Taking the Epistemic Step: Toward a Model of On-line Access to Conversational Implicatures

Richard Breheny [1]

Heather J Ferguson [1 2]

Napoleon Katsos [3]

[1] University College London, UK

[2] University of Kent, UK

[3] University of Cambridge, UK

Correspondence to:
Richard Breheny email: r.breheny@ucl.ac.uk
Tel: +44 (0) 20 7679 4039

UCL Research Department of Linguistics
Chandler House
2 Wakefield Street
London WC1N 1PF
UK

Abstract

There is a growing body of evidence showing that conversational implicatures are rapidly accessed in incremental utterance interpretation. To date, studies showing incremental access have focussed on implicatures related to linguistic triggers, such as 'some' and 'or'. We discuss three kinds of on-line model that can account for this data. A model built around the notion of linguistic alternatives stored in the lexicon would only account for linguistically triggered implicatures of the kind already studied and not so-called 'particularised' implicatures that are not associated with specific linguistic items. A second model built around the idea of focus alternatives could handle both linguistically triggered implicatures and so-called particularised implicatures but would be insensitive to the role that information about the speaker's mental state plays in deriving implicatures. A third more fully 'Gricean' model takes account of the speaker's mental state in accessing these implications. In this paper we present a visual world study using a new interactive paradigm where two communicators (one confederate) describe visually-presented events to each other as their eye movements are monitored. In this way, we directly compare the suitability of these three kinds of model. We show hearers can access contextually specific particularised implicatures in on-line comprehension. Moreover, we show that in doing so, hearers are sensitive to the relevant mental states of the speaker. We conclude with a discussion of how such a model may be developed and of how our findings inform a longstanding debate on the immediacy of online perspective taking in language comprehension.

Keywords: Conversational Implicature, Eye-movements, Pragmatics, Sentence Processing Perspective-taking, Theory of Mind

Introduction

Conversational implicature is a phenomenon that has attracted much attention since the work of Grice. Grice's contribution was to argue that apparently central components of meaning in language could be explained as not deriving from the conventional or encoded meaning of sentences but as inferences about what the speaker means to convey indirectly, over and above what the sentence means in context (Grice, 1989). One of Grice's examples of indirect communication is the case of the uninformative academic reference: Imagine receiving a reference for a candidate for an academic post which states only that the person in question was always punctual for meetings. You would probably infer that the reference-writer did not say anything about the candidate's academic abilities because she had nothing good to say. You would probably also infer that the reference writer must have intended you to infer this and to see that she so intended. Thus, in an indirect way, the reference writer has communicated her estimation of the candidate without explicitly giving it. Grice's pragmatic theory provides a rational reconstruction of such indirect communication in terms of general expectations we have of each other as communicators. Grice proposed that his theory could account for a wide range of phenomena that had previously been seen as purely linguistic. For example, students would normally understand a teacher who says, 'Some of you got an A on the test' to mean that not all of them got an A. Although this example might not seem at first to involve indirect communication as in the uninformative reference case, Gricean theory would derive the implication in a similar way, as illustrated in Table 1a below. In particular, the implication would be derived by making an inference about the speaker's intentions in the specific situation in which the utterance is produced, relative to mutually assumed expectations of relevance.

Recently, some psycholinguistic evidence has come to light suggesting that the results of Gricean reasoning are accessed in incremental online interpretation. For instance, Sedivy

(Grodner & Sedivy, 2011; Sedivy, 2003) argues that the use of pre-nominal modification (as in (1)) to trigger a contrastive inference before the on-set of the noun (see Sedivy, Tanenhaus, Chambers, & Carlson, 1999) is likely to be the effect of Gricean reasoning.

(1) Pick up the tall glass.

Other studies by Huang and Snedeker (2009, 2011) and Grodner et al. (2010) reveal similar on-line effects.

Another much studied phenomenon is so-called scalar Quantity implicatures, as where the teacher saying 'some' implies not all (see Geurts, 2010). Breheny, Katsos and Williams (2006) provide evidence from reading-time studies that where (2) is understood to imply that not all of the consultants had a meeting with the director, this information is integrated while participants are reading the quantificational constituent:

(2) Some of the consultants had a meeting with the director.

Similar evidence for on-line access to scalar implicatures is reported in Panizza et al. (2009; see also Katsos, Breheny, & Williams, 2005; Katsos, 2008).

While studies such as these reveal the effects of pragmatic reasoning in incremental interpretation, it is yet to be determined how this occurs. Psycholinguistic models tend to be set up to account for comprehension in terms of how meaning is *selected* from a range of alternatives provided by information encoded in linguistic stimuli, possibly augmented by contextual information. For example, lexical or syntactic ambiguities give rise to a decision problem generated simply by competing linguistic representations associated with given forms. As mentioned, according to Grice's account of these implicatures, content is added

over and above what is encoded as conventional meaning. In Grice's theory, implicatures are not associated with any linguistic forms but are derived from the use of a given form by a speaker in a context. So, 'x is always punctual' does not encode that x is not suitable for the job. In fact, in some contexts, it can mean the opposite.

One way in which we may begin to bring implicatures into on-line models is suggested in neo-Gricean theory (Horn 1984, Gazdar 1979, Levinson 2000). According to this theory, there are two classes of implicature: generalised and particularised. A generalised implicature is one which is associated with a specific form of words and, in neo-Gricean theory, is available by default and only withdrawn or 'cancelled' under certain circumstances. The much-studied implicature involving the quantificational determiner, 'some', illustrated in (2), is thought to be a generalised, so-called, 'scalar' implicature. Putting aside details of neo-Grican theory, it seems clear that, from a psycholinguistic perspective, generalised implicatures, such as scalars, could be accessed rapidly on-line in virtue of the existence of linguistic triggers.

Levinson (2000) proposes that only implicatures that have specific linguistic expressions as triggers could be accessed incrementally. Both of the types of implicature illustrated in (1) and (2) are plausibly cases that have linguistic triggers. According to Levinson's theory, quantificational determiners such as 'some' and modifier constructions are associated with alternative constructions in memory. The contrast between the expression used (e.g. 'some') and an alternative construction that was not used suffice to automatically trigger access to the relevant inference. So, the specific prediction of Levinson's theory is that these implicatures are always generated automatically in the right linguistic environments. This prediction has been found to be incorrect in numerous studies, using a variety of reaction time and visual world methods (Bott & Noveck, 2004; Breheny et al., 2006; Grodner & Sedivy, 2011). Specifically is has been shown that even these linguistically

triggered implicatures seem to be only accessed where there is sufficient bias in the context. While Levinson's theory has been found to be wanting, it is still an open question whether such rapid on-line access is mediated through anything other than linguistic-level representations. For example, it is plausible that in previous studies where on-line access has been detected, the implicature trigger is in sentence focus or contrastive topic. Focus and contrastive topic are notions that belong to a level of linguistic analysis known as 'information structure' (see Buring, 2007). According to linguistic theories of focus and contrastive topic (see Krifka, 1999; Rooth, 1993), the marking of a constituent as focus or contrastive topic would trigger the search for alternatives and a presupposition that these alternatives are relevant in context in certain ways. There is on-line research showing that linguistic marking of focus leads to rapid contrastive inferences, as illustrated in (1) (Dahan et al., 2002; Ito & Speer, 2008). Although little research has looked at information structure and scalar implicatures (but see Zondervan, 2010) it is not unreasonable to suppose that on-line access to these implicatures may also be mediated via linguistic marking of information structure.

If we assume that on-line access is mediated by focus or contrastive topic and we adopt neo-Gricean re-interpretation of Gricean theory (Horn, 1984, 2006) we could account for rapid access without the need to involve to any further information. For example, where focus falls on a constituent containing a scalar term like 'some' in a positive linguistic context[1], access to lexical information for 'some' would make available a Horn scale, <all, some>. This scale could be seen a providing a pre-defined set of alternatives to satisfy the presupposition of focus or contrastive topic and would lead to the generation of the 'not all' implicature. Thus, one could imagine an account according to which sentence focus or

[1] The details of which linguistic context allow for Gricean quantity implicatures are outlined in Geurts (2010). These are contexts in which the replacement of the lexical term 'some' by a term higher on the Horn scale (like 'all') would yield a more informative proposition.

contrastive topic together with the Horn scale associated with a lexical items have combined to yield rapid access to the scalar implicature in studies reported up until now.

Let us consider two types of model of on-line access to quantity implicatures based on the above considerations. An account in the spirit of Levinson (2000) would hold that only in cases where linguistically triggered scales are involved are the relevant alternatives sufficiently activated for rapid on-line access to the implicatures. Thus information structural cues and linguistic scales would be necessary and jointly sufficient for on-line access. We will call this the *linguistic scales* account. A second alternative account would suppose that only information-structure cues are necessary for on-line access. These linguistic triggers would prompt the search for alternatives and implicatures could be accessed where context makes relevant alternatives sufficiently salient. On this account, linguistic scales would qualify as sufficiently salient but in principle, any set of alternatives could be made sufficiently salient in context. We will refer to the second account as the *information structure* account.

To get a feel for what the information structure account involves, imagine a situation where you know that a woman had two sets of objects, forks and spoons, and you know that she was placing some of these in either one of two boxes, Box A or Box B. Suppose now a third party sets out to tell you what items she put in which box and says, 'The woman put a spoon into box A and a spoon and a fork into Box B'. It seems normal to assume that this description exhausts what happened. In particular, you would infer that the woman put nothing else into box A. According to Gricean pragmatics, this 'nothing else' inference is explained in the same way as the 'and not all' inference associated with 'some' - as set out in Table 1b below. According to the information-structure account of on-line access to implicatures, the 'nothing else' implication could be accessed rapidly since the phrase 'a spoon' is plausibly contained in the focus constituent and since context makes alternatives to

'a spoon' highly salient. These alternatives are not associated with any lexical item or linguistic construction but would form a partial order with the meanings of, 'two spoons', 'a spoon and a fork', 'two spoons and a fork', and so forth (see Hirschberg, 1985, for an account of these ad hoc scales).

It is important to note that the sentence, 'The woman put a spoon into box A' could potentially be associated with any number of quantity implicatures, depending on context. For example if a speaker was telling you about where the woman put spoons, or who put spoons into box A, then different implicatures are available. Levinson (2000) calls quantity implicatures not associated with linguistics scales Particularised Implicatures. We shall adopt this terminology here.

It is possible to test between the two accounts mentioned above precisely by exploring on-line access to these 'and nothing else' quantity implicatures that do not have a linguistic trigger. The experiment reported below does just that. If we find evidence that Particularised Implicatures are accessed on-line, then the Information Structure account would be supported. Moreover, access to such an 'and nothing else' inference would demonstrate greater contextual flexibility with regards to on-line access than has previously been shown.

Both the linguistic-scales and the information-structure accounts are built on the idea that the primary constraints on on-line access to implicatures are linguistic. They differ only in the number of linguistic factors involved. As suggested above, Grice's original derivation of these implications appealed to reasoning about the speaker in context. Table 1a,b gives Gricean derivations of both linguistically-triggered and particularised quantity implicatures exemplified above and show that information about the speaker is crucial in providing grounds for such inferences.

Table 1:

a. Key steps in the derivation of the *and not all* quantity implicature according to Gricean pragmatics.

I.	The teacher has said that some of the students did well on the test. For all that is said, it could be true that all of the students did well.
II.	However, given (i) that the utterance is telling us about how the students did on the test and (ii) the mutually assumed expectation that the teacher will give as much information as is relevant modulo her own knowledge and preferences...
III.	It would clearly be deficient of the speaker to have said what she did if she had known that all of the students did well.
IV.	So we can conclude that the teacher does not know that all of the students did well.
V.	Given that the speaker knows all about how the students did, we can conclude that not all of the students did well.
VI.	The speaker intends me to reason as above.

b. Key steps in the derivation of the *and nothing else* quantity implicature according to Gricean pragmatics.

I.	The speaker has said that the woman put a spoon into the box. For all that is said, the woman could have put many things in addition to a spoon into the box.
II.	However, given (i) that the utterance is telling me about what the woman put into the box and (ii) the mutually assumed expectation that the speaker will give as much information as is relevant modulo her own knowledge and preferences...
III.	It would clearly be deficient of the speaker to have said what she did if she had known that the woman put other things into the box.
IV.	So I can conclude that she does not know that the woman put other things into the box.
V.	Given that the speaker knows all about what the woman put into the box, I can conclude that the woman did not put other things there.
VI.	The speaker intends me to reason as above.

In particular, according to the Gricean account, both the 'and not all' and the 'and nothing else' implications require the assumption that the speaker is in a position to know the relevant facts (step V). Geurts (2010) calls this the competence assumption; and passing from IV to V in these cases is often referred to as the Epistemic Step (see Chierchia et al., 2008; Sauerland, 2004). Note also that, according to the Gricean derivation, the contextual specificity of the alternatives for quantity implicatures derives from the fact that these are determined relative to the speaker's conversational purpose (step II, Table 1a,b) and that taking the epistemic step involves being aware of what these speaker-defined relevant alternatives would be.

Looking at how Gricean theory derives these quantity implicatures, it seems clear that, compared to a linguistic scales model or an information structural model, a richer model of on-line access based on Gricean theory would build in more constraints reflecting the role of the speaker in these inferences. In the study reported in this paper we will investigate not only whether particularised implicatures are accessed online but also whether there is any evidence that access proceeds via the steps described in Table 1. In particular, we will look at the Epistemic Step (step V). If we find evidence that access to particularised implicatures is modulated by the speaker's knowledge about the relevant alternatives, then we will have motivation for a more thoroughly 'Gricean' model of on-line implicatures.

For the study below, we developed a new interactive paradigm where communicators watch short videos on separate computer monitors and then describe the events to each other. On each trial, a still image of the last frame of the video is on the hearer's screen while the speaker describes the events in the video. Hearers' eye movements around the visual scene are monitored and time-locked to related auditory input to examine the anticipation of forthcoming target referents. Thus our method applies the visual-world paradigm (Cooper, 1974; Tanenhaus et al., 1995) as it is developed in the 'look and listen' studies of Altmann and Kamide (1999; see also Altman, & Kamide, 2007; Ferguson, Scheepers, & Sanford,

2010; Kamide, Scheepers, & Altmann, 2003). In this paradigm, participants are simply asked to listen to what is being said while watching a visual display. It has been found that participants' anticipatory looks to objects in the visual display can be directed by the auditory input reflecting the cognitive processes that underlie language comprehension. Our experiments extend this paradigm by testing whether Gricean pragmatic inferences influence anticipatory looks.

In all of the video items (including fillers) shown to participants, an agent is shown from the shoulders down behind a table containing two opaque boxes (labelled 'A' and 'B'), as in Figure 1 below. Between these boxes are two sets of objects (for example, three spoons and three forks). The videos depict the agent transferring some of these objects into the boxes. For example, a video could show the agent transfer a spoon into box A and then a fork into box B. In all of our video items, the focus is on the agent transferring objects into boxes and so any description of such a sequence of events should say what was put into the different boxes. Thus, our items provide a context for deriving an 'and nothing else' quantity implicature as set out in Step II of our inference schema in Table 1b above.

In one of our critical conditions, the video shows the agent put an object of one kind into one box (e.g. a spoon into box A), then one of the same kind of object into the other box (a spoon into box B) and then one object of the second kind into one of the boxes (e.g. a fork into box A). To provide an adequately informative description of the events in the video, the speaker could use or (3) or (4) among other possibilities:

(3) The woman put a spoon and a fork into box A and a spoon into box B.

(4) The woman put a spoon into box B and a spoon and a fork into box A.

Notice that for both (3) and (4), there is an 'and nothing else' implicature available if it is assumed that the sentence in question is being used to describe what happened in the video. In the case of (4), it is implicated that the woman put nothing else into box B, among other things. Focussing on (4), observe that once the speaker has uttered 'The woman put a spoon into...', having seen the video, one can predict the completion of the first clause ('...box B') by making a quantity implicature inference. To see why, note that the contexts of our items get us to Step II in table 1b. When the speaker produces the directional preposition 'into', the hearer can assume that she has exhausted mention of the list of objects going to the destination in question (about to be mentioned in the prepositional phrase).[2] Assuming further that the speaker is being suitably informative (Step III), we get to Step IV, concluding that the list of objects mentioned prior to 'into' exhausts the speaker's knowledge on the topic of what went to the destination about to be mentioned. But, assuming the speaker knows what was put where, we conclude nothing else was put in the destination about to be mentioned (Step V).

Using the look and listen paradigm, items like (4) in the context of the appropriate video will allow us to determine whether participants are able to access the particularised quantity implicature in incremental interpretation. We predict that if the implicature is accessed soon after the point where the direct object noun phrase is completed and the preposition begins, participants' looks around the visual scene will be biased to the target destination before the critical point of disambiguation (the letter 'A' or 'B'). We tested this prediction in the 'full disclosure' or 'open' condition in the study below.

If participants do predict the correct target before the point of disambiguation in the utterance, we can ask whether this is in virtue of applying the constraints as set out in the

[2] This assumption is not necessary, of course, since the description could take the form of conjoined VPs ('put a spoon into box A and put a fork into box A...') or even conjoined sentences ('The woman put a spoon into box A and she put a fork into box A,...') among other possibilities. However, given the results we report below it is most likely that our participants did tend to expect conjoined noun phrases, as in (3) and (4). Moreover, in the practice trials before the main study, participants themselves used this form of words exclusively.

information structure account or whether we need a richer model based more closely on steps I-VI above. In our study, there is also an ignorance condition in which the speaker only sees the first part of the video. In this condition, it is common ground between speaker and hearer that the speaker does not see all of the events in the video. For example, a participant watches the same video as described by (3) or (4) but also knows that (it is common ground that) the speaker only saw the video up until the point before the actor puts the fork into one of the boxes. If the speaker describes what happened as well as she can, she might utter (5) or (6):

(5) The woman put a spoon into box A and a spoon into box B.

(6) The woman put a spoon into box B and a spoon into box A.

In either case, according to the Gricean derivation, hearers will not be able to go beyond step IV in Table 1b above and so will not be able to access the relevant *and nothing else* implicature. So, Gricean pragmatics predicts that any effect of access to the *and nothing else* implicature in the 'open' condition described above will disappear in this ignorance condition. By contrast, a model that only looks at information structure and salience of alternatives would not predict any modulation of the effect in the ignorance condition.

Method

Participants

Forty participants from the University of Cambridge were paired with one of two confederates for the experimental testing. These two confederates were also recruited from the Cambridge student population and were blind to the purpose of the experiment. All were paid to participate in the study.

Stimuli and Design

Forty sets of experimental videos and pictures were paired with an auditory passage in one of five conditions. Video clips were recorded in a single session involving one male and one female 'actor' and edited using Adobe Premier. Subsequent pictures were created from the final frame in the video clips. All visual images were presented on a 17 inch colour monitor in 1024 x 768 pixels resolution.

Three different video scenarios were set up as a context for the subsequent visual scene and auditory description. All video scenarios began with two classes of objects in the centre of a table (e.g. spoons and forks) and two possible target locations (opaque boxes labelled A and B). In the first scenario (XY), the actor moved an object into one of the boxes, paused, then moved an object of the other type into the other box (e.g. a fork into box A and a spoon into box B). The second video scenario (XX) depicted the actor putting an object into one of the boxes, pause, then move an object of the same type into the other box (e.g. a fork into box A and box B). In the third video scenario (XXY), the actor first moved an object into one of the boxes, paused, then moved an object of the same type into the other box, paused, then moved an object of the other type into one of the boxes (e.g. a fork into box A and a fork and spoon into box B). Subsequent screen images then depicted the final state from each of these scenarios (i.e. the two closed boxes and remaining objects), and were created by extracting the final frame from each video clip. Systematic viewing strategies were prevented by counterbalancing the spatial arrangement and order of movement of the object across items.

Initially the two communicators completed an extensive practise block (10 trials), in which each experienced the role of speaker and listener, including trials that involved being an ignorant speaker and a hearer when the speaker was ignorant. Subsequently the communicators were 'randomly' assigned roles for the rest of the experimental session: a

'speaker' (always the confederate) who described events in the video, and a 'listener' (the participant) who listened to the speaker's description while viewing the final image from the video.[3] The monitors were arranged so that neither person could see the other's screen.

To create ignorance trials, we manipulated the completeness of the knowledge held by the speaker about the XX and XXY videos by using a screen to obscure the speaker's (but not the listener's) view at a critical point of the video. In this way, we created X/X and XX/Y trials (note that '/' represents the point in the video at which the speakers' view was blocked). This manipulation established different levels of knowledge across the five experimental conditions (fully informed *vs*. under-informed).

The confederate speaker was asked to describe the events in the video only as far as they knew. Under these conditions, in the X/X scenario, the speaker's limited knowledge will only enable them to describe the first part of the video, meaning that the listener should be able to accurately predict the target referent from the earliest opportunity. In the XX/Y scenario, since the speaker did not see the last part of the video, their description will only include the XX information, meaning that participants could not go beyond step IV of the implicature derivation and so be unable to anticipate the appropriate target before the point of disambiguation ('A' or 'B'). Auditory descriptions were scripted to ensure consistency of the object names and description types for analysis and were constructed to answer the question, "what did the man/ woman do?". Table 2 provides example descriptions for each experimental condition.

Table 2:

Examples of experimental sentences.

[3] Note that each was given further turns at the other's role at two 'switch' points during the experiment.

XY	The woman put a fork into box A and a spoon into box B.
XX	The woman put a fork into box A and another fork into box B.
X/X	The woman put a fork into box A.
XXY	The woman put a fork into box A and a fork and a spoon into box B.
XX/Y	The woman put a fork into box A and another fork into box B.

One version of each item was assigned to one of five presentation lists, with each list containing forty experimental items, eight in each of the five conditions. In addition, thirty-six filler items were added to each list. Of these, twenty depicted events that elicited a false belief in the speaker (elicited by switching the location of the transfer object while the screen was up), and sixteen depicted events that were shared by the two communicators. All involved a transfer action. Importantly, these fillers included 16 trials with three-event videos, which rotated the order of events (XYX, YXX and XXY). Eight of these fillers were 'full disclosure' trials where the speaker and listener both witnessed the full video sequence, but auditory descriptions mentioned the box containing two objects first (e.g. "The woman put a lipstick and a cup into box A and a lipstick into box B"). In the other eight three-event fillers the speaker's screen was blocked part-way through the video sequence (i.e. before the last transfer event), creating XY/X or YX/X items to counterbalance the XX/Y conditions.

To clarify, the purpose of these filler items was to counterbalance the order of events and descriptions to eliminate any low-level contingencies between videos and descriptions. The resulting lists contained as many three-event items where the conjunction is mentioned first as second. In addition, overall, there are as many items where the first-occurring transfer event is not mentioned first as there are items where the first-occurring transfer event is mentioned first.[4]

The filler items were interspersed randomly among the forty experimental trials to create a single random order. Each subject only saw each target video once, in one of the five

[4] Please contact Heather Ferguson (h.ferguson@kent.ac.uk) for further information on these filler items.

conditions. Binary-choice comprehension questions followed half of the experimental and half of the filler trials and both the listener (participant) and speaker (confederate) responded to these questions (see appendix). All participants scored at or above 80% accuracy on the comprehension questions, and the average score was 94%.

Procedure

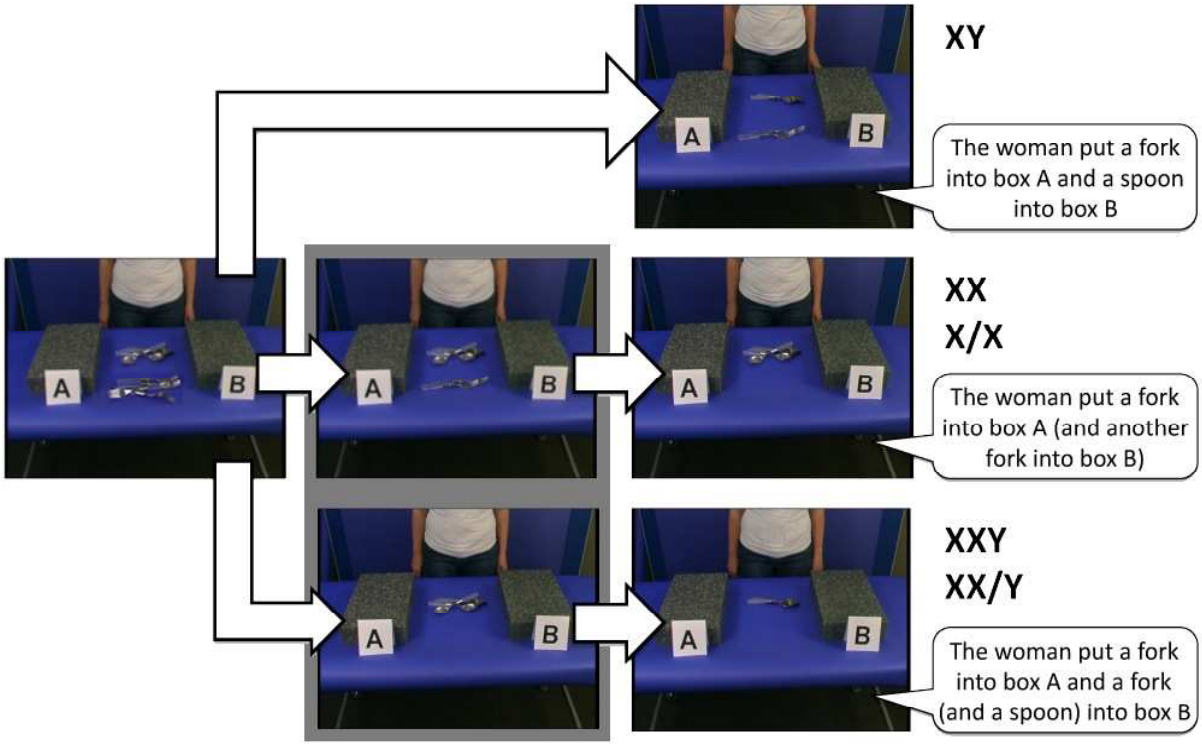
Prior to the main testing period, both confederates completed training in a pilot test so that they fully understood the experimental procedures. Note that auditory descriptions were scripted to ensure consistency of the object names and description types for analysis, and were presented to speakers using Powerpoint software. During the experiment, participants sat in front of a colour monitor while eye movements were recorded using a stand-alone eye tracking system (Tobii X120) running at 120 Hz sampling rate. Viewing was binocular and eye movements were recorded from both eyes simultaneously. The confederate sat at a separate monitor and spoke into a microphone, which recorded audio responses to file.

The experiment was controlled using e-Prime (Schneider, Eschmann, & Zuccolotto, 2002). Each trial began with the presentation of a single centrally-located cross and participants were asked to fixate it for 1000msec before the trial was initiated. At this point, the cross was replaced by the video depicting a transfer scenario. Note that on X/X and XX/Y trials, a message appeared on both monitors part-way through the video instructing the experimenter to put the "Screen up", thus covering the speaker's (confederate) screen during the second part of the video. Both confederate and participant pressed a button to continue when prompted by the experimenter. Video clips lasted on average 25 seconds (range = 19s to 34s) in total and were followed by a pause/ blank screen for 500ms. Next, the corresponding picture was presented as the confederate gave a spoken description of events. Confederates were prompted to begin their description when a 'beep' was heard, 500ms after

picture onset. This ensured that the onset of the picture preceded the onset of the corresponding spoken description by at least 500ms. The picture stayed on-screen until the description was finished and both confederate and participant pressed a button to move on. Note that, in the X/X and XX/Y trials, it is shared knowledge that the speaker will describe what they saw prior to their video monitor being blocked from view.

Figure 1:

Schematic trial sequence of visual displays presented to participants for each condition. Stage 1 (far left) depicts the 'start state' for all videos. During the video, some of the objects in the centre were moved into the boxes. However, note that on X/X and XX/Y trials, the final part of the video was only seen by the participant (the listener), thus establishing a discrepancy in the accuracy of the knowledge held by the speaker. Thus, stage 2 (middle) depicts the final state according to the speaker on these trials. Stage 3 (far right) shows the final state image that participants saw while they listened to their partner's description of events for each video condition.



At the beginning of the experiment, and once every ten trials thereafter, the eye-tracker was calibrated against nine fixation points. This procedure took about half a minute and an entire session lasted for about an hour and a half. After the experiment, we tested the participants' belief in the confederate and experimental design by asking them to rate how strongly they agreed or disagreed with four statements (on a scale of 1-7, with 7 being 'strongly agree'):

1. My partner in the experiment gave me accurate descriptions of the videos, as far as their knowledge allowed;
2. Apart from when the screen was up, I believe that my partner was watching the same video clips as me;
3. I believe that my partner could not view the videos when the screen was covered;
4. I believe that my partner was a real participant.

No participants responded below 5 on any of these questions; mean responses were 6.82 (SD = 0.4) for question 1, 6.79 (SD = 0.8) for question 2, 7 (SD = 0) for question 3, and 6.93 (SD = 0.4) for question 4.

Results and Discussion

Data Processing

Eye-movements that were initiated during the auditory description were processed according to the relevant picture and word onsets for the purpose of aggregating the location and duration of each sample from the eye tracker. For analysis, we removed any sample that was deemed 'invalid' due to blinks or head movements. The spatial coordinates of the eye movement samples (in pixels) were then mapped onto the appropriate object regions; if a fixation was located within 20 pixels around an object's perimeter, it was coded as belonging to that object, otherwise, it was coded as background. Finally, temporal onsets and offsets of the gazes were recalculated relative to the corresponding picture onset.

Probabilities of fixating the Target or 'Alternative' box as a function of time were analysed using the log-ratio measure (see Arai, van Gompel, & Scheepers, 2007): log(Target/Alternative) = $ln(P_{(\text{Target})} / P_{(\text{Alternative})})$. Here, $P_{(\text{Target})}$ refers to the probability of fixating on the target box (i.e. the box that was *actually* described first) and $P_{(\text{Alternative})}$ to the probability of fixating on the alternative box; *ln* refers to the natural logarithm. The output is therefore symmetrical around zero such that a positive score reflects higher proportions of fixations on the target box and a negative score reflects higher proportions of fixations on the alternative box; a score of zero indicates equal bias towards the two boxes.

Log(Target/Alternative) scores were analysed for five consecutive critical regions, determined according the onsets and offsets of words in the corresponding auditory input. These word-regions were identified and synchronised for each participant on a trial-by-trial

basis, relative to the onsets and offsets of relevant words in the appropriate item-condition combination. Note that for all analyses, these word regions were offset by 200ms to allow for the time it takes to program and launch an eye-movement (see Hallett, 1986). The resulting average word-region durations are detailed in Table 3. Importantly, none of these word lengths differed significantly across the five conditions (*F*s < 0.8). Figure 2 plots the average log(Target/Alternative) data for each condition, for every 20 ms time-slot. Eye movements and auditory input have been resynchronized according to individual word onsets (see Altmann & Kamide, 2009), and as such represent more accurate plots of evolving visual biases around the scene.

Table 3:

Average word durations for each condition (timings in ms).

	XY	XX	X/X	XXY	XX/Y
[Object]	599	578	601	596	601
into	228	220	229	237	228
box	284	283	297	292	291
[pause]	99	100	96	103	95
A/B	403	412	407	399	394