Journal of Memory and Language 69 (2013) 165–182

Contents lists available at SciVerse ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Language processing in the visual world: Effects of preview, visual complexity, and prediction

Fernanda Ferreira [a,*], Alice Foucart [b], Paul E. Engelhardt [c]

[a] Department of Psychology and Institute for Mind and Brain, University of South Carolina, Columbia, SC 29201, USA
[b] Department of Technology, Universitat Pompeu Fabra, Barcelona, Spain
[c] School of Psychology, University of East Anglia, Norwich NR4 7TJ, United Kingdom

ARTICLE INFO

Article history:
Received 17 April 2012
revision received 31 May 2013
Available online 4 July 2013

Keywords:
Language-vision interactions
Processing strategies
Learning
Syntactic ambiguity resolution
Eye movements

ABSTRACT

This study investigates how people interpret spoken sentences in the context of a relevant visual world by focusing on garden-path sentences, such as *Put the book on the chair in the bucket*, in which the prepositional phrase *on the chair* is temporarily ambiguous between a goal and modifier interpretation. In three comprehension experiments, listeners heard these types of sentences (along with disambiguated controls) while viewing arrays of objects. These experiments demonstrate that a classic garden-path effect is obtained only when listeners have a preview of the display and when the visual context contains relatively few objects. Results from a production experiment suggest that listeners accrue knowledge that may allow them to have certain expectations of the upcoming utterance based on visual information. Taken together, these findings have theoretical implications for both the role of prediction as an adaptive comprehension strategy, and for how comprehension tendencies change under variable visual and temporal processing demands.

Introduction

One of the most influential findings in the field of psycholinguistics over the last 20 years is that listeners presented with a garden-path sentence in the presence of relevant visual context tend to use the visual information to constrain their linguistic interpretations and avoid a syntactic misanalysis (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). For example, consider the imperative sentence *Put the apple on the towel in the box*. At the point at which the listener hears *on the towel*, two interpretations are possible: Either *on the towel* is the location to which the apple should be moved, or it is a modifier of *apple*. The phrase *into the box* forces the latter interpretation because it is unambiguously a location. Referential Theory (Altmann & Steedman, 1988) specifies that speakers should provide modifiers only when modification is necessary to establish reference (e.g., we do not generally refer to a **big** car if only one car is discourse-relevant). From Referential Theory, it follows that if two apples are present in the visual world and one of them is supposed to be moved, then right from the earliest stages of processing the phrase *on the towel* will be taken to be a modifier, because the modifier allows a unique apple to be picked out. The listener faced with this visual world containing two referents should therefore immediately interpret the phrase as a modifier and avoid being garden-pathed, and this is indeed what the data seem to show (Farmer, Cargill, & Spivey, 2007b; Novick, Thompson-Schill, & Trueswell, 2008; Spivey, Tanenhaus, Eberhard, & Sedivy, 2002; Tanenhaus et al., 1995; Trueswell, Sekerina, Hill, & Logrip, 1999).

This result has led to a large body of research in which researchers make use of what is now referred to as the Visual World Paradigm (VWP) (for a detailed review of the VWP, see Huettig, Rommers, & Meyer, 2011). In the VWP, participants listen to sentences, while at the same time viewing visually relevant displays. Eye movement behavior is treated as a dependent measure for evaluating

[*] Corresponding author. Address: Department of Psychology, University of South Carolina, Columbia, SC 29208, United States.
E-mail address: fernanda@sc.edu (F. Ferreira).

0749-596X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jml.2013.06.001

hypotheses about the kinds of interpretations that are built and the timing of their activation. For example, if a listener hears *put the apple*... in the context of a set of objects including an apple, he or she is likely to make an eye movement towards the mentioned apple. The linking hypothesis is that lexical activation causes a shift of attention towards the object represented by that word, which in turn triggers a saccade to the object (Allopenna, Magnuson, & Tanenhaus, 1998; Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Huettig et al., 2011; Tanenhaus, Spivey-Knowlton, & Hanna, 2000). The situation becomes more interesting when there is some type of linguistic ambiguity, because the pattern of eye movements indicates which sources of information are used to disambiguate the reference. For example, Chambers, Tanenhaus, and Magnuson (2004) examined utterances such as *pour the egg*... heard in the context of a visual world containing both liquid and solid eggs. They observed that participants were able to use the information about the affordances of the objects to immediately constrain their interpretations – in this case, they tended to look at the liquid egg rather than the solid egg when they heard the verb *pour*.

In the VWP experiments originally designed to examine the processing of syntactic ambiguity (e.g. Spivey et al., 2002; Tanenhaus et al., 1995), participants were presented with a 2 × 2 arrangement of real objects (not photos or images) to be manipulated in response to auditory instructions. Two quadrants contained the target and the distractor object and were the objects moved first (Engelhardt, Bailey, & Ferreira, 2006; Spivey et al., 2002). The other two quadrants contained two potential goal locations. Participants then received either a syntactically ambiguous or unambiguous instruction containing a prepositional phrase modifier. The critical finding was that when participants heard an utterance, such as *Put the apple on the towel in the box* in the context of a display containing a single apple, they tended to look at the incorrect goal (i.e. the empty towel) a few hundred milliseconds after hearing the first prepositional phrase. These fixations are interpreted as evidence that participants momentarily considered the goal analysis of *on the towel.* But when the display contained two apples (the "two-referent" condition), participants almost never looked at the empty towel; instead, they looked at the two apples and then they looked directly to the correct goal (i.e. *the box*). This fixation pattern has been taken as evidence that the visual context (i.e., the presence of two apples and the consequent need for modification) can be immediately used to resolve the temporary ambiguity, and is assumed to be evidence for an interactive processing architecture in which visual context informs syntactic decision-making mechanisms (MacDonald, Pearlmutter, & Seidenberg, 1994; MacDonald & Seidenberg, 2006).

The findings from these VWP experiments (e.g. Spivey et al., 2002) highlight a broader set of theoretical issues concerning the interaction of two cognitive systems: the visual system and the language comprehension system (Huettig et al., 2011). In other words, the VWP can be viewed as more than a tool for studying how linguistic ambiguities are resolved, and indeed, the use of the paradigm presupposes some understanding of the interface between vision and language (Ferreira & Tanenhaus, 2007; Henderson & Ferreira, 2004). The context effects that occur in the presence of a visual world are potentially different from those that have been studied previously using manipulations of discourse (or linguistic) context (e.g., Altmann & Steedman, 1988; Ferreira & Clifton, 1986; Trueswell, Tanenhaus, & Garnsey, 1994). Typically, when linguistic context is the focus of study, the context is presented first and is fully processed before the critical sentence is encountered. For example, in the Ferreira and Clifton (1986) experiments, participants read a set of sentences that established the presence of certain discourse entities. That context was presumed to then influence processing of the sentence immediately following, which was either a garden-path sentence or some type of control. The context and the linguistic ambiguity were thus processed **sequentially**. Of course, readers can and occasionally do re-read the context, but generally text is read from top to bottom and left to right, so typically, linguistic contexts intended to bias the interpretation of a critical sentence will be processed first.

In contrast, in the VWP, the context (which is visual rather than linguistic) is available at the same time as the critical sentence, and the ability to process that visual context prior to the utterance is not always controlled or manipulated. In the original VWP studies (e.g. Spivey et al., 2002), the participant was allowed to watch as the experimenter placed the objects for the upcoming trial in their respective positions. Thus, the amount of time available to preview the visual context could be many seconds, and the time interval varied from one trial to the next. Then, once the utterance begins, the visual world and the linguistic material are co-present: The visual context remains visible while the auditory sentence is heard. Unfortunately, little attention has been paid thus far to issues related to the timing of this information and its potential effects on processing. For example, during the preview period, what information do participants extract before hearing the sentence, and how might it potentially guide their expectations about the upcoming utterance? Is the preview period important, or is it just a by-product of the way the paradigm has been implemented when real-world objects are used? Given that the gist of a scene is typically available within a hundred milliseconds (e.g. Castelhano & Henderson, 2008) and given that the visual system can extract information about object identities in as little as 120 ms (Kirchner & Thorpe, 2006), it seems possible that preview is not a prerequisite for context effects, suggesting that the VWP might generalize to a range of situations in which people use language in visual contexts – that is, to situations in which people have fully processed a scene before encountering sentences about it, and also to situations in which people must simultaneously process both the linguistic content and the visual world, and also to situations in which the visual world is dynamic, so that objects move, appear, disappear, etc. In addition, the eye movements that are made to mentioned objects may have different functions when a context is established early compared to cases in which it is extracted at the same time as the linguistic content. These are all largely unexplored questions.

It is also important to appreciate that both the linguistic information and the visual contexts in these experiments are highly constrained across experimental and filler trials. In syntactic ambiguity studies, listeners typically view only four objects, two of which are likely moveable (e.g., an apple and a crayon) and two of which can be treated as locations or goals (e.g., a towel and a box). In addition, most or all of the utterances are imperatives consisting of a transitive verb, a noun phrase, and at least one prepositional phrase. Thus, after experience with some trials, the participant may learn that his or her task is to figure out which of the four objects will be moved, and which one will likely be the goal. Given the affordances of the objects in the display (Chambers et al., 2004), the possibilities are fairly constrained. It is therefore, plausible that the preview period included in many VWP studies gives the listener time to encode the visual information and then use it to generate expectations about the form and content of the upcoming sentence. Both the visual display and the sentences conform to predictable patterns, which participants can potentially learn after a number of trials (Jaeger, 2010). Object names, affordances, and syntactic patterns have been shown to accrue over the course of an experiment (Farmer, Fine, & Jaeger, 2011).

The sequence of events for a participant in a VWP study with multi-second preview might proceed along these lines: (1) Look at the display and identify the likely moveable objects and goals given the objects' affordances. This process might include accessing the names of the objects (Meyer, Belke, Telling, & Humphreys, 2007; Meyer & Damian, 2007; Morsella & Miozzo, 2002; Navarrete & Costa, 2005), as well as their locations. (2) Retrieve a likely syntactic frame. Given the nature of these experiments, it would be something like null subject – transitive verb – noun phrase – prepositional phrase – possibly a second prepositional phrase. (3) Map the visual display to the syntactic frame—associate moveable objects with the direct object position, and goals with the prepositional phrase(s). (4) Compare the input to the predicted utterance, revising and editing as necessary. Finally, (5) execute the action. Of course, these steps are likely to be executed in a cascade-type process; for example, steps (2)–(4) do not have to be completed before the participant begins to execute the action (step (5)). And as participants are doing all this, they are making eye movements to the objects, which reflect their understanding of what action they should perform, and which will be highly influenced by their expectations. Indeed, the absence of a garden path in the two-referent condition is a type of expectation of the linguistic input: The comprehender expects a modifier, and therefore, rarely makes an eye movement to the incorrect goal (for an alternative explanation, see Novick et al., 2008).

It is important to note that, since the publication of work investigating the prepositional phrase attachment ambiguity, some VWP studies have used more complex visual worlds than the ones discussed thus far. However, none of these has explored how complex visual worlds are processed so that they can influence the online resolution of syntactic ambiguity. Instead, most have focused on how conversational partners generate expectations about what object is or will soon be mentioned in a dialogue. For example, Hanna and Tanenhaus (2004) showed that listeners who acted as a cook's helper used knowledge about the cook's pragmatic constraints to narrow their interpretation of what object was being referred to. The visual world consisted of about ten real world objects, and the relevant affordances changed from trial to trial. Similarly, Brown-Schmidt, Campana, and Tanenhaus (2005) had four pairs of participants interact with 56 different objects. Again, the aim was to see how conversational partners collaborated to establish reference, and syntactic ambiguity was not manipulated or tested (see also Brown-Schmidt & Tanenhaus, 2008).

In addition, as mentioned previously, these earlier studies used real objects rather than computer displays, which makes it difficult for the experimenter to control how long the visual information is present before the linguistic information is heard and processed (Farmer et al., 2007b). In a recent study, Andersson, Ferreira, and Henderson (2011) used computer displays to examine the processing of spoken sentences referring to objects in complex real-world scenes. For example, subjects viewed a typically cluttered garage interior and at the same time heard a context-establishing sentence and then either *I like the old and dust-covered* ***sailboat****, the* ***plane****, the* ***sombrero****, and the* ***uniform*** *that's surprisingly mint* or *I like the* ***sailboat*** *that's old and dust-covered, the* ***plane****, the* ***sombrero****, and the surprisingly mint* ***uniform***. No scene preview was provided. The dependent measure was saccades to each of the mentioned objects (in boldface) located in the scene. The first version places the object modifiers in the sentence in such a way that mentioned objects are close together in the utterance; the second switches the modifier types so the first and second as well as the third and fourth objects are more linguistically separated. Eye movement patterns revealed that whereas the first and last of the four named objects had about an 85% chance of being fixated over a 22-s time window, the probability of fixation in that same time window for the middle two objects was about 10% lower. Moreover, in a 5-s time window starting at word onset, the middle two objects were much less likely than the other two objects to be fixated at all, and this tendency was exaggerated when the object names were bunched together in the sentence rather than spread out. These results suggest that, in scenes containing very large numbers of objects, some objects are fixated only after a few seconds have passed, and objects mentioned in the middle of utterances might not get fixated at all.

In the current study, we focused on the integration of visual and linguistic information, and specifically, we asked how the timing and complexity of visual information affects language comprehension. The first hypothesis focused on the role of preview and whether preview is critical for the garden path and non-garden path effects established in previous studies (e.g. Spivey et al., 2002). More specifically, we hypothesized that preview allows listeners to generate certain expectations (or predictions) concerning upcoming linguistic information (Experiments 1–3). The second hypothesis was that if there are many objects, rather than just four or five, then even with preview, there will be too many possibilities concerning which objects are moveable and which objects are goals to allow useful expectations to be generated (Experiment 4). We examined these hypotheses in four experiments. In the first two, participants saw VWP displays with four objects and they heard sentences that were either syntactically ambiguous or unambiguous. The first experiment included preview, and the second did not. The third experiment used a production paradigm to assess whether listeners' acquire knowledge needed to generate expectations based on prior experience and the configuration of objects typical of VWP studies. The final experiment tested comprehension with preview, but expanded the number of objects to assess the effect of visual complexity on syntactic ambiguity resolution.

Experiment 1

Our first step was to attempt to replicate the results reported previously in the literature in which participants viewed a simple visual world and received a few seconds of preview. Recall that, in previous studies, when participants heard *put the apple on the towel in the box* in the context of a display containing an apple on a towel, another apple, an empty towel, and an empty box, participants rarely fixated the empty towel (e.g. Tanenhaus et al., 1995). However, when the single apple was replaced with, for example, a crayon, participants often fixated the empty towel shortly after hearing the ambiguous prepositional phrase. On the other hand, if participants were given an unambiguous instruction (i.e. *put the apple that's on the towel in the box),* then they rarely looked at the empty towel, and the number of referents had no effect. Thus, an increased probability of fixating the incorrect goal (e.g. the empty towel) in the one-referent, ambiguous condition is predicted.

Our first experiment was similar to previous studies: Participants were given 3 s to preview the objects prior to the onset of the auditory instruction, and each visual display contained four objects. An example two-referent visual display is shown in Fig. 1; the corresponding one-referent display was similar except that the lone book was replaced with an unrelated object (e.g. a football). Example instructions for this display are given in (1) and (2). If we observe the same interaction in looks to the incorrect goal as has been reported in previous studies, then it would suggest that we can replicate the original findings using a computerized version of the VWP, which allows more precise control over the timing of trial events.

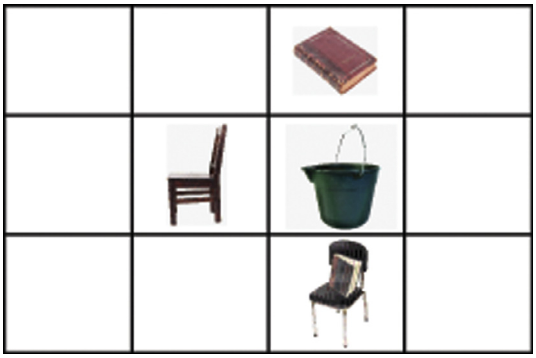

Fig. 1. Example display for the two-referent condition in Experiments 1 and 2. Grid lines were not shown in the experiments).

(1) Put the book on the chair in the bucket.
(2) Put the book that's on the chair in the bucket.

Method

Participants

Thirty-two students from the University of Edinburgh participated in the experiment. Participants were native speakers of British English, and had normal or corrected-to-normal vision. Participants were recruited through the University of Edinburgh employment service, and were paid £3.00.

Materials

We created a total of 120 visual displays: 24 were experimental items, and 96 were fillers. (Seven practice displays were also created.) Each display consisted of four or five common objects which were taken from the Hemera Photo Objects database (displays contained five objects when one of the objects was located on or in another, such as a book on a chair). All images were full color and averaged 700 × 700 pixels. The visual array on average subtended 22° degrees of visual angle horizontally and 17° vertically, for a viewing distance of 90 cm.

On all 24 of the experimental trials and on 32 of the filler trials, one of the objects in the array was a compound object; that is, it consisted of two spatially related objects (e.g. *a book on a chair*). This was the object that had to be moved to another location in the display on critical trials. For each of the 24 experimental displays, both one- and two-referent versions were created. A counterbalancing procedure was used to ensure that target objects had an equal probability of appearing in each grid location. Therefore, a target object appeared twice in each grid location across trials of the experiment. The location of the other three items in the display was randomly determined.

On each trial, a visual display was presented and subjects heard a single auditory instruction to move one object. The critical instructions were either ambiguous or unambiguous. The ambiguous utterances were created by digitally excising the complementizer *that's* from the unambiguous instruction (*put the book ~~that's~~ on the chair in the bucket*). For critical trials, half used the verb *put* and half used *place.* The instructions were recorded by a female native speaker of British English who was naïve with respect to the experimental hypotheses. The sentences were uttered at a normal speaking rate and with the prosody the naïve speaker considered to be normal. The forty-eight critical utterances were placed into four lists that were counterbalanced with display type. Lists were rotated in a Latin Square design, so that each subject saw each display in only one of the conditions.

The 96 filler items contained both a variety of different objects and instructions designed to mask the experimental items. In the fillers, 32 displays contained one compound object and three single objects, and 64 contained four single objects. Twenty-four of the filler displays contained multiple referents (i.e. two objects that would be described with the same word). Twelve of the filler instructions included a prepositional phrase that modified the indirect object, rather than the direct object. An example was *put the coin into the can on the tray*, in the context of a display containing a single coin, a single can, and another can on a tray. Another 12 fillers used a relational modifier to distinguish between two potential goal locations. An example was *move the coin into the glass on the top.* The display in this case contained two glasses, one of which was located above the other. The remaining 72 fillers contained a single prepositional phrase (e.g., *put the apple on the tray*). The initial verb in the filler items consisted of 65% *put* or *place*, and 35% *drag*, *move*, or *push*.

In the practice trials, five displays contained one compound object and three single objects, and two contained four single objects. Four of the practice instructions contained single prepositional phrase (e.g., *put the apple on the tray*), and two contained a prepositional phrase modifier (e.g., *put the apple on the tray in the box*). Of those containing a modifier, one was ambiguous and other was unambiguous. There were 288 objects used in the experiment in total, and the location of objects in both filler and practice trials was randomly determined.

Apparatus

Eye movements were recorded with an SR research Eyelink 1000 eye tracker sampling at 1000 Hz. Viewing was binocular, but only the position of the right eye was tracked. Stimulus presentation was programmed using SR research Experiment Builder software. The eye tracker and a 19″ CRT display monitor (refresh rate of 140 Hz) were interfaced with a 3-GHz Pentium 4 PC, which controlled the experiment and logged the position of the eye throughout the experiment.

Design and procedure

We used a 2 × 2 mixed design. Instruction type was either ambiguous or unambiguous, and was manipulated within subjects. Display type contained either one or two referents and was manipulated between subjects.[1] Participants completed six practice trials, 24 experimental trials, and 96 fillers. Trials were presented in a fixed pseudo-random order in four different lists. At least two filler trials always separated the critical trials.

Participants were told that on each trial they would see an array of pictures and that those would be accompanied by instructions to perform certain actions on the objects shown. Participants were to execute the instructions as quickly and as accurately as possible. At the beginning of each trial, participants were required to fixate a drift correction dot in the center of the screen. The experimenter then initiated the trial. The visual display appeared 3 s prior to the onset of the spoken instruction, and participants pressed the space bar once they had finished moving the object. The session lasted approximately 25 min.

Results and discussion

Results were analyzed using logit mixed effects models (Baayen, 2008; Baayen, Davidson, & Bates, 2008; Barr, 2008; Jaeger, 2008). Logit mixed models have been advocated as more appropriate for binomial data than are ANOVAs performed over transformed proportions (Jaeger, 2008). We included subjects and items as random effects, as well as by-subject random slopes for instruction type and by-items random slopes for instruction type and display type (Barr, Levy, Scheepers, & Tily, 2013). In cases in which the maximal model failed to converge, we sequentially simplified the item and subject random slopes until convergence was achieved. The first dependent variable was whether participants performed the movement specified in the instruction correctly (see total correct in Table 1). Correct performance was coded with a 0 and errors with a 1.[2] The second dependent variable was whether a fixation was made to a particular object during a specific time window. We analyzed three 1000 ms time windows, which were time locked to the onset of each of the nouns in the instruction (Altmann & Kamide, 2004). Depending on the length of the noun, there was some overlap of the three analysis windows. In cases where there was overlap, the mean duration of that overlap was 120 ms. In accordance with standard practice, each time window was shifted forward by 200 ms to account for saccade planning time (Altmann & Kamide, 2004).

For each analysis, we first created a base model, which included an intercept and the two random factors. To the base, we sequentially added display type, instruction type, and their interaction as fixed effects. If the inclusion of additional factors significantly improves fit over the base, then we can conclude that it accounts for a significant amount of variation in the data.[3] Variables were entered one after the other, and a third model tested the interaction. We assessed model improvement via log-likelihood ratio tests using the lme4 package in R (Bates, Maechler, & Dai, 2008). This test compares models using a χ^2 which determines whether an additional predictor improves model fit. In cases where a predictor significantly improved fit, the Wald statistic was used to show that coefficients differed significantly from zero (Agresti, 2002).

Errors

We began the analysis by examining trials that resulted in the correct movement compared to trials that contained an error (see Table 1). The results of the mixed model analyses showed that model fit was not improved with the

[1] We manipulated display type between participants to minimize the possibility that participants would notice the contrast between one- and two-referent displays. Thus, this design allowed us to mask the main experimental manipulation from participants.

[2] If there were multiple errors of the same type on one trial, then only one was counted. If there was a distractor pick up and an incorrect goal drop on one trial, then both were counted, which means that the percentages in Table 1 will not always sum to 100%. Finally, the "other" category contains error pickups and error drops. It is likely that some of these are attributable to accidental mouse clicks and mouse releases.

[3] In addition to conventionally significant effects, we report any marginal *p*-values between .05 and .07.

Table 1
Summary of mouse movements for each of the three eye tracking experiments.

	Ambiguous (%)		Unambiguous (%)	
	1-Referent	2-Referent	1-Referent	2-Referent
Experiment 1				
Total correct	93.2	86.5	95.3	94.3
Distractor pickups	0.0	8.2	0.0	1.5
Incorrect goal drops	0.0	0.5	0.0	1.0
Others	6.8	4.7	4.7	3.6
Experiment 2				
Total correct	89.6	80.2	91.7	89.1
Distractor pickups	0.0	16.1	0.0	2.6
Incorrect goal drops	1.0	0.5	0.0	0.0
Others	9.4	4.2	8.3	8.3
Experiment 4				
Total correct	96.5	89.2	97.9	97.2
Distractor pickups	0.0	9.4	0.0	0.7
Incorrect goal drops	0.7	0.7	0.3	0.3
Others	2.8	0.7	1.7	1.7

Note: Distractor pickups in the one-referent condition represent the distractor-control object.

inclusion of either display type or instruction type (see Table 2).

In examining the error types, we were particularly interested in the cases in which participants moved the target object to the incorrect goal location (incorrect goal drops), and the cases in which they picked up the wrong object (distractor pickups). These errors, and especially incorrect goal drops, are what one might expect if participants misinterpreted the garden path sentences and did not recover in time to prevent the incorrect action. From Table 1 it can be seen that subjects began moving the distractor object on approximately 8% of trials in the two-referent ambiguous condition, but they almost never actually placed the distractor onto the incorrect goal. This low error rate is consistent with previous studies (Farmer, Cargill, Hindy, Dale, & Spivey, 2007a; Novick et al., 2008; Trueswell et al., 1999).

Eye movements

For the eye movement analysis, we began by examining the first time window which corresponded to the onset of the first noun (e.g. Put the book (that's) on the chair in the bucket). Here we were interested in examining looks to the target object (e.g. the book on the chair) and to the distractor (e.g. the single book or the football). We refer to the single book in the two referent display as the distractor, and we refer to the unrelated object in the one-referent display (e.g. the football) as the distractor-control. (The distractor-control in one-referent trials always appeared in the same region as the distractor object in two referent trials.) The proportion of trials with a fixation to the target is shown in Table 3. The mixed model analysis showed that model fit was significantly improved with the inclusion of display type, and specifically, there were more looks to the target with one-referent displays (see Table 2). The analysis of looks to the distractor and distractor-control showed that the model containing display type was a significantly better fit over the base model. This is expected because, in the two-referent condition there should be competition between the two potential referents (e.g. the two books); but in the one-referent condition, the distractor-control is an irrelevant item and never mentioned, so looks to this object should be at or near chance. The proportion of trials with a fixation over time is shown in Fig. 2. The probability graphs were created by dividing time into 100 ms time

Table 2
Logit mixed model analyses for display type and instruction type in Experiment 1.

Predictor	Estimate	*SE*	Wald-*Z*	*p*
Errors				
Model fit was not improved with either variable				
Target fixations				
Best fit model with display: $\chi^2(1) = 16.37$, $p < .001$				
(Intercept)	2.528	0.215	11.744	<.001***
Display – two referent	−1.177	0.276	−4.258	<.001***
Distractor fixations				
Best fit model with display: $\chi^2(1) = 8.52$, $p < .001$				
(Intercept)	−0.810	0.147	−5.517	<.001***
Display – two referent	0.668	0.213	3.135	.002***
Incorrect goal fixations				
Best fit model with display, instruction, and interaction: $\chi^2(1) = 4.20$, $p < .05$				
(Intercept)	2.374	0.233	10.17	<.001***
Display – two referent	0.380	0.255	1.493	0.136
Instruction – unambiguous	−1.211	0.281	−4.304	<.001***
Display × Instruction	0.659	0.315	2.093	.036*
Correct goal fixations				
Best fit model with instruction: $\chi^2(1) = 3.76$, $p = .05$				
(Intercept)	2.103	0.246	8.553	<.001
Instruction – unambiguous	−.0486	0.222	−2.189	.029*

Note: Target and incorrect goal fixations failed to converge and so random slopes were simplified.

Table 3
Means and standard errors for proportion of trials with a fixation across all three visual world experiments.

	Ambiguous		Unambiguous	
	1-Referent	2-Referent	1-Referent	2-Referent
Experiment 1				
Target object	.91(.02)	.75(.03)	.94(.02)	.79(.04)
Distractor object	.32(.04)	.48(.04)	.32(.03)	.45(.04)
Incorrect goal	.46(.04)	.35(.04)	.33(.04)	.36(.06)
Correct goal	.88(.02)	.83(.05)	.82(.03)	.81(.04)
Experiment 2				
Target object	.91(.02)	.77(.03)	.93(.01)	.71(.04)
Distractor object	.34(.04)	.65(.04)	.39(.03)	.70(.03)
Incorrect goal	.43(.04)	.48(.02)	.38(.03)	.43(.04)
Correct goal	.86(.03)	.77(.03)	.87(.02)	.82(.04)
Experiment 4				
Target object	.67(.03)	.47(.05)	.63(.04)	.54(.04)
Distractor object	.23(.03)	.62(.04)	.19(.03)	.64(.04)
Incorrect goal	.22(.04)	.38(.04)	.26(.04)	.29(.03)
Correct goal	.66(.04)	.58(.05)	.61(.05)	.71(.04)

Note: The target and distractor looks correspond to the first analysis window, the incorrect goal corresponds to the second window, and the correct goal corresponds to the third window. The incorrect goal is the key object with respect to garden path misinterpretations, and has been highlighted for convenience.

bins, and then calculating the proportion of trials that contained a fixation to an object.

The second window was time locked to the onset of the second noun in the instruction (e.g. Put the book (that's) on the chair in the bucket). Again, we analyzed a 1000 ms time window, but in this window the object of interest is the incorrect goal (e.g. the empty chair), as this indexes the garden-path misinterpretation (Spivey et al., 2002). The results of the mixed model analysis showed that model fit was improved with the inclusion of the interaction. As can be seen in Table 3, the interaction is driven by the increased probability of fixating the incorrect goal in the one-referent ambiguous condition (Spivey et al., 2002; Tanenhaus et al., 1995). The comparison of ambiguous vs. unambiguous instructions with one-referent displays was significant [$\chi^2(1) = 6.97$, $p < .01$; (Intercept): Estimate = −0.186, SE = 0.214, Wald-Z = −0.872, p = .383; Instruction – unambiguous: Estimate = −0.585, SE = 0.217, Wald-Z = −2.69, $p < .01$], and the comparison of the one- vs. two-referent displays with ambiguous instructions was also significant [$\chi^2(1) = 4.36$, $p < .05$; (Intercept): Estimate = −0.179, SE = 0.197, Wald-Z = −0.922, p = .356; Display – two referent: Estimate = −0.471, SE = 0.216, Wald-Z = −2.18, $p < .05$].

The third, and final, window was time locked to the onset of the third noun (e.g. Put the book (that's) on the chair in the bucket), which is the correct goal. The results showed that model fit was improved by instruction type, and in this window, the ambiguous instructions resulted in more looks to the correct goal.

To summarize, this experiment replicated the results found in previous work with a computerized version of the VWP and instructions carried out via mouse movements (see also Farmer et al., 2007b). We found that participants were more likely to fixate the incorrect goal when there was a single referent in the display and when the instruction was ambiguous. In contrast, there was no effect of ambiguity and few looks to the incorrect goal with the two-referent display. This result has been taken as evidence that the language processing system can use visual context (i.e. the presence of two books) to immediately guide the interpretation of the ambiguous prepositional phrase to the more complex modifier alternative. The presence of two referents means that a modifier is necessary to allow one book to be distinguished from the other, and so listeners interpret the ambiguous prepositional phrase as a modifier rather than a goal. In the one-referent condition, the modifier appears to be unmotivated, and so the goal analysis of the ambiguous prepositional phrase is preferred. We also observed few movement errors, which is consistent with previous work. Our next step was to examine whether this pattern would be obtained if the participants did not preview the visual display before hearing the sentences.

Experiment 2

The influence of preview on processing of syntactic ambiguity has not yet been investigated, but some indication of its potential role comes from a study that manipulated preview to observe its effects on phonological processing (Huettig & McQueen, 2007). In one experiment, participants were given a 3 s preview of four unrelated objects before hearing a sentence that mentioned one of those objects; in a follow-up experiment, the preview was reduced to 200 ms. Preview, in this case, seemed to provide time for the phonology of the names of each of the objects to become activated, as the advantage for phonological competitors in the first experiment disappeared when the preview was eliminated in the second. This pattern of results suggests that, during the preview, participants are able to get a head start on linguistic processing – they begin anticipating names of the objects that could occur in the upcoming utterance.

In our second experiment, we also eliminated the preview in order to assess its effects on the use of context to resolve syntactic ambiguity. The visual display appeared at the same time that the spoken instruction began, so participants had to process both the visual and the linguistic information simultaneously. If preview is used to generate expectations concerning not just likely object names but also the form and content of the utterance, then we expect to find that participants will be less able to use the visual context to assess whether a prepositional phrase is likely to be a modifier or a goal. In addition, performance errors involving movement of the distractor object or placement of the target object on the wrong goal should be more frequent than they were in Experiment 1.

Method

Thirty-two students from same participant pool participated in Experiment 2, and were compensated in the same manner. None had participated in Experiment 1. The mate-

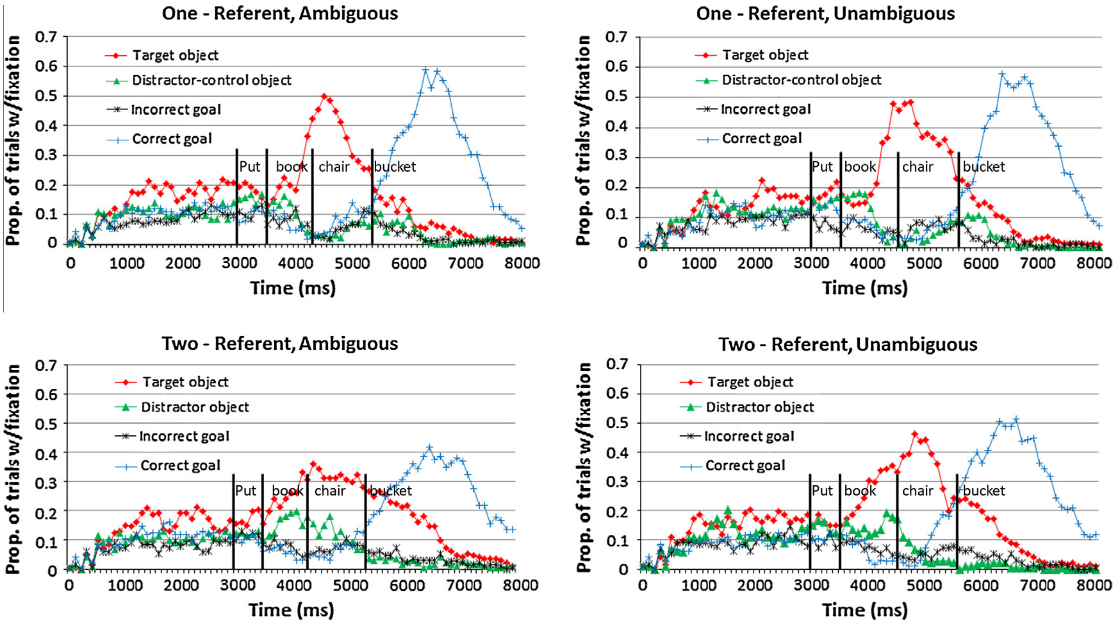

Fig. 2. Proportion of trials with a fixation to each object broken down by the four conditions in Experiment 1. The vertical black lines represent the mean onset of the verb and the three nouns.

rials and apparatus were the same as in Experiment 1. The design and procedure were also the same, except that the visual display appeared at the same time as the onset of the auditory instruction.

Results and discussion

Data analysis procedures were the same as in Experiment 1.

Errors

We began the analysis by examining cases in which participants performed an incorrect action given the instruction (see Table 1). The mixed model analysis showed that model fit was significantly improved with the inclusion of instruction (see Table 4). The overall pattern of results was similar to the previous experiment, except that elimination of the preview phase almost doubled the number of distractor pickups (attempts to move the wrong book). Distractor pickups again occurred primarily in the two-referent ambiguous condition, and there were almost no trials on which an incorrect goal drop occurred.

Eye movements

As before, the first time window was aligned with the onset of the first noun (e.g. Put the book (that's) on the chair in the bucket). We were interested in looks to the target object (e.g. the book on the chair), the distractor object (e.g. the single book, which occurs in the two-referent condition), and the distractor-control (e.g. the football, which replaced the single book in the one-referent condition). The proportion of trials with a fixation is shown in Table 3. The mixed model analysis showed that model fit was significantly improved over base with the inclusion of display type, instruction type, and the interaction. The comparison of the one- vs. two-referent displays with ambiguous instructions was significant [$\chi^2(1) = 5.03$, $p < .05$; (Intercept): Estimate = 2.269, $SE = 0.248$, Wald-$Z = 9.163$, $p < .001$; Display – two referent: Estimate = −0.886, $SE = 0.349$, Wald-$Z = -2.54$, $p < .05$], and the comparison of one- vs. two-referent displays with ambiguous instructions was also significant [$\chi^2(1) = 13.32$, $p < .01$; (Intercept): Estimate = 4.046, $SE = 0.634$, Wald-$Z = 6.382$, $p < .001$; Display – two referent: Estimate = −2.824, $SE = 0.715$, Wald-$Z = -3.953$, $p < .001$]. The two one-referent conditions were not significantly different from one another, and neither were the two two-referent conditions. Thus, in this time window, there were fewer looks to the target object with the two-referent displays and when the instruction was unambiguous (see Fig. 3).

Looks to the distractor object (e.g., the lone book) and the distractor-control objects (e.g., the football), showed strong effect of display type, similar to what we found in Experiment 1. The one-referent displays resulted in approximately 36.5% of trials with a fixation to the distractor control (34% in the ambiguous and 39% in the unambiguous conditions), and the two-referent displays resulted in approximately 67.5% of trials with a fixation to the distractor (65% for ambiguous and 70% for unambiguous conditions). This shows the expected competition between the two books with two-referent displays.

In the second time window, which corresponded to the onset of the second noun in the instruction (e.g. Put the book (that's) on the chair in the bucket), we examined looks to the incorrect goal (i.e. the chair). Analyses showed no significant improvement over the base model.

The third time window was aligned with the onset of third noun (e.g. Put the book (that's) on the chair in the

Table 4
Logit mixed model analyses for display type and instruction type in Experiment 2.

Predictor	Estimate	*SE*	Wald-*Z*	*p*
Errors				
Best fit model with instruction: $\chi^2(1) = 6.05$, $p < .05$				
(Intercept)	−2.417	0.291	−8.307	<.001***
Instruction – unambiguous	−1.922	0.577	−3.332	<.001***
Target Fixations				
Best fit model with display, instruction, and interaction: $\chi^2(1) = 4.96$, $p < .05$				
(Intercept)	2.304	0.259	8.882	<.001***
Display – two referent	−0.900	0.369	−2.413	.016*
Instruction – unambiguous	1.776	0.674	2.637	.008**
Display × Instruction	−1.952	0.783	−2.494	.013*
Distractor fixations				
Best fit model with display: $\chi^2(1) = 34.06$, $p < .001$				
(Intercept)	−0.604	0.139	−4.354	<.001***
Display – two referent	1.397	0.202	6.913	<.001***
Incorrect goal fixations				
Model fits were not improved with either variable.				
Correct goal fixations				
Best fit model with display: $\chi^2(1) = 4.56$, $p < .05$				
(Intercept)	1.898	0.173	10.990	<.001***
Display – two referent	−0.511	0.230	−2.216	.027*

Note: Correct goal fixations failed to converge and so the random slopes were simplified.

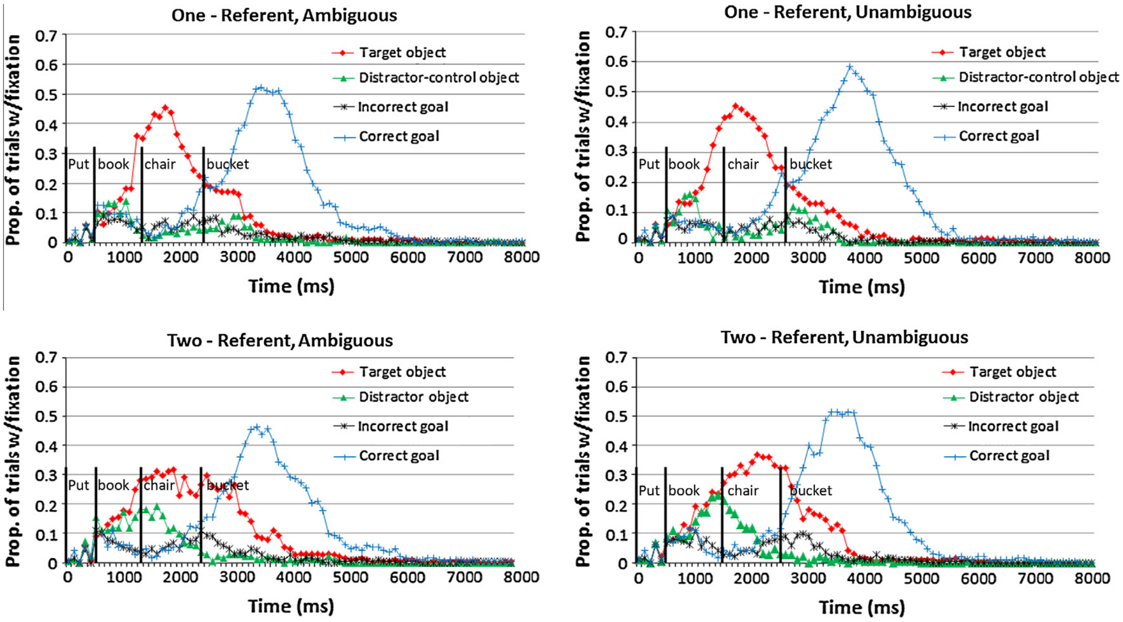

Fig. 3. Proportion of trials with a fixation to each object broken down by the four conditions in Experiment 2. The vertical black lines represent the mean onset of the verb and the three nouns.

bucket). The results showed that model fit was significantly improved when display type was included in the model (see Table 4). The pattern was more looks to the correct goal with the one-referent displays compared to the two-referent displays. This finding is different from Experiment 1, but consistent with the one-referent conditions being easier to complete compared to the two-referent conditions because there is less opportunity to become confused over which object should be moved (e.g., which book).

In summary, in this experiment, we eliminated the preview phase, and therefore, participants had to process the visual and linguistic information simultaneously. The lack of preview changed performance in two important ways. First, there were nearly twice as many distractor pickups in the two-referent ambiguous condition as in Experiment 1. This increased error rate suggests a greater difficulty in resolving the referential ambiguity about which of the two potential referents needed to be moved. It appears that, in this situation, there was a tendency to move the more prominent of the two potential referents (i.e., the single book). Consistent with the increase in distractor pickup errors, we also observed more fixations on the distractor object in the two-referent displays compared to what was found in the previous experiment (∼.68 vs. ∼.47).

The second main difference in results across the two experiments was that, in this one, fixations on the incorrect goal were relatively high and similar across all four conditions, and performance errors involving the incorrect goal were rare. Recall that, in the first experiment, we found an interaction in which there were more looks to the incorrect goal in the one-referent condition with an ambiguous instruction – the classic pattern. The current results indicate that when participants are deprived of preview, they are less able to use the visual context to resolve the temporary ambiguity. On the face, the lack of preview seems to make it **more** likely that participants activate the misinterpretation. However, we think this is unlikely, but instead, suggests that the preview is necessary for the ambiguity-avoidance effect that is normally observed with the two-referent contexts. Before going into more detail about what processes might be taking place during the preview, we turn to a production experiment that was designed to assess whether participants accrue information concerning the relationship between the visual contexts and the utterances that go with them.

Experiment 3

The purpose of this experiment was to empirically evaluate the hypothesis that participants in the standard version of the VWP (which includes preview) are able to make predictions about both the structure and content of the utterance that will be associated with the visual displays. Huettig and McQueen (2007) have already shown that listeners activate phonological representations of objects during preview. The current experiment extends this idea by examining the content of participants' syntactic expectations as well. To explore the possibility that listeners can generate predictions, we conducted a production experiment in which participants saw one- and two-referent visual displays, and they then generated what they thought the instruction for that display might be.[4] For example, participants would see a display like the one shown in Fig. 1, and their task was to guess the instruction that would go with it. Given that the display contains five objects, chance performance would be 20% for each of the potential nouns in the sentence. (In the critical displays, one object was located on or in another object, and those two objects could be referred to separately.) If people are better than chance at predicting the form and content of the instructions, then we can infer that the display and task constraints provide enough information to allow listeners to make reasonably accurate predictions concerning utterance content.

Method

Participants

Ten undergraduate students from the University of Northumbria agreed to participate. They were all native speakers of British English and all had normal or corrected-to-normal vision. Each was paid £3.00 for their participation.

Materials

We utilized 82 displays from Experiments 1 and 2. Twenty-four were the critical items, and 58 were taken from the fillers. The experiment had a training phase and a test phase. In the training phase, participants saw 34 filler displays and heard the accompanying sentence. There were six one-referent and six two-referent trials that contained the same features of the critical items used in Experiments 1 and 2. Twenty-two consisted of a direct object noun phrase followed by single prepositional phrase indicating the goal location (e.g. *put the book in the bucket*). The test phase consisted of 48 visual displays with no accompanying speech. Twenty-four of these were the critical displays from Experiments 1 and 2 (see Fig. 1), and the other 24 were taken from the filler trials. Of the 24 filler displays in the test phase, six contained two-referents of the same type and 18 did not.

Design and procedure

The experiment consisted of a single variable (display type) with two levels (i.e. one- or two-referent). This variable was manipulated within subjects. In the training phase of the experiment, participants' task was simply to view 34 displays and listen to the accompanying instruction. Participants were not required to perform any overt response. The critical one- and two-referent trials were always separated by two filler trials, and items were rotated across two lists. In the test phase, participants viewed a visual display, and they had to say what they thought the corresponding instruction for the display was. No feedback was given. The dependent variable was the number of correctly identified objects included in these utterances, referenced to the original instructions used in Experiments 1 and 2.

Data coding

We were primarily interested in the target object, distractor, and the goal, as well as any prepositional phrase modifiers produced. Responses were coded 0 for an incorrect answer and 1 for a correct answer. If the utterance produced in response to a display such as Fig. 1 were *place the bucket on the chair*, then this response would be coded as 0 for the target object, goal, and modifier. On the other hand, if participants produced *place the book on the chair in the bucket*, then this would receive a coding of 1 for the target object, goal, and modifier. Coding in the two referent condition is more complicated because there were a substantial number of ambiguous references. For example, if the participant produced an instruction such as *put the book in the bucket* with a display like Fig. 1, then the instruction could be described as under-specified because it does not contain sufficient information to determine

[4] We do not differentiate between the terms expectation, bias, anticipation, or prediction.

which object the participant was referring to. To deal with this problem, we combined the target and distractor references, which affected our estimate of chance performance (see below).

Data analysis

We conducted two sets of analyses. In the first, we performed one-sample *t*-tests in which we compared the proportion of correct responses to chance performance. There were five objects in each critical display, and so chance performance is .20 correct. The one exception is that we combined the target and distractor references in the two-referent condition (because of the under-specification problem mentioned above), and so chance in this case was .40. The second set of analyses compared performance between the one- and two-referent conditions, and here we conducted logit mixed effects analyses, in the same way as in the previous experiments.

Results and discussion

The data are presented in Table 5. We use Fig. 1 as the example display. Beginning with the target object (e.g. the book on the chair), in the one-referent condition, participants were no better than chance at guessing which object would be the target (or the moved) object. However, considering just the one-referent condition, they were significantly more likely than chance to choose the distractor-control object (e.g. a football). In the two-referent condition, as mentioned previously, we observed a substantial number of ambiguous references (e.g. *put the book in the bucket*, with no modification to indicate which book.). In Table 5, the Target and Distractor columns contain only the unambiguous responses. One-sample *t*-tests showed that participants were no different from chance at unambiguously identifying the target object in both the one- and two-referent conditions. However, in the two-referent condition, 52% of the utterances contained an ambiguous reference to one of the two books. To deal with under-specification, we summed the target and distractor references and then conducted one-sample *t*-tests with a test value of .40. Results in both conditions were significant, showing that participants have a better than chance likelihood of guessing which objects in the display are likely to be the ones that will need to be moved.

As can be seen in Table 5, participants produced a prepositional phrase modifier on approximately 40% of trials (Engelhardt et al., 2006), and they to include a modifier more in the two-referent condition. Finally, the correct goal was guessed significantly more often than chance in both the one- and two-referent conditions, with no difference between them. Thus, participants can correctly guess which object will be the goal more than half of the time.

In the two-referent condition, people seemed to be able to anticipate that they would have to move one of the two contrasting objects (e.g. books) to another location. In the one-referent condition, participants were biased more towards the distractor-control object than to the target. However, in both one- and two-referent conditions, participants were fairly good at determining which object(s) would need to be moved and which would serve as goals. Thus, the production results indicate that participants in visual world experiments with not a great deal of exposure to the display–instruction pairings are able to predict with a reasonable degree of accuracy what type of instruction they will hear, especially in two-referent contexts. One reason not to expect perfect accuracy is that it appears participants have little trouble violating real-world plausibility constraints when they make these predictions. For example, they came up with sentences such as *put the fish in the cage and put the stool on the cake*. Interestingly, the same was true in our first two VWP experiments, in that we sometimes asked participants to execute an instruction that would be difficult if not impossible to make in the real world. In previous visual world experiments, such as in Spivey et al. (2002), participants manipulated real objects, not computer images, and so these affordances were obviously more constraining. Thus, it is possible that instruc-

Table 5
Summary of production data Experiment 3. Percentages represent the number correct divided by total number of trials per condition (i.e. 12).

	Target	Distractor	Ambiguous	Target + Distractor	Modifier	Goal
One-referent	32.8%	51.2%	–	84.1%	28.0%	59.4%
One-sample *t*-test	$t(9) = 1.21$	$t(9) = 3.98^{**}$	–	$t(9) = 8.00^{**}$	$t(9) = .74$	$t(9) = 4.59^{**}$
Two-referent	40.1%	0.0%	52.2%	92.3%	40.1%	62.5%
One-sample *t*-test	$t(9) = 1.62$	–	$t(9) = 1.12$	$t(9) = 14.65^{**}$	$t(9) = 1.62$	$t(9) = 4.20^{**}$
1 vs. 2 referent	$\chi^2(1) = 0.11$	–	–	$\chi^2(1) = 6.25^{*}$	$\chi^2(1) = 1.04$	$\chi^2(1) = 0.31$
(Intercept)						
Estimate				−0.894		
SE				0.928		
Wald-*Z*				−0.963		
p				.335		
Display type – two referent						
Estimate				2.870		
SE				0.828		
Wald-*Z*				3.467		
p				$.001^{***}$		

Note: All one-sample *t*-tests were conducted with a test value of .20, except for the Ambiguous (Target + Distractor) comparisons, which had a test value of .40.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

tions were even more predictable in the original studies using real objects manipulated with hand movements than in the current experiments using images presented on a computer screen and manipulated using a computer mouse.

One important cautionary note is that this experiment has shown only that participants are capable of anticipating many aspects of the form and content of the instructions that would likely be associated with the visual displays. We have not provided direct evidence that listeners in VWP experiments, such as our Experiment 1, do in fact make these kinds of predictions. However, it would not be unreasonable to postulate that they do given the increasing evidence for prediction in language processing (e.g., Altmann & Kamide, 1999; van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005).

One idea that naturally follows from our hypothesis concerning prediction is that performance should change over the course of the experiment. This is because prediction during the preview (in Experiment 1) would only be possible after exposure to a certain number of critical trials. Thus, an obvious post hoc analysis is to investigate "trial order" as a predictor variable. Specifically, we investigated whether there were changes in performance in how listeners dealt with syntactic ambiguity given increased experience to instruction–display combinations. We hypothesized that looks to the incorrect goal in the one-referent (ambiguous) condition would not show an effect of trial order. This is because looks to the incorrect goal were relatively high in both Experiments 1 and 2, and also, there were fewer prepositional phrase modifiers produced in this condition in Experiment 3. In contrast, in the two-referent (ambiguous) condition, looks were relatively low in Experiment 1, relatively high in Experiment 2, and nearly half of utterances produced in Experiment 3 contained a prepositional phrase modifier. Thus, our conjecture is that prediction is easier with two-referent contexts, and at least part of listeners' ability to avoid the garden path in two-referent contexts is dependent on within-experiment experience, which gives rise to expectations or predictions during the preview.

In the follow-up analysis, we utilized the same statistical procedures as the previous experiments. The design of this analysis included trial order as a continuous variable and display type (one- vs. two-referents) as a categorical variable. We tested only the ambiguous conditions. Results showed that model fit was significantly improved over the base with the inclusion of both display type and trial order (see Table 6). To provide an indication of the magnitude of the trial order variable, we split the trials into first half and second half. For the proportion of trials with a fixation to the incorrect goal, there was a decrease from .50 to .42 with one-referent contexts, and a decrease from .41 to .30 with two-referent contexts.[5] Therefore, both conditions showed approximately similar trial order effects in which the tendency to fixate the incorrect goal decreases over time.

Table 6
Logit mixed model analyses for display type and trial order in Experiment 1.

Predictor	Estimate	*SE*	Wald-*Z*	*p*
Incorrect goal fixations				
Best fit model with display and trial order: $\chi^2(1) = 5.34$, $p < .05$				
(Intercept)	0.382	0.312	1.23	.221
Display – two referent	−0.482	0.217	−2.219	.027*
Trial order	−0.009	0.004	−2.396	.017*

This suggests that participants have a decreasing tendency to be misled by the ambiguous prepositional phrase over time, suggesting that they are less garden pathed in both contexts. However, it is important note that the effect of display type remained robust even at the end of the experiment.

To summarize, in the three experiments described so far, we have replicated the standard referential garden-path effect, and then eliminated it by depriving participants of preview. We also showed that participants exposed to visual arrays can predict relatively accurately the structure and content of the instruction likely to be associated with a particular configuration. In the Introduction, we also hypothesized that if there are many objects rather than just four or five, then even with preview, there will be too many possibilities concerning which objects are moveable and which objects are goals to allow useful predictions (or expectations) to be made concerning the content of the upcoming utterance. Experiment 4 tests this idea.

Experiment 4

This experiment was similar to Experiment 1, but we added eight objects to the array for a total of 12 (see Fig. 4). As in Experiment 1, participants previewed the displays before hearing the utterance. This change in number of objects increased the visual complexity of the displays, and we assumed that the increase in complexity would lead to more difficulty overall. Not only is visual search for the mentioned objects now potentially more difficult, but the greater number of objects in the array increases the number of potential objects and goal locations, making it more difficult for participants to anticipate the content of the spoken instruction. On the other hand, if it is preview that is critical for the integration of visual and linguistic information, then it is possible that participants in this experiment will perform the same way they did in Experiment 1.

Method

Thirty-two students from the same participant pool as in Experiments 1 and 2 were tested in this experiment. None had participated in the previous experiments. The materials were the same as in Experiment 1, except that eight additional objects were added to each display, so that all 12 cells of the array contained an object. Twenty-six filler trials contained one compound object, 46 contained two compound objects, and 24 contained no compound objects. The critical trials all contained two compound

[5] The unambiguous one-referent condition showed almost no change over the course of the experiment (i.e. .34 vs. .32), and the unambiguous two-referent condition showed a slight increase in looks to the incorrect goal (i.e. .33 vs. .39).

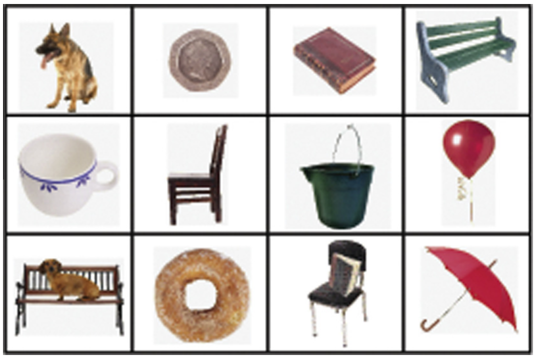

Fig. 4. Example display for the two-referent condition in Experiment 4. Grid lines not shown to participants.

objects, one that was the target (in Fig. 4, the book on the chair), and another that was included to make sure the target compound object did not stand out (the dog on the bench). In addition, for both compound objects, different tokens of the individual constituent objects were also included (that is, a single book, a single chair, a single dog, and a single bench). The procedure was the same as in Experiment 1, except that participants were given a slightly longer period (3.5 s instead of 3 s) in which to preview the objects prior to the onset of the auditory instruction.

Results and discussion

Errors

We began the analysis by examining correct trials compared to trials that contained an error of any type (see Table 1). The mixed model analysis showed that model fit was significantly improved with the inclusion of both instruction type and display type (see Table 7). There were more errors with the two-referent displays and with the ambiguous instructions. Examining the error types in Table 1 indicates that the number of distractor pickups was similar to that obtained in Experiment 1, and therefore, lower than what we observed in Experiment 2. It therefore appears that the 3.5 s preview was sufficient to allow participants to arrive at an interpretation to accurately guide their overt actions, even when they were confronted with many more objects. The next question is whether eye movement performance will also be the same as in Experiment 1.

Eye movements

The first time window was aligned with the onset of the first noun (e.g. Put the <u>book</u> (that's) on the chair in the bucket), and we were interested in looks to the target object (e.g. the book on the chair), the distractor (e.g. the single book), and the distractor-control (e.g. the football). The proportion of trials with a fixation is shown in Table 3. The mixed model analysis showed that model fit was significantly improved over base with the inclusion display type. In this time window, there were more looks to the target object with the one-referent displays compared to the two-referent displays, which is the expected pattern. Looks to the distractor and distractor-control objects again showed a robust effect of display type. The one-referent displays resulted in approximately 21% of trials with a fixation to the distractor control, and the two-referent displays resulted in approximately 63% of trials with a fixation to the distractor. The model containing display type was a significantly better fit over the base model (see Table 5).

In the second time window, which corresponded to the onset of the second noun in the instruction (e.g. Put the book (that's) on the <u>chair</u> in the bucket), we examined looks to the incorrect goal. Analyses showed that model fit was not improved with the inclusion of either display type or instruction type. The pattern of means however,

Table 7
Logit mixed model analyses for display type and instruction type in Experiment 4.

Predictor	Estimate	*SE*	Wald-*Z*	*p*
Errors				
Best fit model with display and instruction: $\chi^2(1) = 7.26$, $p < .01$				
(Intercept)	−3.712	0.466	−7.966	<.001***
Display – two referent	1.775	0.535	3.315	<.001***
Instruction – unambiguous	−2.970	0.664	−4.473	<.001***
Target fixations				
Best fit model with display: $\chi^2(1) = 4.61$, $p < .05$				
(Intercept)	0.755	0.229	3.302	<.001***
Display – two referent	−0.680	0.304	−2.238	.025*
Distractor fixations				
Best fit model with display: $\chi^2(1) = 37.02$, $p < .001$				
(Intercept)	−1.624	0.231	−7.035	<.001***
Display – two referent	2.255	0.309	7.296	<.001***
Incorrect goal fixations				
Fit not improved with either variable				
Correct goal fixations				
Best fit model with display, instruction, and interaction: $\chi^2(1) = 4.15$, $p < .05$				
(Intercept)	0.863	0.306	2.818	.005**
Display – two referent	−0.496	0.394	−1.259	.208
Instruction – unambiguous	−0.292	0.356	−0.819	.413
Display × Instruction	1.023	0.484	2.115	.034*

was different from that observed in the first two experiments. Specifically, there were more looks to the incorrect goal in the condition in which the visual context contained *two* referents and the instruction was ambiguous.

The third window was aligned with the onset of third noun (e.g. Put the book (that's) on the chair in the bucket). The results showed that model fit was significantly improved when the interaction was included in the model. The pattern is that looks to the correct goal were about equally likely in the one-referent condition regardless of instruction ambiguity, but in the two-referent condition, looks to the correct goal were more likely when the instruction was unambiguous. However, none of the paired comparisons were significant (*ps* > .20). This pattern suggests that the two-referent condition is more difficult than the one-referent condition, and thus listeners benefit more from the presence of linguistic disambiguation.

In summary, this experiment has provided evidence that listeners have difficulty using visual context to constrain their interpretations online when the context is moderately more complex. However, we also found that overt errors were as unlikely with 12 objects as with four, and we also observed more saccades to the named target object than to random objects in the display. Both of these results suggest that a great deal of useful information was obtained from the visual displays even though they were more complex than the ones used in Experiments 1 and 2.

General discussion

In this section, we begin by summarizing the results and relating them to previously reported studies of syntactic ambiguity resolution in the context of a relevant visual world. Then we will consider the implications of the results for issues relating to the timing and complexity of visual information during online comprehension. Finally, we will discuss what these experiments tell us about the ability of the language processing system to anticipate or predict particular structures on the basis of information accrued over the course of an experiment. We feel that this work contributes to the growing body of literature on the role of prediction as an adaptive language comprehension strategy (e.g. Altmann & Kamide, 1999; Swets, Desmet, Clifton, & Ferreira, 2008).

Summary of main findings

The first key result is our replication of the classic finding that listeners are garden-pathed when they hear a sentence such as *Put the book on the chair in the bucket*, in which the phrase *on the chair* can be analyzed either as a goal or as a modifier of *book*. This occurred when the visual world was previewed and consisted of just a few objects. The misinterpretation is triggered when viewers see only one book in the relevant context, and therefore, the modifier is referentially unmotivated. The measure of garden-pathing is fixations on the empty chair, which are more likely when the visual display contains only one book and much less likely when it contains two books. In the second visual world experiment, when viewers were denied preview of the display, we observed no significant differences in looks to the incorrect goal between any of the four conditions. However, the overall fixation rates were relatively high, that is, they were similar to the one-referent ambiguous condition in Experiment 1 (see Table 3). We did not interpret this pattern as evidence for garden pathing in all conditions, but instead, when visual search is engaged, the system tends to rely on a matching-type comprehension strategy (i.e., when hearing a particular word, identify/fixate all exemplars present). We return to this issue below in the adaptive and flexible processing strategies section. One potential concern is that, without preview, participants may not notice the contrast set, and thus, the two-referent contexts would essentially be perceived as one-referent contexts. However, it is important to note that looks to the distractor were much higher in Experiment 2 than in Experiment 1, suggesting more competition (rather than less) between the two possible referents. This pattern rules out the possibility that participants simply did not have time to apprehend the contrast set.

The production experiment suggests one possibility concerning what might take place during the preview period to allow comprehenders to avoid being garden-pathed in the presence of two referents. Our production results showed that naïve subjects can predict which objects in the two-referent condition will likely need to be moved, and which objects in both one- and two-referent conditions will constitute goals. This idea is further supported by the comprehension data in Experiment 1, where looks to the target object during the preview were higher compared to the other three objects in the array (see Fig. 2). In addition, 40% of utterances produced in the two-referent condition in Experiment 3 contained prepositional phrase modifiers. Thus, our data confirm that participants possess the knowledge needed to make relatively accurate predictions about modification when contrasts are present. This notion is hardly new, but rather, has been part of the logic underlying the VWP since its inception. However, we feel that our production data provide concrete evidence, for this assumption, whereas previous work simply relied on the assumption that the processing system had the ability to make these kinds of pragmatic inferences. Moreover, our data also show that these expectations are crucially reliant on preview of the visual arrays.

In the fourth experiment, in which the number of objects was increased, we observed a slightly different pattern of results. Here, looks to the target with the two-referent contexts were actually lower than looks to the distractor, which suggests that visual search was ongoing during the first time window. This is confirmed, in that looks to the target increase much later in Experiment 4 compared to Experiment 1 (see Figs 2 and 5). The delayed competition associated with the more complex visual arrays has knock-on effects for looks to the incorrect goal. In Experiment 4, looks to the incorrect goal were relatively low compared to Experiments 1 and 2, and again, there was no difference between the four conditions. If anything, the means trended toward an interaction driven by an increased probability of fixating the incorrect goal in the two-referent ambiguous condition, which again suggests

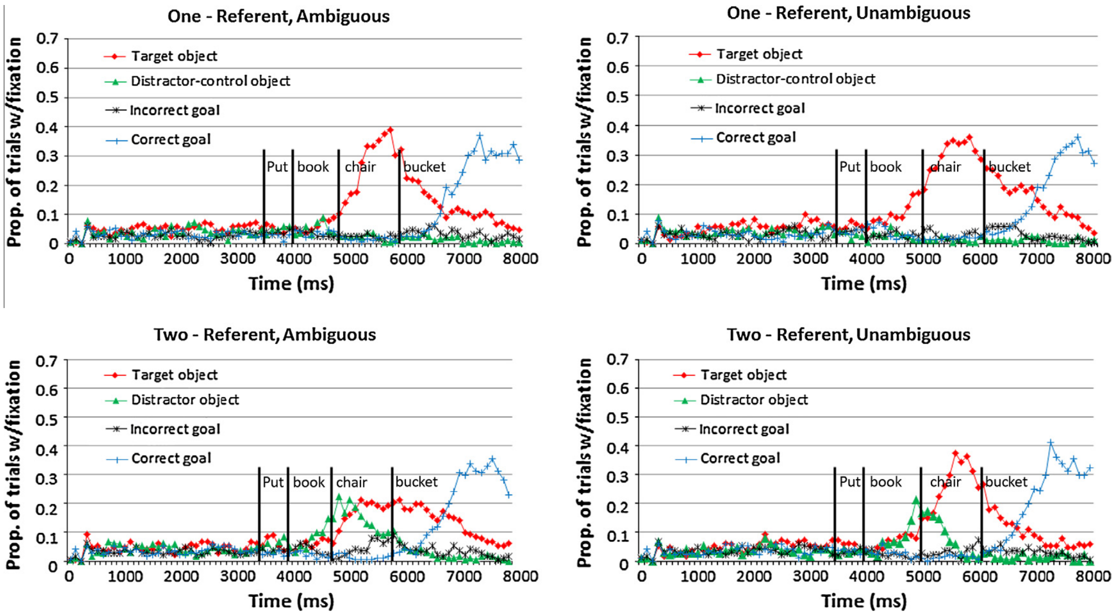

Fig. 5. Proportion of trials with a fixation to each object broken down by the four conditions in Experiment 4. The vertical black lines represent the mean onset of the verb and the three nouns.

that participants do not make the types of inferences concerning modification that they do when the visual array is simple.

Language–vision interactions

People often process language in the context of a relevant visual environment. Language can be used as a tool to help us locate objects in the world, and the visual world provides a background against which linguistic expressions are interpreted. Psycholinguistic studies have shown that language comprehension is incremental: That is, meaning is built up word by word as the utterance unfolds over time (Altmann & Kamide, 1999; Knoeferle, Crocker, Scheepers, & Pickering, 2005; Magnuson, Dixon, Tanenhaus, & Aslin, 2007; Sedivy, 2003; Sedivy, Tanenhaus, Chambers, & Carlson, 1999). What is less often appreciated is that visual information processing is incremental as well (Henderson & Ferreira, 2004; Zelinsky & Schmidt, 2009). Although it is true that the gist of a scene is apprehended in as little as 100 ms (Castelhano & Henderson, 2008; Oliva, 2005; Potter, 1975; Potter & Levy, 1969), sequential fixations are nonetheless required to establish the identities of the specific objects in the scene (Hollingworth & Henderson, 1998; Võ & Henderson, 2011). For visual arrays in the VWP, there is often no scene, and therefore, little opportunity for scene gist to be extracted. It is possible that the object array constitutes some type of semantic or ad hoc category, but it is still unlikely that information about positions and relative locations can be obtained from such arrays as compared to real scenes. Therefore, the processing of the visual information in the VWP is likely incremental – object identities and locations are built up over time similar to how linguistic meaning is built up in language comprehension.

It is useful to distinguish among three vision–language situations. The first is a situation in which linguistic material is processed first and in advance of visual information. This situation is rare in studies of language processing, but is standard in studies of visual search (e.g., the viewer knows to look for a green T) either through verbal instruction or presentation of the target and then a display containing some colored letters is presented (see also Spivey, Tyler, Eberhard, & Tanenhaus, 2001). This situation is of less interest to psycholinguists because they tend to be interested in the real-time processing of linguistic information, which this experimental set-up ignores.

The second vision–language situation is the one that is typical for studies of language processing using the VWP, and is the one that we used in Experiment 1: The visual world is processed first and known in advance of the linguistic information being presented. As we saw from the results of the production experiment, this situation may allow comprehenders to predict some of the content of the upcoming utterances. Of course, this situation is not unusual in the real world. For example, if two people are cooking together, the set of objects with which they will interact and to which they will likely refer may be known (Brown-Schmidt et al., 2005). This second situation is also the one that most easily supports the creation of a contrast set (e.g., two books) in which one of the objects in the set will require a modifier to allow for unique identification.

The third situation is the one that we examined in Experiments 2 and 4, as well as in Andersson et al. (2011). Here, the visual world has to be processed incrementally, as does the spoken utterance. Therefore, the two streams of information must be mapped onto each other and integrated, at least to some extent. The findings from the current experiments suggest that this integration is somewhat challenging. When visual and linguistic processing must operate in parallel, it appears that language comprehenders sometimes adopt a more superficial strategy for interpreting the utterances: Rather than engaging in deep inferencing about the likelihood of a prepositional phrase being a modifier given the referential situation, and rather than attempting to predict the form and content of the utterance, a more rational strategy for the comprehender might be to listen for keywords and make saccades to the objects mentioned (e.g. our Experiment 2). This third situation is one that needs to be understood if we are to have a complete picture of how the visual and language systems interact, because it is certainly one that occurs often in the real world, particularly given that viewers and environments are dynamic rather than static. For example, to return to the example of two people cooking a meal together, someone might move ingredients or tools from one area of the workspace to another, or one person might add or remove items when the other person is not looking. Given that the world is a dynamic place, it is important to understand how the visual and linguistic systems work together in situations that require the two streams of information to be processed and integrated incrementally.

In addition, this analysis highlights the important differences between linguistic and visual context, as discussed in the Introduction. Recall, our observation that most studies exploring the effects of linguistic context on processing allow participants to listen to or read a discourse in order to set up expectations concerning the form or content of subsequent sentences. In contrast, in the latter two situations just described, visual context is entirely co-present with the critical utterances, and so listeners are usually examining that context at the same time that they process the utterance. This is true even when there is visual preview, but it is even more of an issue when preview is denied, because with no preview the visual context must be processed and consulted as the utterance unfolds in time. And, unlike discourse contexts, visual contexts can change, making it even more critical for people to be able to integrate visual and linguistic information incrementally and dynamically. Given these important differences, it is important not to assume that results from studies of discourse context necessarily generalize to studies of visual context and vice versa.

Language adaptation and flexible processing strategies

One important contribution of the current study is to add to the growing body of evidence suggesting that language processing strategies are not fixed and architecturally determined, but instead change in response to task demands (Beckner et al., 2009; Farmer et al., 2011; Kleinschmidt, Fine, & Jaeger, 2012; Scott-Phillips & Kirby, 2010; Swets et al., 2008; Wells, Christiansen, Race, Acheson, & MacDonald, 2009). In a situation in which participants can pre-process a visual world and then receive a sequence of instructions which are similar to each other, they might adopt the following strategy (as argued in the Introduction): First, people will look at the display and identify the likely moveable objects and goals given the objects' affordances. At the same time, perhaps via a mechanism such as syntactic priming (Pickering & Branigan, 1999), they will retrieve the syntactic frame they have been using throughout the experiment, which will be something like null subject – transitive verb – noun phrase – prepositional phrase – (second prepositional phrase). Then they will try to map the entities in the visual display to the syntactic frame—they will associate moveable objects with the direct object position, and goals with the prepositional phrase(s). They then will compare the input to the predicted utterance, revising and editing as necessary, allowing them to execute the action.

However, if the visual situation is more demanding, either because it is not provided in advance or because the visual world contains more than a handful of objects (or both), then comprehenders might rationally change their strategy. As a result, instead of just comparing the actual utterance to the utterance predicted and then tweaking their representation as necessary, they will have to simultaneously identify the objects and goals in the visual world as they are mentioned so they can execute the required action. Due to these task constraints, comprehenders might eventually choose a more passive approach of simply waiting to hear object names and trying to locate those objects once their linguistic labels have been interpreted. This more passive, superficial strategy would not allow for a great deal of deep inferencing or prediction. This is not because the system is architecturally incapable of engaging in these processes, but because the situation will not support them, or because there is no payoff – the listener is likely to be correct too infrequently to justify the effort.

Finally, these four experiments have implications for the current debate concerning the extent to which language processing involves prediction (Altmann & Kamide, 1999; Altmann & Mirkovic, 2009; Engelhardt, Demiral, & Ferreira, 2011; Kamide, Altmann, & Haywood, 2003; Lau, Stroud, Plesch, & Phillips, 2006; Levy, 2008; Sedivy, 2003; Staub & Clifton, 2006; van Berkum et al., 2005). Our contribution is to suggest that certainly comprehenders can engage in prediction, and under some circumstances it will make sense for them to do so because there is time and because it promotes efficient processing. On the other hand, in some situations prediction will not be possible, and these situations are by no means rare or atypical. Thus, again, we suggest that issues concerning language processing strategies should be framed not in terms of whether or not some processing strategy can be used (including prediction), but rather in terms of what the situations would be that would support that strategy or discourage it. In addition, this work highlights the importance of understanding language processing in rich environments as the product of complex interactions between different cognitive systems.

Acknowledgments

The authors would like to thank E. Matthew Husband, Fiona Allan, Manon W. Jones, and Laura J. Speed for the running experiments and the preparing stimuli. We would also like to thank Oliver Stewart, Jens Apel, Alex B. Fine, and Christoph Scheepers for help with the statistical analyses, and three anonymous reviewers for their helpful comments on previous versions of the manuscript. This work was supported by ESRC Grant RES-062-23-0475 awarded to Fernanda Ferreira.

References

Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ: Wiley.

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

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

Altmann, G. T. M., & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and the visual world. In J. M. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action* (pp. 347–386). New York: Psychology Press.

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

Altmann, G. T. M., & Steedman, M. J. (1988). Interaction with context during human sentence processing. *Cognition, 30*, 191–238.

Andersson, R., Ferreira, F., & Henderson, J. M. (2011). I see what you're saying: The integration of complex speech and scenes during language comprehension. *Acta Psychologica, 137*, 208–216.

Baayen, R. H. (2008). *Analysing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for participants and items. *Journal of Memory and Language, 59*, 413–425.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*, 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*, 255–278.

Bates, D., Maechler, M., & Dai, B. (2008). Lme4: Linear mixed-effects models using S4 classes. <http://lme4.r-forge.r-project.org/> (Computer software manual).

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., et al. (2009). Language is a complex adaptive system: Position paper. *Language Learning, 59*, 1–26.

Brown-Schmidt, S., Campana, E., & Tanenhaus, M. K. (2005). Real-time reference resolution by naïve participants during a task-based unscripted conversation. In J. C. Trueswell & M. K. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language as product and language as action traditions*. MIT Press.

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive Science, 32*, 643–684.

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

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

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research, 24*, 409–436.

Engelhardt, P. E., Bailey, K. G. D., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean maxim of quantity. *Journal of Memory and Language, 54*, 554–573.

Engelhardt, P. E., Demiral, S. B., & Ferreira, F. (2011). Over-specified referential expressions impair comprehension: An ERP study. *Brain and Cognition, 77*, 304–314.

Farmer, T. A., Cargill, S. A., Hindy, N. C., Dale, R., & Spivey, M. J. (2007a). Tracking the continuity of language comprehension: Computer mouse trajectories suggest parallel syntactic processing. *Cognitive Science, 31*, 889–909.

Farmer, T. A., Cargill, S. A., & Spivey, M. J. (2007b). Gradiency and visual context in syntactic garden paths. *Journal of Memory and Language, 57*, 570–595.

Farmer, T. A., Fine, A. B., & Jaeger, T. F. (2011). Implicit context-specific learning leads to rapid shits in syntactic expectations. In *Proceedings of the 33rd annual meeting of the cognitive science society (CogSci 2011)* (pp. 2055–2060).

Ferreira, F., & Clifton, C. E. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25*, 348–368.

Ferreira, F., & Tanenhaus, M. K. (2007). Introduction to the special issue on language–vision interactions. *Journal of Memory and Language, 57*, 455–459.

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science, 28*, 105–115.

Henderson, J. M., & Ferreira, F. (2004). *The interaction of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press.

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General, 127*, 398–415.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language, 57*, 460–482.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*, 151–171.

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

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology, 61*, 23–62.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*, 133–156.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research, 46*, 1762–1776.

Kleinschmidt, D., Fine, A. B., & Jaeger, T. F. (2012). A belief-updating model of adaptation and cue combination in syntactic comprehension. In *Proceedings of the 34th annual meeting of the cognitive science society (CogSci 2012)* (pp. 599–604).

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition, 95*, 95–127.

Lau, E., Stroud, C., Plesch, S., & Phillips, C. (2006). The role of structural prediction in rapid syntactic analysis. *Brain and Language, 98*, 74–88.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*, 1126–1177.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*, 676–703.

MacDonald, M. C., & Seidenberg, M. S. (2006). Constraint satisfaction accounts of lexical and sentence comprehension. In *Handbook of psycholinguistics*, pp. 581–611). London: Elsevier Inc..

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science, 31*, 1–24.

Meyer, A. S., Belke, E., Telling, A. L., & Humphreys, G. W. (2007). Early activation of object names in visual search. *Psychonomic Bulletin & Review, 14*, 710–716.

Meyer, A. S., & Damian, M. F. (2007). Activation of distractor names in the picture–picture interference paradigm. *Memory & Cognition, 35*, 494–503.

Morsella, E., & Miozzo, M. (2002). Evidence for a cascade model of lexical access in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 555–563.

Navarrete, E., & Costa, A. (2005). Phonological activation of ignored pictures: Further evidence for a cascade model of lexical access. *Journal of Memory and Language, 53*, 359–377.

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

Oliva, A. (2005). Gist of the scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *The Encyclopedia of neurobiology of attention* (pp. 251–256). San Diego, CA: Elsevier.

Pickering, M. J., & Branigan, H. P. (1999). Syntactic priming in language production. *Trends in Cognitive Sciences, 3*, 136–141.

Potter, M. C. (1975). Meaning in visual search. *Science, 187*, 965–966.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology, 81*, 10–15.

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences, 14*, 411–417.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research, 32*, 3–23.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual interpretation. *Cognition, 71*, 109–147.

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

Spivey, M. J., Tyler, M. J., Eberhard, K. M., & Tanenhaus, M. K. (2001). Linguistically mediated visual search. *Psychological Science, 12*, 282–286.

Staub, A., & Clifton, C. (2006). Syntactic prediction in language comprehension: Evidence from *either…or*. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 425–436.

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition, 36*, 201–216.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632–1634.

Tanenhaus, M. K., Spivey-Knowlton, M. J., & Hanna, J. E. (2000). Modeling thematic and discourse context effects on syntactic ambiguity resolution within a multiple constraints framework: Implications for the architecture of the language processing system. In M. Pickering, C. Clifton, & M. Crocker (Eds.), *Architecture and mechanisms of the language processing system*. Cambridge: Cambridge University Press.

Trueswell, J. C., Sekerina, I., Hill, N., & Logrip, M. (1999). The kindergarten-path effect: Studying on-line sentence comprehension in young children. *Cognition, 73*, 89–134.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language, 33*, 285–318.

van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology Learning Memory and Cognition, 31*, 443–467.

Võ, M. L.-H., & Henderson, J. M. (2011). Object-scene inconsistencies do not capture gaze: Evidence from the flash-preview moving-window paradigm. *Attention, Perception, & Psychophysics, 73*, 1742–1753.

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence comprehension: Statistical learning and relative clause comprehension. *Cognitive Psychology, 58*, 250–271.

Zelinsky, G. J., & Schmidt, J. (2009). An effect of referential scene constraint on search implies scene segmentation. *Visual Cognition, 17*, 1004–1028.