonational break after the second prepositional phrase (2PP)
[The girl will put the orange] [on the tray in the bowl]

For the NP modifier interpretation, the intonational break is placed after the first PP, and *the orange* and *on the tray* are uttered as a single intonational phrase triggering the unambiguous interpretation *the orange that is on a tray, is put in the empty bowl*. In terms of eye-movements, we expect more looks to ORANGE ON TRAY at 1PP *on the tray*, compared to the PP modifier condition. For the PP modifier interpretation, an intonational break is placed after the NP direct object, *the orange*, thus combining the first and second modifier in the same intonational phrase. The compound object TRAY IN BOWL is expected to be in referential focus and receive more looks, when the first modifier *on the tray* is uttered, compared to the NP modifier condition.

As in Experiment 1, *Number of Referents* refers to the number of visual objects corresponding to the direct object of the sentence, e.g., *the orange*. In a 2 Referent visual context, there is a single ORANGE and AN ORANGE ON A TRAY; whereas in the 1 Referent condition, only an ORANGE ON A TRAY as depicted in Figure 4.

Materials. We used the images from the No Saliency condition of Experiment 1, as saliency was not manipulated in this experiment. For the spoken stimuli, we used the same list of sentences as in Experiment 1, without splicing, for the two types of intonational breaks. For the

NP modifier condition, a mean pause of 413 ms occurred between the end of 1PP *on the tray* and the beginning of 2PP *in the bowl*; whereas for the PP modifier condition, there was a mean pause of 637 ms between the end of direct object *the orange* and the beginning of 1PP *on the tray*.

Since it is known that pauses are positively correlated with the duration of the word preceding the break, and give rise to different pitch accents (Ferreira, 1993; Snedeker & Trueswell, 2003), we carried out t-tests to validate our materials. We found that the duration of the direct object, e.g. *the orange*, was longer in the PP modifier condition (mean = 689.22; SD = 110.05) than in the NP modifier condition (mean = 506.75; SD = 133.31; $t(140) = 8.95$, $p = 0.0001$). In contrast, the duration of the first modifier, e.g., *on the tray*, was longer in the NP modifier condition (mean = 776.94; SD = 117.96) than in the PP modifier condition (mean = 673.58, SD = 128.76; $t(140) = 5.02$, $p = 0.0001$).

We also looked at the pitch accent of the two different phrases for the two intonational break condition. We computed mean pitch frequency of each phrase using `Praat` (Boersma & Weenink, 2013). We found that the direct object, e.g., *the orange* had a higher pitch frequency in the NP modifier condition (mean = 226.38, SD = 7.70) than the PP modifier condition (mean = 214.75, SD = 9.53; $t(67), = 5.69$, $p = 0.0001$). The opposite pattern was found for the first preposition, e.g., *on the tray*: here, for the PP modifier condition, we found a higher pitch frequency (mean = 219.86, SD = 11.45) than in the NP modifier condition (mean = 206.53, SD = 6.36; $t(67), = 6.10$, $p = 0.0001$). The first word of the intonational phrase received a more marked stress than its second word, as it is the referent to which the modifier is going to operate.

Procedure, Participants, and Analysis. Thirty-two new participants from the same population as in Experiment 1 (i.e., students at the University of Edinburgh) took part in the study. Calibration was carried out at the beginning of each session and manual drift correction was done between trials. As in Experiment 1, we analyzed looks on individual target objects across conditions, which were aligned at the onset of the different linguistic regions of interest.

The experimental procedure was the same as that used in Experiment 1, except that we only had one preview condition of 1000 ms instead of two, as the manipulation of the preview had no effect on visual responses (see Section).

As in Experiment 1, we analyze the data using linear mixed effect models, using time-course probabilities of fixations to the target object during a given linguistic region of interest across experimental conditions. The model predictors were *Intonational Break*, *Referents*, and *Time*, and the random effects were *Participant* and *Item*. Time was encoded as before using an orthogonal polynomial of order two (i.e., a quadratic spline), and the same model selection procedure as described previously was used.

For this experiment, we will consider the first prepositional phrase (ROI:1PP, *on the tray*), as this is where the effect of the intonational breaks is expected (Snedeker & Yuan, 2008). We analyze looks to the visual objects referred to by this linguistic region, i.e., TRAY IN BOWL; Supplementary Material, Experiment 2: Linguistic Prominence Effect on ROI:1PP, *on the tray* on ORANGE ON TRAY, reports corroborating results on the ORANGE ON TRAY object, where effects of intonational breaks are also expected to emerge.

The effect of Number of Referents at the direct object ROI *orange* on the ORANGE/DISTRACTOR, observed in Experiment 1 and expected from previous literature (e.g., Spivey-Knowlton et al., 2002), was replicated in Experiment 2, but will be omitted from the results reported below, as it offers no new insights beyond what was described in Experiment 1. There was no effect of Intonational Break or Number of Referents on other target objects at this ROI.

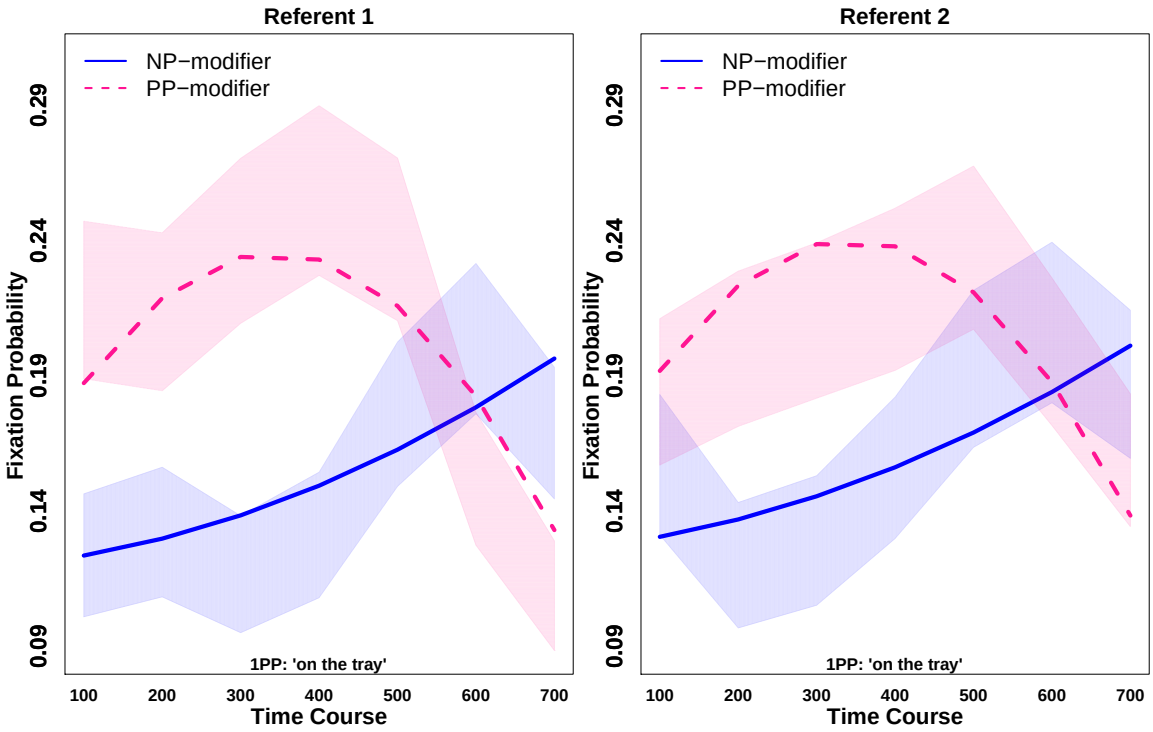

Figure 5. Experiment 2. Time course plot of fixation probability for the object TRAY IN BOWL (corresponding to the Compound Location) from 100 ms to 700 ms at ROI:1PP *on the tray*. Left panel: 1 Referent condition, right panel: 2 Referent condition. The intonation conditions (NP-modifier, PP-modifier) are marked through line types and colors. The shaded bands indicate the standard error around the observed mean. The lines represent the predicted values of the LME model reported in Table 2. Note that the offset of the region of analysis varied by items, but fixations crossing the offset were excluded, see Analysis section for details.

Results

On the object TRAY IN BOWL we expect more looks in the PP modifier intonation condition, in which a break is placed between the direct object NP and 1PP, enclosing 1PP and 2PP within the same intonational phrase *on the tray in the bowl*. This contrasts with the NP modifier condition, which should receive less looks, as the intonational breaks are now placed at the end of 1PP and the beginning of 2PP, resulting in *in the bowl* as an intonational phrase of its own. This prediction is confirmed in Figure 5, which shows the probability of fixations on the object TRAY IN BOWL at the linguistic region of interest *on the tray*.

The mixed model confirms this observation: the target object fixated significantly more in the PP modifier condition (see Table 3). Over time, looks to the target decrease, as evidenced by the interaction of Intonation with the quadratic term of time. This finding confirms previous research, where intonational breaks were found to enhance the likelihood of fixating the visual object which they are associated with (e.g., Snedeker & Trueswell, 2003).

The results on TRAY IN BOWL are corroborated by the pattern we find on ORANGE ON TRAY. Here, we expect more looks in the NP modifier intonation condition, in which there is no break

Table 3
Experiment 2. Mixed model analysis of the fixations on the object TRAY IN BOWL *at ROI:1PP* on the tray. Intonation *is coded as* -0.5 *for NP modifier and* 0.5 *for PP modifier,* Number of Referents *is coded as* -0.5 *for 1 Referent,* 0.5 *for 2 Referents. Time (100 to 700 ms, in 100 ms intervals) is represented as an orthogonal polynomial of order two (Time^1, Time^2).*

Predictor	β	*SE*	*t*	*p*
Intercept	0.178	0.010	16.891	$<$ 1e-04
Intonation	0.041	0.020	2.034	0.04
Time^2	-0.032	0.012	-2.664	0.007
Referent	0.006	0.019	0.332	0.7
Intonation:Time^2	-0.082	0.024	-3.454	0.0005

Formula: (1 | item) + (1 | participant) + Intonation + (0 + Intonation | item) + (0 + Intonation | participant) + Time^2 + (0 + Time^2 | item) + Referent + (0 + Referent | participant) + (0 + Referent | item) + Intonation:Time^2

between *the orange* and *on the tray*, compared to the PP modifier intonation. This should be true especially in the 1 Referent condition, in which there is less referential competition. We refer the reader to Supplementary Material, Experiment 2: Linguistic Prominence Effect on ROI:1PP, *on the tray* on ORANGE ON TRAY, for results on this second target object.

Discussion

In Experiment 2, we investigated how intonational breaks are used to resolve syntactic ambiguity. Intonational breaks are interruptions in the speech stream marking the boundaries of phrases, which correlate with a longer duration of the word preceding the break, and are marked by changes to pitch accents (Ferreira, 1993). We found that such boundaries boosts the visual prominence of the object enclosed within the resulting intonational phrase (e.g., a boundary around the phrase *the orange on the tray* boosts looks to the object ORANGE ON TRAY). In line with previous studies (Snedeker & Trueswell, 2003; Bailey & Ferreira, 2007; Snedeker & Yuan, 2008), we observed that differences in the placement of intonational breaks modulate the prominence of the objects that are visually attended.

This effect was especially evident during ambiguous regions, such as the ROI:1PP *on the tray*, in which the need for resolution forces visual attention to rely on intonational break information. Crucially, the effect of intonation interacted with the factor of Time, indicating that the parser incrementally integrates the referring expression enclosed by the intonational phrase.

To summarize, in Experiments 1 and 2, we observed that two forms of saliency, visual saliency and linguistic prominence, are used during the situated language understanding of ambiguous sentences. Crucially, however, their effects on the visual responses are distinct. Visual saliency is used by the processor to aid the prediction of upcoming linguistic material, and hence does not directly interact with ambiguity resolution. Linguistic prominence, in contrast, directly correlates with ambiguity resolution. As argued previously, visual saliency is predictively used to forecast upcoming arguments. Linguistic prominence instead is locally related to a linguistic referent, and has therefore a more direct application to the resolution of syntactic ambiguity.

A key theoretical issue remains: What happens when these two types of saliency are both available? They can either be consistent (give prominence to the same target object) or be incon-

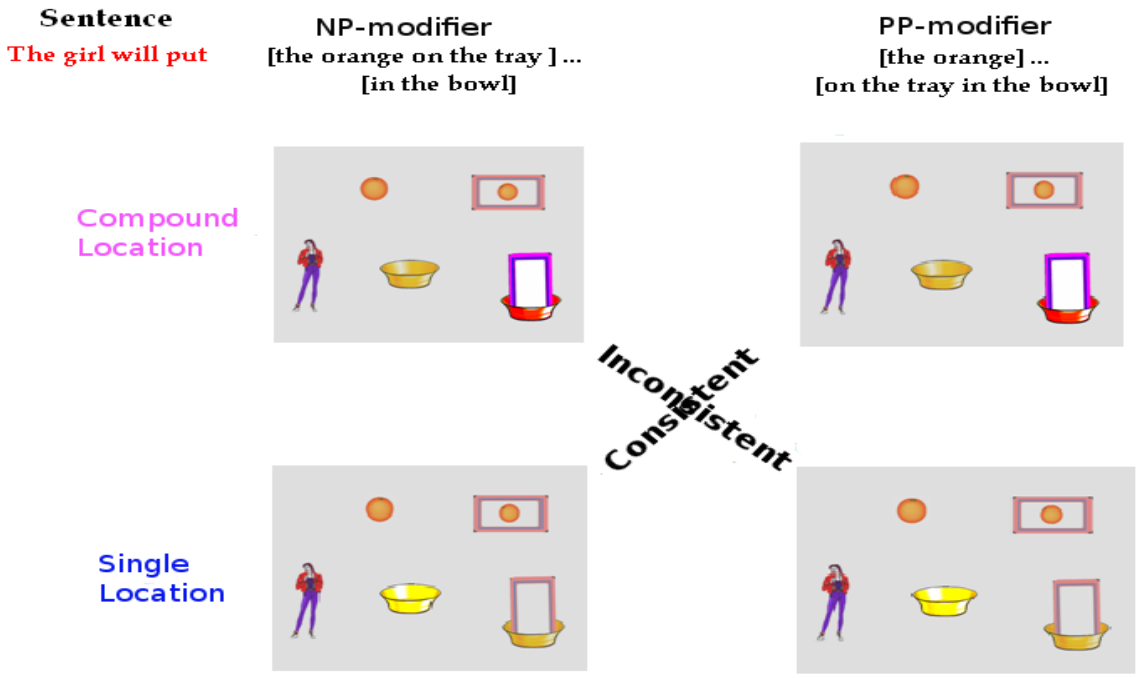

Figure 6. Example stimuli for Experiment 3. The design crosses Intonational Break (NP modifier, PP modifier) with Saliency (Single Location, Compound Location). This results in a stimulus which is either consistent (the two modalities resolve the syntactic ambiguity in the same way) or inconsistent (the two modalities resolve the ambiguity in different ways).

sistent (give prominence to the different target objects). Under an adaptive view of the language processor (see Introduction), we would expect the processor to be able to deal with inconsistent case by utilizing the two types of information at different points of time during incremental sentence comprehension, and for different sub-tasks (prediction vs. ambiguity resolution, as shown in Experiments 1 and 2). The following experiment tests this claim by manipulating visual saliency and linguistic prominence in the same design.

Experiment 3: Interaction of Visual and Linguistic Saliency

In Experiment 3, we bring together visual and linguistic saliency in the same design and investigate their pattern of interaction. Our aim is to assess how information from the two modalities is combined during sentence processing. Under an adaptive view of the language processor, we expect visual and linguistic saliency to be used to optimize separate aspects of the comprehension task. If this is the case, then we expect the present design to replicate the results of Experiment 1 and 2, in which visual saliency and intonational breaks were manipulated separately.

Method

The experimental design crossed the factors *Intonational Break* (NP modifier or PP modifier) and *Saliency* (Single Location or Compound Location). As in Experiments 1 and 2, participants listened to sentences with ambiguous PP attachments such as (1) while concurrently viewing a visual context that supported both interpretations.

Figure 6 illustrates the design of this experiment. Crossing the factors *Intonational Break* and *Saliency* results either consistently between the two modalities (the two modalities point to the same target object, resolving the ambiguity in the same way) or inconsistently (they point to different target objects, resolving the ambiguity in different ways). An example for consistency is the combination of NP modifier and Single Location, in which the phrase *in the bowl* is given prominence by the NP modifier break and it is also visually salient. An example for inconsistency is the PP modifier and Single Location. Here, the phrase enclosed by the intonational break is *on the tray in the bowl*, while the visually salient object is the BOWL, rather than the TRAY IN BOWL.

Materials, Procedure, Participants, and Analysis. The materials were reused from the previous experiments: the visual stimuli were taken from Experiment 1 and the linguistic stimuli from Experiment 2. For the visual stimuli, we retained only the 2 Referent condition, as it allows a larger number of possible interpretations. As in Experiment 2, we did not manipulate preview, but fixed it at 1000 ms. Thirty-two new participants from the same population as Experiments 1 and 2 took part in the study.

As in Experiments 1 and 2, we show plots of probabilities of fixation to the target object during a given linguistic region of interest, and then fit a mixed effects model for statistical inference. The model predictors were *Intonational Break*, *Saliency*, and *Time*, and the random effects were *Participant* and *Item*. Time was encoded as before using quadratic splines, and the same model selection procedure as described previously was used.

Results

If visual and linguistic saliency are used in a complementary fashion by the sentence processor, then we expect the patterns of results in this experiment to be a combination of the patterns observed in Experiments 1 and 2. This should be the case even when the two sources of information are inconsistent with respect to the target object they give prominence to, as the processor adaptively uses visual and linguistic saliency at different point during sentence processing, and for different sub-tasks of comprehension.

We utilize the same analyses we have presented for Experiments 1 and 2. First, we look at the linguistic ROI:NP direct object (*the orange* in our running example) and investigate looks to the visually salient objects Single Location (BOWL). In this region, we expect to replicate the anticipatory effects of visual saliency observed in Experiment 1. (For results on the Compound Location objects, refer to Supplementary Material, Experiment 3: Visual Saliency Effect on ROI:NP, *the orange* on TRAY ON BOWL).

We then move to the 1PP region, at which ambiguity is expected to emerge. At this region, we observed effects of intonational breaks in Experiment 2, but no effect of visual saliency in Experiment 1. Thus, in line with Experiment 2, we expect looks to increase on the target objects associated with the intonational phrase marked by the intonational break. The target object TRAY IN BOWL is expected to receive more looks for the PP modifier intonation, which encloses *the tray in the tray* within intonational boundaries. (Results the alternative object region ORANGE ON TRAY are reported in the Supplementary Material, Experiment 3: Linguistic Prominence Effect on ROI:1PP *on the tray* on ORANGE ON TRAY)

As in Experiments 1 and 2, we analyze the probability of fixations to the target object across experimental condition using linear mixed effects models.

Figure 7 plots the time course of fixations on the target object BOWL for the ROI:NP *the orange*. We observe an anticipatory effect triggered by visual saliency: the BOWL is looked at sig-

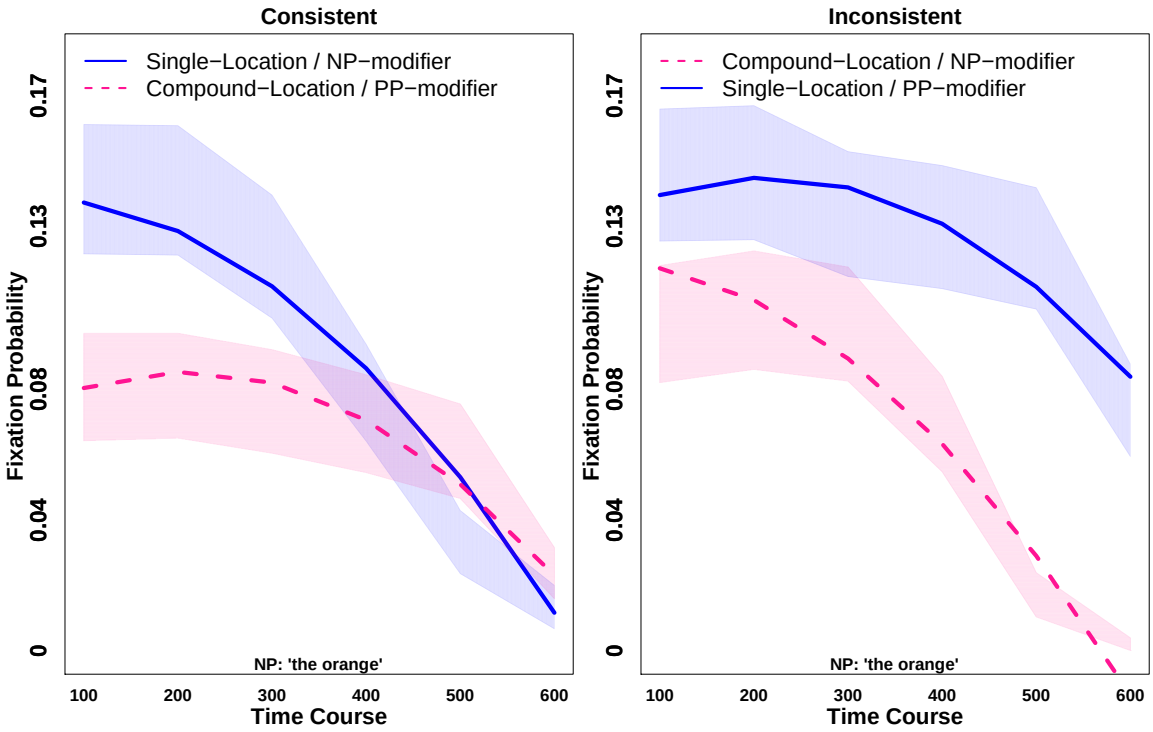

Figure 7. Experiment 3. Time course plot of fixation probability for the object BOWL (corresponding to the Single Location) from 100 ms to 600 ms at ROI:NP *the orange*. Left panel: Consistent, right panel: Inconsistent condition. The four experimental conditions are marked through line types and colors. The shaded bands indicate the standard error around the observed mean. The lines represent the predicted values of the LME model reported in Table 4. Note that the offset of the region of analysis varied by items, but fixations crossing the offset were excluded, see Analysis section for details.

nificantly more when visually salient (Single Location condition). This matches our findings from Experiment 1: compare with Figure 2 (right panel). Looks to the visually salient object decrease over time, though this trend does not reach significance, as Table 4 shows (two-way interaction Time^1:Saliency). We also find a main effect of intonation: in the PP-modifier break condition, more looks to the target object are observed. A PP-modifier break implies a longer duration of the direct object, which may make the Single Location object a likely goal location of the direct object.

Crucially, however, we found no interaction between visual saliency and intonation on this target object.

We will now turn to the results on the object TRAY IN BOWL for the region ROI:1PP *on the tray*. Here, we expect to replicate the effect of intonational breaks observed in Experiment 2. Figure 8 plots looks to this target object across the different conditions. We find a main effect of intonation which increases linearly over time (interaction Intonation:Time^1; see Table 5). The intonational break gives prominence to the visual object enclosed within the intonational phrase, and this effect becomes stronger over time. Saliency is included as it improves model fit, but its coefficient fails to reach significance. These results match closely what we found in Experiment 2, compare with Figure 5 (right panel) and Table 3. We refer the reader to Supplementary Material,

Table 4
Experiment 3. Mixed model analysis of the fixations on the object BOWL *at ROI:NP* the orange. Intonation *is coded as* −0.5 *for NP modifier and* 0.5 *for PP modifier,* Saliency *is coded as* −0.5 *for Single Location and* 0.5 *for Compound Location. Time (100 to 600 ms, in 100 ms intervals)is represented as an orthogonal polynomial of order two (Time^1, Time^2).*

Predictor	β	*SE*	*t*	*p*
Intercept	0.085	0.011	7.297	< 1e-04
Time^1	-0.076	0.014	-5.177	< 1e-04
Saliency	0.044	0.016	2.760	0.005
Time^2	-0.025	0.007	-3.321	0.0008
Intonation	0.017	0.008	2.088	0.03
Time1:Saliency	-0.02	0.015	-1.331	0.1

Formula: (1 | item) + (1 | participant) + Time^1 + (0 + Time^1 | item) + Saliency + (0 + Saliency | item) + (0 + Saliency | participant) + Time^2 + Intonation + (0 + Intonation | item) + (0 + Intonation | participant) + Time^1:Saliency

Table 5
Experiment 3. Mixed model analysis of the fixations on the object TRAY IN BOWL *at ROI:1PP* on the tray. Intonation *is coded as* −0.5 *for NP modifier and* 0.5 *for PP modifier,* Saliency *is coded as* −0.5 *for Single Location and* 0.5 *for Compound Location. Time (100 to 700 ms, in 100 ms intervals)is represented as an orthogonal polynomial of order two (Time^1, Time^2).*

Predictor	β	*SE*	*t*	*p*
Intercept	0.174	0.014	11.917	<1e-04
Time^2	-0.069	0.012	-5.353	<1e-04
Intonation	0.041	0.018	2.218	0.02
Time^1	-0.046	0.022	-2.044	0.04
Saliency	-0.006	0.016	-0.406	0.6
Intonation:Time^1	-0.075	0.019	-3.959	<1e-04
Time^2:Intonation	-0.052	0.019	-2.741	0.006

Formula: (1 | participant) + (1 | item) + Time^2 + (0 + Time^2 | item) + Intonation + (0 + Intonation | participant) + (0 + Intonation | item) + Time^1 + (0 + Time^1 | item) + (0 + Time^1 | participant) + Saliency + (0 + Saliency | item) + (0 + Saliency | participant) + Intonation:Time^1 + Time^2:Intonation

Experiment 3: Linguistic Prominence Effect on ROI:1PP *on the tray* on ORANGE ON TRAY, for corroborating results of on the target object ORANGE ON TRAY.

Discussion

In Experiment 1, we observed that at the direct object region (*the orange* in our example), saliency is utilized to predict upcoming linguistic information. In Experiment 2, in line with previous studies (e.g., Snedeker & Trueswell, 2003), we observed that intonational breaks give prominence to the referent enclosed within the intonational phrase, hence triggering more looks to the corresponding visual object.

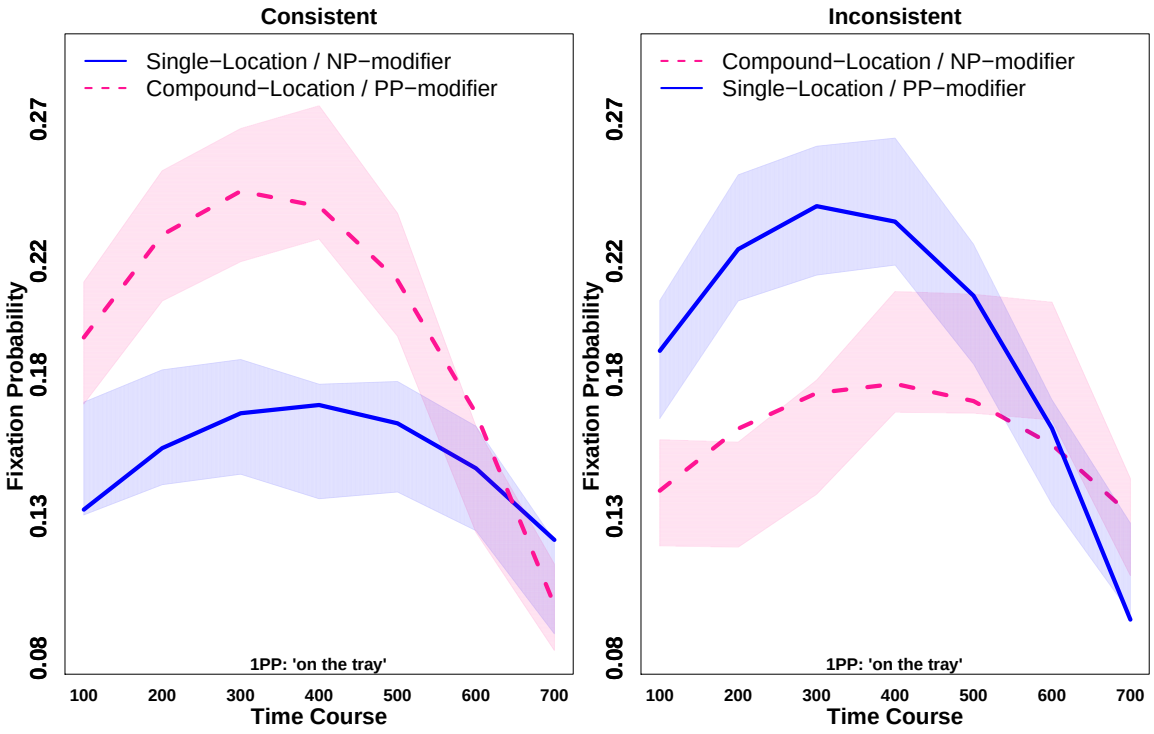

Figure 8. Experiment 3. Time course plot of fixation probability for the object TRAY IN BOWL (corresponding to the Compound Location) from 100 ms to 700 ms at ROI:1PP *on the tray*. Left panel: Consistent, right panel: Inconsistent condition. The four experimental conditions are marked through line types and colors. The shaded bands indicate the standard error around the observed mean. The lines represent the predicted values of the LME model reported in Table 5. Note that the offset of the region of analysis varied by items, but fixations crossing the offset were excluded, see Analysis section for details.

In Experiment 3, we looked at the interaction between visual and linguistic saliency, with the two forms of prominence either consistent or inconsistent with each other. We obtained results that closely matched what we observed in Experiments 1 and 2, when we looked at the two forms of saliency separately. Visual saliency triggered anticipatory effects during the processing of the verb, such that a salient object is attended to more, because it is expected to appear in as one of the arguments of the verb. Intonational breaks, on the other hand, modulate the mapping between visual and linguistic referents. They give prominence to the referring expression enclosed within the intonational phrase, which is then attended more in the visual context.

Crucially, we did not find significant interactions between visual saliency and intonation. These results suggest that the two types of information do not influence each other during sentence understanding. Visual and linguistic information both constrain sentence processing, with the interpretation of the sentence being revised incrementally as new information becomes available. The two modalities therefore are used at different points during the incremental processing of a sentence, and they are used for different sub-tasks (argument prediction vs. ambiguity resolution).

General Discussion

Language understanding often occurs synchronously with other modalities, e.g., vision, which raises the question of how information is integrated across modalities, and more specifically how the sentence processor utilizes cross-modal input when performing tasks such as ambiguity resolution.

Previous work on sentence processing situated in a visual context (e.g., Cooper, 1974; Tanenhaus et al., 1995; Altmann & Kamide, 1999) has shown that visual responses are influenced both by linguistic information and by properties of the depicted objects, i.e., by the visual referents that form the context. The interplay between the two modalities encompasses different levels of processing, such as intonational information (e.g., Snedeker & Yuan, 2008) or lexical semantics (e.g., Huettig & Altmann, 2005) and is bi-directional in nature: visual information mediates sentence processing (e.g, Gleitman et al., 2007), but sentence processing also constraints the allocation of visual attention (e.g., Kukona et al., 2011).

More recent work has focused on the interaction of different types of information (visual and linguistic). Vogels et al.'s (2012) study, for example, investigated how the visual saliency of a referent (background vs. foreground), and its prior linguistic context, influence how participants refer to that referent. Vogels et al. found that both linguistic and visual saliency influence the referring expressions produced, but there was no interaction between the two. This finding is consistent with results by Ferreira et al. (2013) on the role of preview and visual complexity in situated language understanding. Ferreira et al.'s study suggest that the human language processor, rather than using a fixed processing strategy, is adaptive, and can use visual and linguistic information as and when it becomes available, tailoring its strategy to the current task and the processing stages within that task.

The present study tested this adaptive view of situated language processing by investigating what happens when linguistic and visual information are concurrently available, but are inconsistent with each other, e.g., pointing to different interpretation for an ambiguous sentence in language processing. In such a setting, the adaptive view predicts that information from the two modalities is used at different point in the incremental processing of a sentence, and for different sub-tasks of comprehension.

We used a classic syntactic attachment ambiguity paradigm, which has been extensively used in psycholinguistics to uncover mechanisms of linguistic ambiguity resolution by studying attention in visual scenes (Spivey-Knowlton et al., 2002; Snedeker & Trueswell, 2003; Farmer, Anderson, & Spivey, 2007; Novick et al., 2008; Ferreira et al., 2013). In such a setup, visual and linguistic information can point either to the same or to a different regions in the display, hence promoting either the same or a different resolution of the ambiguity. We first investigated the mono-modal case to establish how visual and linguistic information are used for ambiguity resolution when information of only one the two modalities is experimentally manipulated (Experiments 1 and 2). We then looked at the cross-modal case to investigate what happens when information from both modalities is concurrently available and is either consistent or inconsistent in terms of which interpretation is favored by each modality (Experiment 3).

All three experiments manipulated saliency, i.e., low-level properties of the stimulus. Visual saliency was operationalized in terms of low-level image features such as color, intensity, and orientation, which can give prominence to regions in a visual display (Itti & Koch, 2000). To operationalize linguistic saliency, we used intonational breaks. Intonational breaks can be considered

a form of linguistic saliency, as they do not carry explicit semantic information, and are used to enhance the prominence of the referring expression enclosed by the breaks (Snedeker & Trueswell, 2003).

Experiment 1 showed that visual saliency is utilized at the beginning of the direct object of a sentence to predict upcoming linguistic information. This finding suggests that situated sentence understanding is divided into two main phases: a free-viewing phase occurring during the preview of the visual context, and a more task-oriented phase when the sentence is incrementally understood. In particular, the more linguistic input is parsed, the more visual attention is constrained to the referents being mentioned, which then become search targets. This explains why we find effects of visual saliency after the verb (at the beginning of the direct object): saliency does not directly affect ambiguity resolution, but instead acts upon the selection of possible search targets, which manifests itself as argument prediction in our results. This also confirms results that suggest that both top-down and bottom-up mechanisms play a role in visual guidance (see Coco et al., 2014). Note that the effect of visual saliency was not driven by non-language related shifts of visual attention, as we observe purely visual effects of saliency at the scene onset and these rapidly decay prior to the onset of the sentence (see Appendix for details). Instead, we found that anticipatory looks were related to the beginning of the speech stream, and especially confined to a precise segment, i.e., the onset of the direct object, which indicates a selective use of visual saliency by the sentence processor. By the end of the direct object region, in fact, we observed a visually salient object to receive the same amount of fixation as its non-salient version.

Experiment 2 investigated the effect of intonational breaks on ambiguity resolution using the same materials as in Experiment 1. We found that a visual object that corresponds to a referring expression enclosed by intonational breaks is looked at more compared to a condition in which the intonational breaks induce a different mapping between linguistic and visual referents. The results of our Experiment 2 extend previous findings to another type of syntactic construction: Snedeker and Trueswell (2003) and subsequently Snedeker and Yuan (2008) investigated instrument modifiers, e.g., *tap the frog with the flower*, whereas our study focused on goal modifiers. The fact that we obtained comparable results suggest that the function of intonational breaks is fairly broad, and applies to a range of syntactic contexts. The use of breaks relates, more generally, to the introduction of new linguistic information into the discourse, and to the consolidation of previous information with the available context. Thus, a referring expression preceding the break becomes less likely to be modified by subsequent material, whereas a referring expression following the break is more likely to be a new contextual entity. More research testing different syntactic context is needed to assess the generality of this claim.

Finally, in Experiment 3, we tested visual and linguistic saliency (intonational breaks) together using an experimental design in which the two forms of saliency are either consistent, i.e., they both point at the same target object, or inconsistent, i.e., they point at different target objects. We replicated the pattern of results observed in Experiment 1 (visual saliency is predictively used) and Experiment 2 (intonational breaks give visual prominence to the enclosed referent and thus aid ambiguity resolution). Crucially, we did not find any significant interactions between visual saliency and intonational breaks in Experiment 3. This result supports an adaptive view of cross-modal processing, in which visual and linguistic information are used when needed during sentence comprehension. The two modalities play complementary roles: the visual saliency of targets is used to suggest upcoming arguments in the linguistic stream before they occur, whereas linguistic saliency gives prominence to the referents enclosed by breaks.

Our results corroborate the findings of Vogels et al. (2012), who studied language production rather than comprehension and found that the visual saliency of a referent, and the contextual expectations that relate to it, both influence the rate and type of referring expression produced. Our results are expected under the traditional constraint-satisfaction view of language processing, where constraints are simultaneously evaluated to optimize the interpretation at the current point in time (e.g., MacDonald et al. (1994)). The results are also compatible with more recent situated understanding constraint-satisfaction accounts, which postulate locality, and temporal dependence of constraint activation (e.g, Knoeferle and Crocker (2006); Kukona et al. (2011)). Such accounts assume that visual and linguistic information are used for active forecasting; and in line with Ferreira et al. (2013), they assume that the use of cross-modal information is adapted to the current state of processing. Adaptive processing emerges by using available resources in a complementary fashion for different sub-tasks of language comprehension: visual saliency of an object is used to predict possible arguments, while intonational breaks are used to resolve the referential ambiguity by giving prominence to one of the candidate objects.

From a modeling perspective, our results suggest that saliency, in its linguistic, and especially in its visual form, forms an integral part of situated sentence processing, and needs to be accounted for by computational models of this process. This is not the case for existing models such as the Coordinated Interplay Model (Crocker et al., 2010) or the Impulse Processing Model (Kukona & Tabor, 2011), which both work on linguistic input in the form of transcribed speech combined with a symbolic representation of the visual context. This means that they have no way of representing either linguistic or visual saliency; how the models (and their input representations) can be adapted to account for results such as the ones presented in this study is a subject for future research.

To conclude, the present set of results imposes constraints on theories of situated language processing. We found that visual and linguistic information are brought to bear at different points during comprehension. They complement each other, and eventually are integrated into an overall interpretation of the sentence. This is the case even if the information provided by the two modalities is inconsistent, giving prominence to different parts of the sentence. We can therefore rule out a fixed architecture in which the processing on one modality takes precedence over the processing in the other modality. Rather, our results suggest an adaptive architecture in which the two modalities are accessed by the sentence processor for different sub-tasks, and called upon when needed to update the current analysis of the input.

Supplementary Material

This paper is accompanied by the series of additional visualizations and analyses which can be found it the Supplementary Material. Here, we briefly summarize its content.

In *Decay of Visual Saliency Effect during Preview*, we demonstrate that purely low-level effects of visual saliency are observed immediately after the scene onset, and decay during the preview time prior to speech onset.

In *Experiment 1: Referentiality Effect at ROI:NP* the orange *on Object* ORANGE, we successfully replicate the classic referentiality effect first shown by Tanenhaus et al. (1995).

In *Experiment 2: Linguistic Prominence Effect on ROI:1PP* on the tray*s on Object* ORANGE ON TRAY, we show corroborating result of intonational breaks on the other target object to which the manipulation refers.

In *Experiment 3: Visual Saliency Effect on ROI:NP* the orange *on Object* TRAY IN BOWL, we show corroborating results of visual saliency on the other target object to which the manipulation

refers.

In *Experiment 3: Linguistic Prominence Effect on ROI:1PP* on the tray *on Object* ORANGE ON TRAY, we show corroborating results of intonational breaks on the other target object to which the manipulation refer.

In *Comparison Between Linear Mixed Effects Models Aggregated by Trials and by Participants*, we show that a linear-mixed effects model with participants and items both as random effects returns the same coefficient estimates as a model with fixation probability aggregated by participants while having a lower Type 1 error rate.

In *Time-course Plots across the Whole Sentence for Single-Location, Compound-Location and Other objects*, we visualize the fixation proportions to the objects experimentally investigated, and to the other objects available in the display, across the whole sentence.

References

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

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

Altmann, G., & Mirkovic, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*, 583–609.

Arai, M., & Keller, F. (2013). The use of verb-specific information for prediction in sentence processing. *Language and Cognitive Processes*, *28*, 525–560.

Arai, M., van Gompel, R., & Scheepers, C. (2007). Priming ditranstitive structures in comprehension. *Cognitive Psychology*, *54*(3), 218–250.

Ariel, M. (1990). *Accessing noun-phrase antecedents.* London: Routledge.

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.

Bailey, K., & Ferreira, F. (2007). The processing of filled pause disfluencies in the visual world. In *R. Van Gompel & M. Fisher & R. Hill & W. Murray (Eds.), Eye movements: a window on mind and brain.* Oxford: Elsevier.

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

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random-effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Boersma, P., & Weenink, D. (2013). *Praat: doing phonetics by computer [computer program]. version 5.3.39.* `http://www.praat.org/`.

Borji, A., Sihite, D., & Itti, L. (2013). What stands out in a scene? a study of human explicit saliency judgment. *Vision research*, *91*, 62–77.

Bradlow, A., Clopper, C., Smiljanic, R., & Walter, M. (2010). A perceptual phonetic similarity space for languages: evidence from five native language listeners groups. *Speech Communication*, *52*, 920-942.

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 687–696.

Coco, M., & Keller, F. (2009). The impact of visual information on referent assignment in sentence production. In *In N.A. Taatgen and H. van Rijn (Eds.), Proceedings of the 31th Annual Conference of the Cognitive Science Society.* Amsterdam: Cognitive Science Society.

Coco, M., Malcolm, G., & Keller, F. (2014). The interplay of bottom-up and top-down mechanisms in visual guidance during object naming. *Quarterly Journal of Experimental Psychology*, *0*(0), 0.

Cooper, R. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 189–201.

Crocker, M., Knoeferle, P., & Mayberry, M. (2010). Situated sentence comprehension: The coordinated interplay account and a neurobehavioral model. *Brain and Language*, *112*(3), 189–201.

Dahan, D., & Tanenhaus, M. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychological Bulletin & Review*, *12*, 455–459.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*(2).

Farmer, T., Anderson, S., & Spivey, M. (2007). Gradiency and visual context in syntactic garden-paths. *Journal of memory and language*, *57*(4), 570–595.

Ferreira, F. (1993). The creation of prosody during sentence production. *Psychological Review*, *100*, 233–253.

Ferreira, F., Foucart, A., & Engelhardt, P. (2013). Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of Memory and Language*, *69*(3), 165–182.

Fukumura, K., van Gompel, R., & Pickering, M. (2010). The use of visual context during the production of referring expressions. *The quarterly journal of experimental psychology*, *63*(9), 1700–1715.

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

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*.

Huettig, F., & Altmann, G. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition*, *96*, B23–B32.

Huettig, F., & Altmann, G. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. *Visual Cognition*, *15*, 985–1018.

Huettig, F., & Altmann, G. (2011). Looking at anything that is green when hearing ÂŚfrogÂŠ: How object surface colour and stored object colour knowledge influence language-mediated overt attention. *Quarterly Journal of Experimental Psychology*, *64*, 122–145.

Ito, K., & Speer, S. R. (2008). Anticipatory effect of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, *58*, 541–573.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10-12), 1489–1506.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford: Oxford University Press.

Kaiser, E., Runner, J., Sussman, R. S., & Tanenhaus, M. (2009). Structural and semantic constraints on the resolution of pronouns and reflexives. *Cognition*, *112*(1), 55–80.

Kamide, Y., Altmann, G., & Haywood, S. (2003). Prediction and thematic information in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*, 133–156.

Knoeferle, P., & Crocker, M. (2006). The coordinated interplay of scene, utterance and world knowledge. *Cognitive Science*, *30*, 481–529.

Kukona, A., Fang, S., Aichera, K., Chen, H., & Magnuson, J. (2011). The time course of anticipatory constraint integration. *Cognition*, *119*, 23–42.

Kukona, A., & Tabor, W. (2011). Impulse processing: A dynamical systems model of incremental eye movements in the visual world paradigm. *Cognitive Science*, *35*(6), 1009âĂŞ1051.

MacDonald, M., Pearlmutter, N., & Seidenberg, M. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological review*, *101*(4), 676.

Mirman, D., Dixon, J., & Magnuson, J. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–

494.

Mirman, D., & Magnuson, J. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Memory & Cognition*, *37*(7), 1026-1039.

Novick, J., Thompson-Schill, S., & Trueswell, J. (2008). Putting lexical constraints in context into the visual-world paradigm. *Cognition*(107), 850–903.

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

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in s and s-plus.* New York: Springer-Verlag.

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

Snedeker, J., & Trueswell, J. C. (2003). Using prosody to avoid ambiguity: effects of speaker awareness and referential context. *Journal of Memory and Language*, *48*, 103–130.

Snedeker, J., & Yuan, S. (2008). Effects of prosodic and lexical constraints on parsing in young children (and adults). *Journal of Memory and Language*(58), 574-608.

Spivey-Knowlton, M., Tanenhaus, M., Eberhard, K., & Sedivy, J. (2002). Eye movements and spoken language comprehension: Effects of syntactic context on syntactic ambiguity resolution. *Cognitive Psychology*(45), 447–481.

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

Tatler, B., Baddeley, R., & Gilchrist, I. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision research*, *45*(5), 643–659.

Vo, M., & 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.

Vogels, J., Krahmer, E., & Maes, A. (2012). Who is where referred to how, and why? the influence of visual saliency on referent accessibility in spoken language production. *Language and Cognitive Processes*(ahead-of-print), 1–27.

Walther, D., & Koch, D. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*, 1395–1407.