This is the author accepted manuscript of: Barr, D. J., Sirniö, H., Kovács, B., O'Shea, K. J., McNee, S., Beith, A., Britain, H., & Li, Q. (2025). Perspective conflict disrupts pragmatic inference in real-time language comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition . Advance online publication. https://doi.org/10.1037/xlm0001455

Perspective conflict disrupts pragmatic inference in real-time language comprehension

Dale J. Barr*[1], Hanna Sirniö[1], Beáta Kovács[1], Kieran J. O'Shea[2], Shannon McNee[1], Alistair Beith[1], Heather Britain[1], and Qintong Li[1]
[1]School of Psychology and Neuroscience, University of Glasgow, UK
[2]Department of Psychological Sciences and Health, University of Strathclyde, UK

Abstract

In two visual-world eyetracking experiments, we investigated how effectively addressees use information about a speaker's perspective to resolve temporary ambiguities in spoken expressions containing prenominal scalar adjectives (e.g., *the small candle*). The experiments used a new "Display Change" task to create situations where an addressee's perspective conflicted with that of a speaker, allowing the point of disambiguation (early versus late) to be specified independently from each perspective. Contrary to existing perspective-taking theories, the only situation in which addressees resolved references early was when *both* perspectives afforded early disambiguation. When perspectives conflicted, addressees exhibited a lower rate of preferential looks to the target and slower response times. This disruption to contrastive inference reflects either the suspension of pragmatic inferencing or cognitive limitations on the simultaneous representation and use of incompatible perspectives.

Keywords: common ground, pragmatics, visual-world eyetracking, perspective taking

Modern psycholinguistic theories portray language comprehension as an active, efficient, and sophisticated system for decoding meaning from incoming linguistic signals. Two assumptions underpin this view: first, that comprehension informs, and is informed by, a working discourse model (Crain & Steedman, 1985; Zwaan & Radvansky, 1998); and second, that the processing of speech input is highly incremental, meaning that comprehension does not passively wait to assign meaning to structures until they are linguistically complete, but instead actively entertains semantic candidates consistent with the unfolding input (Allopenna et al., 1998; Altmann & Kamide, 1999; Eberhard et al., 1995). These two assumptions are intertwined, inasmuch as it is a discourse model that empowers addressees

*Corresponding author: Dale J. Barr, School of Psychology & Neuroscience, University of Glasgow, 62 Hillhead Street, Glasgow, UK, G12 8QB. Email: dale.barr@glasgow.ac.uk. The raw data, analysis code, materials, and pre-registrations are available through the OSF repository (Barr, Britain, & Li, 2024). Thanks to Camilo Ronderos for comments on an earlier draft of this manuscript.

to swiftly lock onto intended meanings in the face of incomplete input.

To participate successfully in real-world conversational interaction it is not enough to have just any discourse model; critically, one must have a discourse model that is aligned with that of one's conversational partners in certain critical respects (Pickering & Garrod, 2004). From the point of view of comprehension, this means that addressees must have some level of awareness about aspects of a speaker's perspective that are uncertain or that directly conflict with their own. In this paper, we use visual-world eyetracking (Cooper, 1974; Tanenhaus et al., 1995) to investigate how addressees use information about a speaker's conflicting perspective to resolve temporary ambiguity related to the meaning of prenominal scalar adjectives (e.g., the word *large* in definite descriptions such as *the large candle*).

It is widely accepted that the way people speak and understand is shaped by their "common ground" (Clark & Marshall, 1981)—the set of assumptions, suppositions, and beliefs that interlocutors believe they hold in common. However, the impact of common ground on comprehension as it unfolds in real time remains a controversial topic. Early theoretical work predicted that common ground would have a powerful role in reducing the set of linguistic structures and interpretations that must be considered during language processing (Clark & Carlson, 1981). Instead, the first visual-world studies on this topic suggested that processing was fundamentally egocentric and that addressees optionally use common ground to monitor and correct the unfolding understanding (Keysar et al., 2000). Later studies called this view into question, detecting effects of common ground at the earliest moments of comprehension (Hanna et al., 2003; Nadig & Sedivy, 2002). A recent view assumes that addressees simultaneously integrate information from their own and the speaker's perspective, with the weights given to each perspective influenced by the communicative situation (Heller et al., 2016). But this view seems incompatible with findings that pragmatic-level factors exert little or no constraint on lower-level processes such as lexical activation (Barr, 2008b).

These conflicting views may co-exist because the evidence base for perspective taking in spoken language comprehension is largely built on visual-world studies looking at a certain type of perspective difference—one of *embedded perspectives*—in which the information available to the speaker is a proper subset of the information available to the addressee. An embedded perspective situation is one in which an addressee has so-called "privileged" information that a speaker lacks. Usually the perspective difference is created through visual occlusion of objects from the speaker's view (e.g., Keysar et al., 2000). In this situation, addressees must keep track of what the speaker does not know, and ignore this information when interpreting their speech. Embedded perspectives may give rise to the *curse of knowledge*, a phenomenon in which judgments are contaminated by private information (Birch & Bloom, 2007). At the same time, awareness of such asymmetries in perspective are likely to play motivational roles for communication in the first place, leading to information-seeking or information-providing behavior that brings perspectives into greater alignment.

Another type of perspective difference that has been studied less often involves *perspective conflict*, where there are incompatible interpretations of the same situation: *It's not that you know something that I don't know, it's that you and I believe something entirely different.* In contrast to psycholinguistics, where studies of perspective conflict are rare, conflicting perspectives are the norm in developmental research, where a child's abil-

ity to deal with such conflicts provides benchmarks for social development (Flavell et al., 1983; Perner et al., 1987; Wimmer & Perner, 1983). Conflicting perspectives require people to *substitute* another's belief when reasoning about their behavior rather than *suppressing* privileged information. Reasoning about another's perspective in the face of conflict might call upon greater imagination and cognitive flexibility than merely suppressing information not known to one's interlocutor (Westra & Nagel, 2021). Alternatively, perceiving a conflict might actually improve reasoning if the irreconcilable difference makes the distinction in perspectives more salient; indeed a recent theory suggests that mentalizing is oriented toward detecting such conflicts (Deschrijver & Palmer, 2020). In sum, to attain a full understanding of how perspective-taking works, we need to study conflicting as well as embedded perspectives.

To our knowledge, to date there are only three published visual-world experiments involving perspective conflict. However, these experiments have yielded inconsistent findings. Two of these involved addressees witnessing events contrived to make it seem that a (confederate) speaker had false belief about an object's identity. In Experiment 2 of Keysar et al. (2003), addressees were led to believe that a speaker had a false belief about the identity of an occluded object. This false-belief condition was contrasted with an "ignorance" condition where speakers were unaware of the identities of hidden objects. Addressees made errors and experienced egocentric interference in both conditions, with no evidence for any differences across conditions. In contrast, Experiment 2 of Hanna et al. (2003) did find evidence that addressees were able to take a speaker's differing perspective into account. In their study, addressees were led to believe that speakers had false information about the identity of an object after an experimenter mislabelled the object for the (confederate) speaker, who then repeated the mislabelling. This created a situation where addressees could resolve the target earlier if, when interpreting the speaker's instructions, they relied on the mislabelled perspective rather than the information perceptually available to themselves. The eye data supported this prediction.

Mozuraitis et al. (2015) examined the processing of referential expressions in visual arrays containing critical items whose appearance mismatched their true functional identity (e.g., an object that looks like a lightbulb but is actually a candle). This visually confusing object played the role of a potential phonological competitor, given that its functional name (*candle*) overlapped phonologically with the name of a target referent (e.g., *pick up the candy...*). There was no evidence that lexical competition was reduced when knowledge of the object's true identity was not in common ground with the speaker. The experiment contrasted this appearance-reality mismatch situation with a visual occlusion situation where a phonological competitor (either the lightbulb-shaped candle or an actual candle) was occluded from the speaker's view. Their analysis suggested reduced interference in the visual occlusion situation than in the appearance-reality mismatch situation. The competitors in this situation still attracted looks, but at a lower rate.

Creating arbitrary differences in perspective with the Display Change Task

A likely reason for the predominance of visual-world studies using embedded perspectives is the ease of generating such differences through visual occlusion. In this paper, we introduce the Display Change Task, which makes the generation of conflicting perspectives similarly easy, by exploiting temporal rather than spatial disparities in scene perception

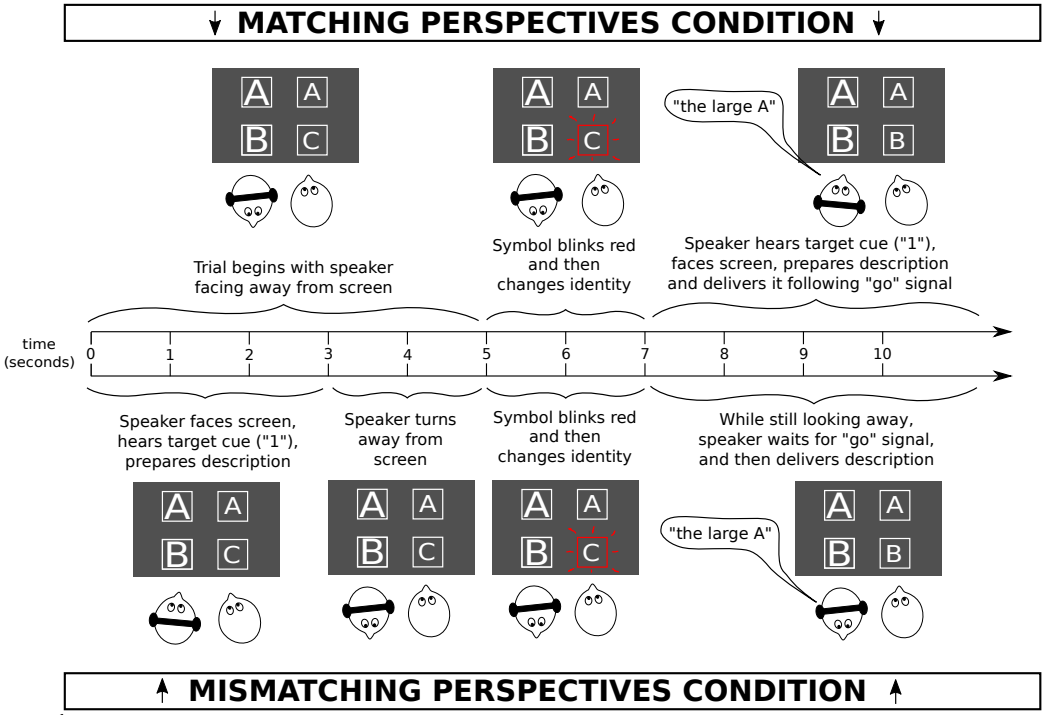

Figure 1

Illustration of how the Display Change task can be used to create matching or mismatching perspectives. Schematic "top" views of the confederate speaker (wearing headphones) and addressee (no headphones) facing toward or away from the screen at various stages of the trial. The middle of the image shows a trial timeline in seconds. Above the timeline are events in the Matching Perspectives condition; below are events in the Mismatching Perspectives conditions. The Matching Perspectives condition gives an example of the Speaker Late / Addressee Late condition, while the Mismatching Perspectives condition gives an example of the Speaker Early / Addressee Late condition.

(see Figure 1 for a schematic overview). In a referential communication task, a speaker and addressee view images on the same computer screen. The basic task is identical to that of a standard referential communication task: a speaker describes an target object that appears in an array of objects, and the addressee's task is to identify the target based on that description, indicating the selection by clicking on it with a computer mouse). The addressee's eye movements may be recorded while performing the task, while also recording response times. The speaker is secretly informed of the location of the target, and she silently prepares her description of it. She then looks away from the display and waits for a signal to speak her prepared description aloud. While she is looking away, the addressee witnesses a single object change its identity. Because the addressee believes the speaker planned her description without knowledge of the change, the interlocutors' common ground consists of the set of objects in pre-change display. Because the changed object itself may attract attention, it can also be useful to include a matching perspectives condition as a control, where the change occurs *before* the speaker looks at the screen (Figure 1). The critical

question is the extent to which response times and eye movements reflect consideration of information in the shared perspective versus privileged information available only to the addressee.

This approach has intrinsic advantages over other approaches for creating conflict in that it does not require deception nor finding objects with an appearance-reality mismatch. It also has an inherent advantage over the use of visual occlusion to study perspective taking in that the spatial location of the changed object still contains a viable referent, and so addressees must still allocate attention to it. This may help remedy the interpretive problems arising in studies that use visual occlusion (Barr, 2008a; Barr et al., 2011).

The Display Change Task works because people are highly attuned to the social nature of attending. Stimuli that are collectively attended are more likely to be attended and remembered than stimuli encoded individually (for review, see Shteynberg, 2015). One way of looking at the perspective discrepancy created by our task is that it gives the addressee reason to believe that the speaker has a false belief about the identity of any changed object or objects. In this respect, it is like the classic "Maxi" false belief task, where a change occurs that the protagonist of a story does not witness (Wimmer & Perner, 1983). The difference is that the event is embedded in a referential game, and the "protagonist" is the participant's partner in a referential communication task. Thus, the participant needs not only to keep track of the speaker's outdated belief, but also to reason about how holding such a belief would affect any descriptions the speaker might produce.

Overview of the current experiments

In the following two visual-world experiments, we used the Display Change Task with visual-world eyetracking to investigate perspective taking in contrastive inference, following the early versus late disambiguation approach of Heller et al. (2008) and Eberhard et al. (1995). Participants played the role of addressee and viewed sets of objects on a computer screen while interpreting descriptions from a research assistant who played the role of speaker. A star was hidden behind one of the images on the screen, and the speaker had to convey its location by describing the image that covered it (the "target" image). Addressees' eyes were tracked while they interpreted these descriptions. Critical descriptions used the prenominal scalar adjectives "small" and "large" (e.g., *the small A* or *the large candle*). Speakers prepared their descriptions while viewing displays containing either one or two size-contrasting image pairs, allowing for early or late disambiguation respectively. Independently of the speaker's perspective, we manipulated the number of size-contrasting image pairs visible to the addressee while they interpreted these descriptions. By doing this, we fully deconfounded the speaker's and the addressee's perspectives, making it possible to estimate their independent influences on the unfolding interpretation.

Scalar adjectives such as *small* or *large* locate items along a continuous dimension; however, the parameters of the dimension depend critically on the semantics of the noun they modify (e.g., a 'big ant' is smaller than a 'small planet', Kamp & Partee, 1995). Although this would seem to require addressees to wait until the head noun to determine the meaning of the adjective, its meaning can also be fixed by the alternatives in the immediate referential context (Sedivy et al., 1999). When addressees view a visual array including a *contrast set*—a pair of objects from the same category contrasting in size—and hear size-modified descriptions like "the large glass," they look at the target earlier than when the

contrast is absent. Thus, addresses calculate contrast sets on the fly and make rapid use of this information to enrich the meaning of scalar adjectives and resolve ambiguity.

Heller et al. (2008) investigated whether this contrastive effect depended upon the addressee's common ground with the speaker. Addressees viewed displays with one or two pairs of objects contrasting in size—for example, a small and large toy duck and a small and large box. The target object was a member of one of these pairs (e.g., the small duck). The key manipulation concerned the privileged or shared status of one of the members of the other pair, the *critical object* (e.g., the large box). In the privileged condition, this object was hidden from the speaker's view by a visual occluder. Thus, although the addressee would know about both contrasting pairs, the speaker only knew about one of them—namely, the one containing the target. When the critical object was privileged, addressees could therefore identify the target based on the size adjective alone if they used information from the speaker's perspective. In the shared condition, the critical object was visible to the speaker. Thus, in this condition, the speaker would know about both contrasting pairs, and addressees would have to wait for the head noun to resolve the reference. During processing of the adjective, addressees showed a higher likelihood of looking at the target in the privileged condition than in the shared condition, supporting the prediction that they could use information from the speaker's perspective to resolve the reference early. A later follow-up study by Heller et al. (2016) yielded similar findings.

Addressees seem able to account for a speaker's differing perspective to resolve temporary ambiguity when the speaker's perspective is embedded in their own—but what happens when perspectives conflict? To investigate this question, we present two experiments using the Display Change Task, which allows for complete independence in manipulating the point-of-disambiguation for each interlocutor. This allows us to add the novel condition in which disambiguation is *early for the addressee, but late for the speaker*, a situation that would be challenging to create using visual occlusion. Fully separating the two perspectives in a factorial design makes it possible to directly test the proposal that addressees integrate both perspectives in a simultaneous, weighted fashion (Heller et al., 2016). This view predicts that the speed with which addressees resolve scalar ambiguity should be fastest when the point of disambiguation is early for both parties, intermediate when it mismatches (i.e., early from one perspective and late from the other), and slowest when it is late from both perspectives. Furthermore, the extent to which the two "mismatching" perspectives conditions differ from each other would give insight into the relative weighting addressees gave to their own versus the other's perspective.

We investigated contrastive inference across two experiments, which differed in two ways: (1) the stimuli used and (2) the way the way in which we manipulated the speaker's perspective. The first experiment involved typographic symbols varying in size, and we manipulated speakers' perspectives by having them formulate target descriptions while looking at either the pre-change display or the post-change display. By manipulating the pre- and post- change displays, we could independently manipulate the number of size-contrasting pairs the speaker and the addressee saw. In the second experiment, we used pictures of everyday objects and included a perspective conflict in every condition. The main difference was whether the changed image also made a difference to the number of size-contrasting sets from either perspective.

To ensure consistent speaker behavior and streamline the logistics of data collection,

the role of the speaker in both experiments was played by confederate research assistants. We took steps to minimize the impact of this feature on the ecological validity of the study (Kuhlen & Brennan, 2013). We introduced the speaker by name and did not say anything about their affiliation to the lab (e.g., "This is Shannon, she is going to help out by playing the role of speaker." We did not script the speaker's descriptions—they participated in the task just like any other participant would, and had to come up with descriptions of the target on the fly. Confederate speakers knew that the only type of modifiers needed to distinguish the target would be size modifiers, and that they should use them prenominally, but only when required. In other words, confederates always strove to provide optimally informative descriptions from the shared perspective. We did not place any constraints on what size adjectives they used, although the speakers almost invariably used 'small' and 'large'. Second, although the speakers generally knew about the experimental hypotheses and the purpose of the study, they had no way of knowing what condition any particular trial was in, as they never knew what object changed identity when they were looking away. In short, they had to perform the task just like a naïve speaker would, and their beliefs about the situation were consistent with the beliefs a naïve speaker would have had in that role.

Analyses of response time and eye gaze data in Experiment 1 suggested that addressees only disambiguated references at the adjective when both their own and the speaker's perspective contained a single size-contrasting pair. In other words, addressees did not resolve references early when this possibility was offered by their own or the other's perspectives—only when both did. Moreover, addressees preferentially looked at the target at a lower rate during the adjective and also were slower to select it when perspectives mismatched compared to when disambiguation was late from both perspectives. However, interpretation of these results was clouded by potential confounds, inasmuch as the mismatching perspectives conditions differed from the matching perspectives conditions in terms of both cognitive load and joint attention.

Experiment 2 removed the confounds by eliminating the matching conditions, such that all trials took place under mismatching perspectives, with the speaker always facing away from the screen when delivering the critical utterance. We manipulated the speaker's point of disambiguation for the speaker by manipulating whether the changed object was part of a contrasting pair. This required the addressee to track a perspective difference across all conditions. If the results of the previous study were attributable to cognitive load, then we should find late resolution of the target in all conditions, including the condition where both speaker and addressee view only a single size-contrasting pair. Instead, we replicated the pattern from Experiment 1: disambiguation took place during the adjective only when both the speaker's and addressee's perspectives contained a single size-contrasting pair. Once again, relative to the late-matching condition, in the mismatching conditions preferential looks to the target were lower during the adjective and response times were delayed.

None of the current theoretical approaches to perspective taking in language comprehension can easily explain this disruption to contrastive inference when perspectives conflict. In the General Discussion, we explore two possible explanations for this result. Under one explanation, when addressees detect conflicts in perspective that are pragmatically consequential, they adopt a metacognitive strategy to suspend contrastive inference. This

possibility assumes that addressees have a sufficient level of control over the incrementality of the comprehension system to be able to override a default, 'greedy' mode of pragmatic inferencing when the communicative environment is perceived as uncertain. Alternatively, the disruption could reflect cognitive limitations related to the ability to simultaneously represent and use information from incompatible perspectives.

Experiment 1

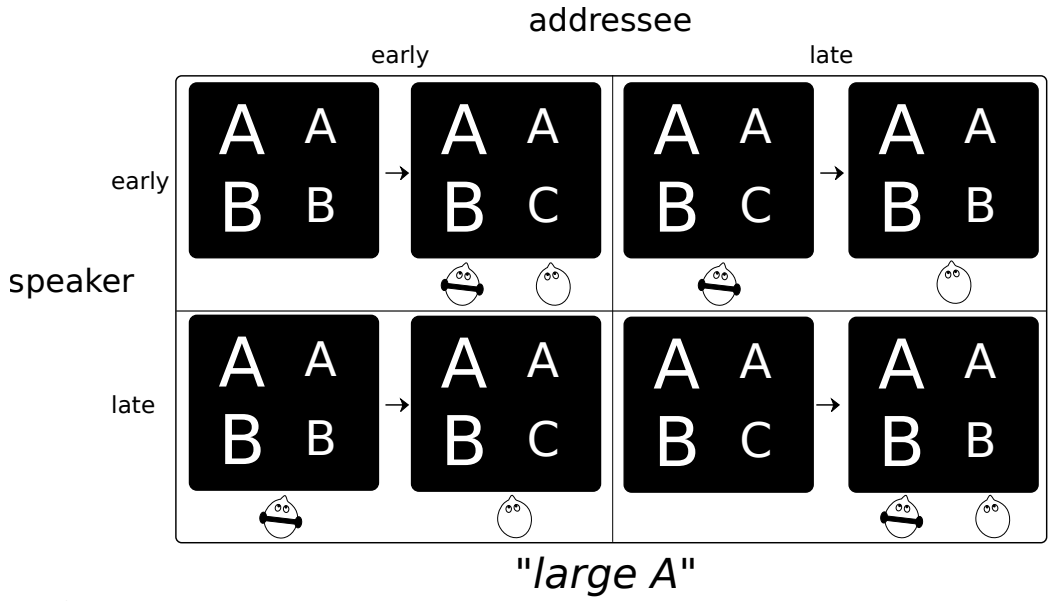

Figure 2

Factorial design of Experiment 1, showing a single stimulus display across all four conditions, where the utterance is "large A." The speaker is represented by the schematic head that is wearing headphones, and the addressee is the head without them.

In the first experiment, participants saw four typographic symbols on the screen varying in size, with a star hidden behind the target. The speaker (who was a research assistant) was informed of the location of the target through a private headphone, and had to convey its location to the addressee by describing it. The addressee selected the target by clicking on it with a computer mouse.

On critical trials across all conditions, there was always one size-contrasting pair in common ground, one member of which was the target. What differed across critical trials was whether or not the two remaining images were contrasting in size, and for whom: for the speaker, for the addressee, or for both. By manipulating one of the images in this second pair, we could independently manipulate whether disambiguation was early or late for the speaker versus early or late for the addressee (Figure 2). Addressees always heard the speaker's description of the target while looking at the post-change display. On critical trials, whatever identity change took place never affected the target or its size-contrasting counterpart. Instead, it affected one of two unrelated symbols, with one of them changing identity (Figure 3). For example, a pre-change display might contain a large 'A' (*target*), a

small 'A' (*target-contrast*), a 'B' and a 'C', with the 'C' changing into a 'B' that contrasts in size with the 'B' that is already present. We refer to the member of this second pair whose size matches that of the target (e.g., the larger 'B') as the *competitor*, and to the other member as the *competitor-contrast*.

The critical manipulations were whether the changing object introduced or removed a size-contrasting object from the second (non-target) pair of symbols, as well as whether speakers formulated their description of the target while looking at the pre-change or post-change display. These manipulations enabled us to independently vary whether one or two size-contrasting pairs were available from the speaker's and the addressee's perspectives. When speakers formulated their descriptions while looking at the post-change display, perspectives matched, with both partners viewing the same display and thus the same number of size-contrasting pairs. When speakers formulated their descriptions while looking at the pre-change display, perspectives mismatched, with one partner considering a display with only one size contrast, and the other considering a display with two.

The key question was: what matters more for contrastive inference, the number of size contrasts in the speaker's or the addressee's perspective? If addressees resolve references based on common ground alone (Clark & Carlson, 1981), then we should see early resolution in the two conditions where the speaker sees a single size-contrasting pair (yielding a main effect of point-of-disambiguation for the speaker). If addressees are egocentric (Keysar et al., 2000), then we should see early resolution whenever there is a single size contrast from the addressee's perspective (yielding a main effect of point-of-disambiguation for the addressee). If addressees simultaneously integrate both perspectives (Heller et al., 2016), then the two mismatching conditions should lie somewhere in between the two matching conditions (yielding main effects of point-of-disambiguation for both speaker and addressee).

Method

Design

This experiment had a two-by-two factorial design, with crossed within-participant and within-stimulus factors of Speaker's Perspective (Early vs. Late Disambiguation) and Addressee's Perspective (Early vs. Late Disambiguation).

Participants

Participants were 34 members of the University of Glasgow community who were paid for their participation. Two participants were removed from the sample for looking at the center of the screen throughout the experiment, leaving 32 in the final analysis. Of these, 16 had the matching perspective conditions for the first block of trials, and 16 had the mismatching perspective conditions. The role of speaker was played by a female undergraduate who was a native speaker of British English. The experiment was approved by the Ethics Committee of the College of Medical, Veterinary, & Life Sciences at the University of Glasgow.

Materials

There were 96 displays in the experiment. Each display presented as stimuli four typographic symbols (e.g., `A`, `B`, `$`, `/`) of two different sizes at the corners of an imaginary

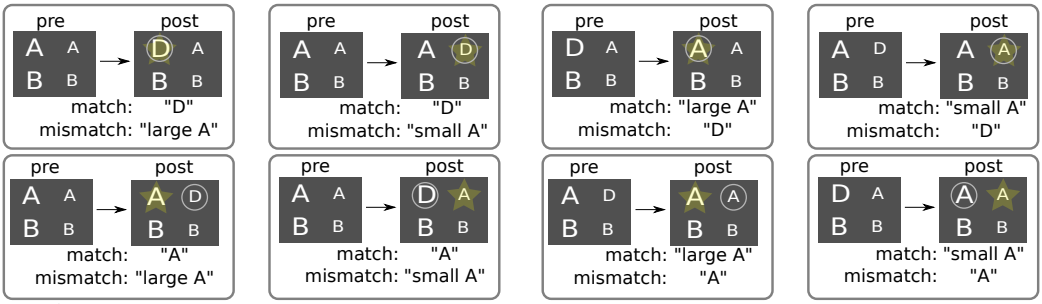

Figure 3

The eight types of filler displays used in Experiment 1. The star indicates the target symbol, but was not depicted on actual displays. The identities of symbols and their locations varied across trials.

square. Each display had two states: a one-contrast state where there was a single size-contrasting pair (e.g., large A and small A) and a large and a small symbol from different categories (e.g., a large B and a small C), and a two-contrast state where there were two size-contrasting pairs (e.g., large A, small A; large B, small B). The displays were designed such that it was possible to transition between these two states in either direction by changing the identity of a single symbol. Half of the displays transitioned from one to two contrasts, and the other half transitioned from two to one. The change always left one size-contrasting pair intact and either removed or created a second size-contrasting pair, depending on the direction of the change. When a symbol changed identity, it was always replaced by another symbol of the same size (e.g., a small A would be replaced by a small C and never by a large C), such that there were always two large-sized symbols and two small-sized symbols in the pre- and post- change displays. Figure 3 shows all configurations used in the experiment.

The target could be any of the four images in the display, with equal probability. Trials where the target was a member of the intact size-contrasting pair comprised the critical trials in the experiment, of which there were 48. We considered as fillers the remaining 48 trials, where the target was either the changed symbol or the symbol that changed from or changed into the same category as the target (see the bottom panels of Figure 3). Addressees' performance on filler displays when perspectives mismatched provides a useful check on whether they are keeping track of perspective disparities. Each of the 16 display types shown in Figure 3 (8 critical and 8 filler) was used exactly six times in the each of the two matching perspectives conditions as well as in each of the two mismatching perspectives conditions.

We manipulated the number of contrasts available to addressees by manipulating whether the changed symbol removed or introduced a second size contrast in the post-change display, with these conditions corresponding to "early" and "late" disambiguation conditions respectively. We manipulated whether speakers' perspectives matched or mismatched addressees' perspectives by having them prepare their descriptions either while looking at either the post-change or pre-change display, which meant that speakers either saw the same or the opposite number of size-contrasting pairs as the addressee.

Symbols were drawn from a larger set of 48 symbols:

!?(@*/&#%→+=$£€123456789ABCDEFGHJKLMNPQRSTUVWXYZ and appeared in one of two sizes, a large size (80 point, or about 150 pixels tall) and a small size (32 point, or about 104 pixels tall). Each symbol was printed in white on a black background in Arial font in a 200x200 bitmap file. The top-left corner of each 200x200 bitmap depicting a symbol was anchored at the following x, y screen coordinates on a 1024x768 pixel display, where (0, 0) corresponds to the top left corner of the display: (228, 100), (596, 100), (228, 468), and (596, 468). Each bitmap was surrounded by a 2 pixel white border to offset it from the rest of the screen. For each display, the symbols were allocated randomly to these locations. For a given participant, each symbol could appear as target no more than once per block, but its appearance as a non-target item was not controlled. Each position on the display had a small number from 1 to 4 next to its outside bottom corner, going from left to right and top to bottom. This was so that on each trial, the speaker could be informed of the target identity by hearing that number in his or her headphones.

Procedure

The EyeLink 1000 eyetracker was calibrated at the start of the first block, and throughout whenever the experimenter deemed it necessary. There was a drift correction before the start of each trial.

To minimize confusion about differences in perspective, the trials were organized into two blocks, with one block of trials containing only those trials where perspectives matched—i.e., where speakers prepared their descriptions while facing the post-change screen (the speaker early / addressee early condition and the speaker late / addressee late condition). The other block of trials were the mismatching perspective trials (speaker early / addressee late; speaker late / addressee early), where speakers prepared their descriptions based on the pre-change screen, but delivered them while looking away from the post-change screen. The order of the blocks was counterbalanced across participants, with 16 participants receiving each order.

At the beginning of each block, there was a practice "walkthrough" trial that allowed the experimenter to give instructions on the procedure for the trials in that block. This trial contained no changing object, because we did not want addressees to believe that speakers were aware that any change would be taking place. After the walkthrough trial of the first block, the participant was given a consent form and asked them to read an additional instruction sheet titled "Secret Instruction", which explained that one of the objects in the display would change, and that the speaker will never be aware that any such change took place. The second trial of each block was a practice trial in which a symbol changed identity.

Figure 1 shows the basic procedure for a single trial in each of the two perspective conditions (matching vs mismatching). The speaker's behavior was guided by auditory cues that were audible only to them, played through a set of closed-back headphones. Two events were common to both conditions. First, five seconds after the beginning of each trial, while the speaker was turned away, addressees were alerted to the impending change of a symbol's identity by the border around that symbol flashing from white to red eight times over a two second period. Second, ten seconds after trial onset, a "go" signal played in the speaker's headphones, cueing them to deliver their description. For the purpose of synchronization, simultaneously with this cue, a "SYNCTIME" message was sent to the eyetracker and the computer began recording audio. If the participant clicked on the correct target symbol, it

was replaced by a star, otherwise by a white X against a red background.

On trials in the block where perspectives matched, the speaker started each trial facing away from the screen. Seven seconds after display onset, after the identity change was complete, the speaker heard a number spoken (1 to 4) in their headphones indicating the location of the target. This served as a cue to turn around, face the screen, and silently prepare their description. They then waited for the 'go' signal, upon which they delivered their description while still facing the screen. Once the addressee clicked on one of the symbols, the speaker heard a cue to turn away from the screen, so that they would start the next trial looking away.

On trials in the block where perspectives mismatched, the speaker started each trial facing the screen. The number identifying the target location was played back in the speaker's headphones simultaneously with the display onset. The speaker silently prepared their description. Three seconds after display onset, the speaker heard a cue to turn away from the screen, so that they would not witness the display change that took place two seconds later. The speaker remained facing away from the screen for the rest of the trial, including when they heard the "go" signal prompting them to deliver their description, ten seconds after display onset. The speaker in this experiment simply described the target using unadorned descriptions such as *small/large X* instead of using imperatives, e.g., *click on the small/large X*. After addressees clicked on one of the symbols, the screen cleared and the speaker heard a cue to face the screen again, so that they would be facing it when the next trial began.

Following completion of the experiment the participant was debriefed about the purpose of the study. On average, sessions lasted about 30 minutes.

Apparatus

We tracked addressees eyes using an EyeLink 1000 tabletop remote eyetracker, which sampled gaze position at a rate of 250 Hz (i.e., every four milliseconds).

Open Practices and Data Availability Statement

For both experiments, we pre-registered the sampling plan (number of subjects and stimuli) as well as the main visual-world eyetracking analysis we performed, which used cluster-based permutation tests. For analyses, only the main analysis was pre-registered; the follow up eyetracking analyses and response time analyses were not pre-registered. The pre-registrations, raw data, analysis code, and materials are openly available at the OSF project repository (Barr et al., 2024). (The archive also includes results from an earlier version of Experiment 1 that was aborted after a design flaw was discovered due to a programming error.)

Data preprocessing

For each of the 48 critical trials for each participant, we used Audacity (audacityteam.org) to view the waveform and code the onset of the adjective, as well as the onset and offset of the noun. To allow synchronization with the eyetracking data, sound recording began simultaneously with a synchronization message sent to the eyetracker. Thus, the

timing of audio events in the soundfile could be directly mapped onto frames in the stream of eye data.

All our analyses used the raw frame data from the eyetracker rather than the processed fixation data. We matched the pixel coordinates of each frame to areas of interest on the display allowing for a 50 pixel border around each image. For each trial, we time-locked the data to the onset of the adjective, and eliminated all frames beyond the point at which the addressee clicked on the target. We identified the point at which 80\

Results

We collected a grand total of 3,072 trials from 32 participants, 1,536 critical trials and 1,536 filler trials. Before any of the analyses reported below, we removed 16 critical trials either because the speaker incorrectly described the target or because there was a problem with the sound recording.

Accuracy

Overall, on critical trials, addressees were highly accurate at identifying the target image with a mean accuracy rate across participants of 99.6% (SD = 0.8%) and a minimum of 97.9%. They were also highly accurate on filler trials, with an average rate across participants of 98.0% (SD = 2.7%) and a minimum of 87.5%. Addressees were also accurate on the subset of filler trials where perspectives mismatched and the target or its potentially contrasting member changed identity, resulting in the speaker's description over- or under-specifying the target from the addressee's perspective or even referring to a symbol no longer present. To accurately identify the target on these trials, addressees would have to remember at least the location where the change took place. Their performance indicated that they did, with accuracy rates well above the 25% chance rate, mean 96.6% (SD = 3.9%) and minimum of 87.5%. These numbers confirm that participants took the task seriously and were motivated to track the speaker's differing perspective.

Eyetracking

Figure 4 shows the overall gaze probability by condition, time-locked to the onset of the adjective. Although the observations are averaged over many different stimulus displays, for ease of exposition we will refer to the plot using a single example where the referring expression is *the small A*, the *target* is the smaller of two As, the *target contrast* is the symbol from the same category which contrasts in size with the target (a larger A). The two remaining symbols mismatched the identity of the target, and of these, the one matching the target on its size dimension is the *competitor* (e.g., a small-sized B or C) and the one matching the target contrast on its size dimension is the *competitor contrast* (e.g., a large-sized B or C). Note that the log ratio plot only shows data for each trial extending up to 200 ms after noun onset, and thus is relatively uncontaminated by any disambiguating information.

What is immediately evident from the plots in the left panel is that at the onset of the adjective, addressees are more likely to be looking at the competitor (small B or C) or competitor-contrast (large B or C) than either the target (small A) or the target contrast (large A). This is not surprising because the change of identity affected one member of

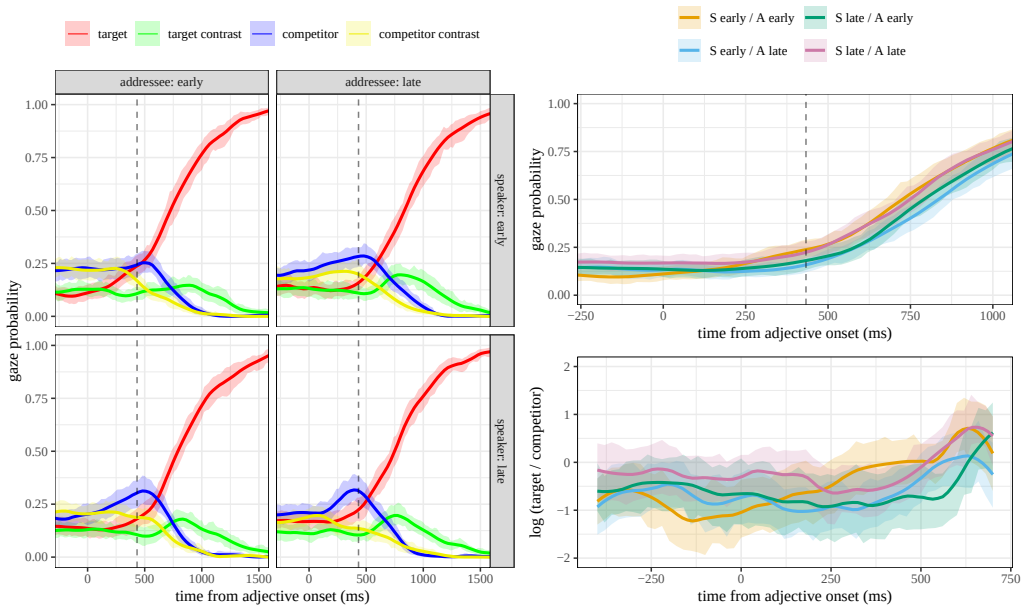

Figure 4

Left panels: Probability of gazing at each image time-locked to adjective onset, Experiment 1. Top right panel: probability of gazing at the target. Bottom right panel: Log ratio of looks to the target versus the competitor. Shaded areas represent 95% confidence intervals for the particular time series, estimated by bootstrapping participants. The dashed vertical line represents the mean onset of the noun at 433 ms.

the former pair. Another discernable pattern is that of relatively late looks to the target contrast (large A) across all conditions, which is also not surprising given that it occurs during the processing of the noun (when "A" is articulated).

Turning to the main question of the study: did looks to the target rise faster in the early disambiguation conditions than in the late conditions? The log-ratio plot (Figure 4, lower right panel) excludes any data following the disambiguation point (i.e., 200 ms after the onset of the noun) and thus any effects related specifically to adjective processing. This window provides the most stringent test of whether addressees can identify the target based on the adjective alone, as claimed in Sedivy et al. (1999). In this plot, higher numbers reflect preference for the target (small A) over the competitor (small B or C). The steepest and earliest (~200 ms) rise in target preference took place in the speaker-early / addressee-early condition, although the difference seems masked by the target likelihood starting off numerically lowest in this condition. Target preference in this condition only numerically exceeded all others in the window between 250 and 500 ms after adjective onset. In the other conditions, a sustained rise only begins around 450 ms after adjective onset. Surprisingly, probabilities seem higher for the condition where disambiguation was late for both parties than for conditions where it was early for either the speaker or the addressee, suggesting that addressees had difficulty managing the conflicting perspectives. The plot of raw target probabilities in the top right panel, which includes data beyond the disambiguation

point (i.e., the onset of the noun), confirms these general impressions, with probabilities in the two matching-perspectives conditions diverging from the two mismatching-perspectives conditions at around 250 ms after adjective onset, and remaining higher than the latter until at least 1000 ms after noun onset.

Pre-registered confirmatory eyetracking analysis

These basic observations were supported by a pre-registered confirmatory analysis using cluster-based permutation tests (Barr et al., 2014; Bullmore et al., 1999; Maris & Oostenveld, 2007), testing at each frame (i.e., one sample every four milliseconds) from 400 ms before adjective onset until 200 ms after noun onset in the corresponding trial. Starting the window early at 400 ms before the noun enabled the potential detection of any statistically significant anticipatory baseline effects. Because adjectives varied in length across trials, this meant that the lengths of the time series available to analyze differed across trials. We opted to end the analysis window at 628 ms, the point at which 50% of the total time series ended. The analysis was performed twice, once with subjects as sampling unit, and once with items.

The analysis at each frame was a baseline-category multinomial logit model, implemented using the `multinom()` function from the `nnet` package (Venables & Ripley, 2002). The dependent variable extracted from each fit was the log odds of fixating the target versus the competitor. The independent variables were the deviation-coded (early $= -\frac{1}{2}$, late $= \frac{1}{2}$) predictors for the factors speaker and addressee and their interaction. The standard errors for each fit were calculated by 500 bootstrap samples over the unit of analysis. As laid out in the pre-registration document, the test statistic was the transformed p-value from the bootstrap randomization ($-\log p$), which was given the same sign as the sign of the parameter estimate in the fit. We calculated cluster-mass statistics by summing up consecutive frames where effects where significant, and generated a null-hypothesis distribution based on 1000 random permutations, where the numeric signs for the two levels of each predictor were randomly swapped for each subject.

The only time region over which there was a significant effect both by subjects and items was from 328–620 ms, where there was a significant interaction between speaker's and addressee's perspective, by-subjects cluster-mass statistic (CMS) of -378.1 from 328 to 620, $p = 0.016$; by-items CMS of -464.5 from 312 to 624, $p = 0.010$. As the lower right panel of Figure 4 illustrates, the pattern largely driving the interaction was such that target advantage was higher in the two matching perspectives conditions than in the two mismatching perspectives conditions.

None of the other clusters detected in the original data were significant after correction, all ps > 0.174.

Follow-up eyetracking analyses

All of the follow-up analyses of the eye-tracking data in this section were not part of the original pre-registration, and were conducted to validate the contrastive effect in the early-matching condition, as well as to aid in interpreting the observed interaction effect.

We ran a follow-up (non-pre-registered) analysis to compare the two matching perspectives condition in the window exhibiting the significant interaction. The purpose was to try to replicate the pattern predicted by Sedivy et al. (1999)—namely, to test

whether the target preference was higher in the early-disambiguation condition than in the late-disambiguation condition. For this analysis, we simply ran a single analysis on the 328–620 ms time window, performed in the same manner as for a single frame in the cluster-permutation test (baseline-category multinomial logistic regression, with bootstrapped standard errors to form a Wald z statistic). Although the trend was in the right direction (.16 logits higher in the early-disambiguation condition), the difference was not statistically significant neither by subject nor by items; respectively, $z_1 = 0.89$, $p_1 = 0.375$; $z_2 = 0.95$, $p_2 = 0.342$.

However, this null result should be interpreted in the context of the tendency (noted above) for addressees to gaze at the competitor during the adjective. Indeed, the log ratio was lowest in the early-matching condition, well before the onset of the adjective. It is also notable that this condition was the only one to exhibit a sustained increase in target preference during the adjective itself. Thus, it is possible that rise in target looks in the early-matching condition was masked by an anticipatory baseline in the opposite direction. To test this possibility, we performed an additional baseline-corrected analysis of the log ratio in the early-matching condition.[1] To account for saccadic programming (Matin et al., 1993), our main window included all frames occurring later than 180 ms after adjective onset until the end of the pre-registered analysis window (184–628 ms). For the baseline, we defined a window of the same size (444 ms) extending backward from 180 ms to -264 ms. The motivation for including a baseline window of the same size is that the range of the log ratio depends on the number of frames included in the calculation. The analysis used a linear mixed-effects model with crossed random factors of participants and stimuli (Baayen et al., 2008), which included by-subject and by-item random intercepts and random slopes for the factors of window (baseline window versus main window) and early versus late, with p-values derived using a likelihood ratio test. Window was coded as 0 for baseline and 1 for main, and early-vs-late as $\frac{1}{2}$ for early and $-\frac{1}{2}$ for late. In the condition where disambiguation was early for both speaker and listener, the log ratio increased by 1.32 units from baseline, which was numerically larger than the 0.44 increase for the condition where disambiguation was late for both parties; however, the window-by-condition difference was not statistically significant, $\chi^2(1) = 2.91$, $p = 0.088$.

The next follow-up analysis confirmed the observation that during the adjective (180–628 ms from adjective onset), overall preference for the target (as measured by the log ratio) was lower in the two mismatching conditions as compared to the late-matching condition. Once again, we used a linear mixed-effects model with by-subject and by-item random intercepts and random slopes for the main factor of interest (mismatching versus late-matching) and p-values derived using a likelihood ratio test. The mean log ratio for the target (versus competitor) in the two mismatching conditions, -0.76, was significantly lower than the mean of -0.01 for the late-matching condition, $\chi^2(1) = 4.96$, $p = 0.026$.

[1] For discussion and justification of statistical correction as an approach for dealing with anticipatory baseline effects, see Barr (2008a) and Barr et al. (2011).

Response time

Table 1. Mean response times (and standard deviations).

	Addressee		
	early	late	
Speaker			
early	1074 (368)	1234 (454)	1154 (421)
late	1222 (414)	1109 (341)	1166 (384)
	1149 (399)	1172 (406)	

Before calculating mean response times on critical trials, we excluded 6 trials where addressees selected something other than the target. Mean response times (and standard deviations) are provided in Table 1. We analyzed the data using linear mixed effects models with crossed random factors of subjects and stimuli and the maximal random-effects structure justified by the design, which meant by-subject and by-item random intercepts and by-subject and by-item random slopes for speaker and addressee and the speaker-by-addressee interaction. The predictors for the factors of speaker and addressee were each coded using deviation coding (early $= -\frac{1}{2}$, late $= \frac{1}{2}$). All models converged with this random effects structure, and we ignored any 'singular fit' messages. We used likelihood ratio tests to derive p-values.

There was a significant interaction between speakers' and addressees' perspectives, $\chi^2(1) = 12.26$, $p < .001$. The cell means in Table 1 suggest that addressees were fastest when disambiguation was early for both the speaker and the addressee. In an (unplanned) analysis suggested by the data, we found a significant effect whereby the speaker-early / addressee-early condition was faster than the average of the three other conditions, $\chi^2(1) = 14.54$, $p < .001$. An additional exploratory analysis revealed that the mean of the two "mixed" perspective conditions was slower than the mean of the speaker-late / addressee-late condition, $\chi^2(1) = 8.98$, $p = 0.003$, which suggests that addressees had difficulty simultaneously managing the two conflicting perspectives.

Turning to the main effects, addressees responded about 12 milliseconds faster when disambiguation was early for the speaker than when it was late, but this main effect of speaker was not statistically significant, $\chi^2(1) = 0.48$, $p = 0.489$. When disambiguation was early for the addressee, responded about 23 milliseconds faster, but this main effect of addressee was also not statistically significant, $\chi^2(1) = 3.77$, $p = 0.052$.

Discussion

Both the eyetracking and response time data suggest that comprehension was disrupted when perspectives conflicted. Addressees took longer to select the target and showed lower rates of target fixation when one or the other perspective afforded early disambiguation than when disambiguation was late from both perspectives. Although early disambiguation was available to addressees in the speaker-early condition, responses appeared to be facilitated only when disambiguation was also early for the addressee. This pattern of findings is

inconsistent with Heller et al. (2008), who found facilitation when disambiguation was early for the speaker but late for the addressee. However, this comes with the caveat that the difference between the early-matching and late-matching conditions was only marginally significant, although the trend was in the right direction.

It is surprising that addressees did not seem to take advantage of information that could have been used to facilitate comprehension. If addressees were attuned to common ground (Clark & Carlson, 1981), there should have been a main effect of speaker, with references resolved early whenever addressees believed speakers saw only one size-contrasting pair. If addressees interpreted speakers' expressions egocentrically (Keysar et al., 2000), they should have only resolved references early when there was a single size-contrasting pair available to them at the moment of interpretation. The overall pattern of eye data and response times is also inconsistent with the proposal that addressees can simultaneously integrate information from multiple perspectives (Heller et al., 2016). Under this view, one would have expected the two mismatching perspectives to fall between the early-matching and late-matching perspectives. Instead, response times were later and looks to the target were less likely than for the late-matching perspective condition.

These findings are unexpected and not predicted by existing theories, and so should be replicated. Also important to note is that the design confounded the manipulation of matching versus mismatching perspectives with cognitive load as well as with joint attention. To perform accurately in the mismatching perspectives conditions, addressees had to remember the previous identity of one of the items, whereas this information could be discarded in the matching perspectives condition. Therefore, mismatching perspective conditions had higher cognitive load. Relatedly, the speaker's perspective was also confounded with joint attention, since speakers were gazing at the screen along with addressees in the matching perspectives conditions, but not in the mismatching perspectives conditions. Given that co-attending may enhance attention to and memory for stimuli (Shteynberg, 2015), it would be worthwhile to confirm the observed patterns in situations where matching and mismatching perspectives are equivalent in terms of joint attention. Experiment 2 removes this confound by having the speaker looking away from the screen in every trial. Since all trials would now be 'mismatching' trials, this also removed the need to block together matching and mismatching trials in the design.

Another concern was that despite finding an significant interaction, our follow-up analysis comparing the early-matching and late-matching conditions did not statistically confirm the contrastive effect originally observed by Sedivy et al. (1999), although there was a marginal trend in the right direction. So while it is clear that perspective conflict did disrupt comprehension, it is less clear that the disruption specifically affected contrastive inferencing. We sought to more clearly confirm this result in the next experiment.

Finally, we observed an overall anticipatory baseline preference for the competitor, which meant that the log ratio measure, which indexes preference for the target over the competitor, was largely negative over the analysis window. One possible reason for this negative bias is that on half of the critical trials, the competitor was the image that changed identity, which drew extra attention to it. We avoided this in Experiment 2, where on critical trials the changed object was always the competitor-contrast.

Experiment 2

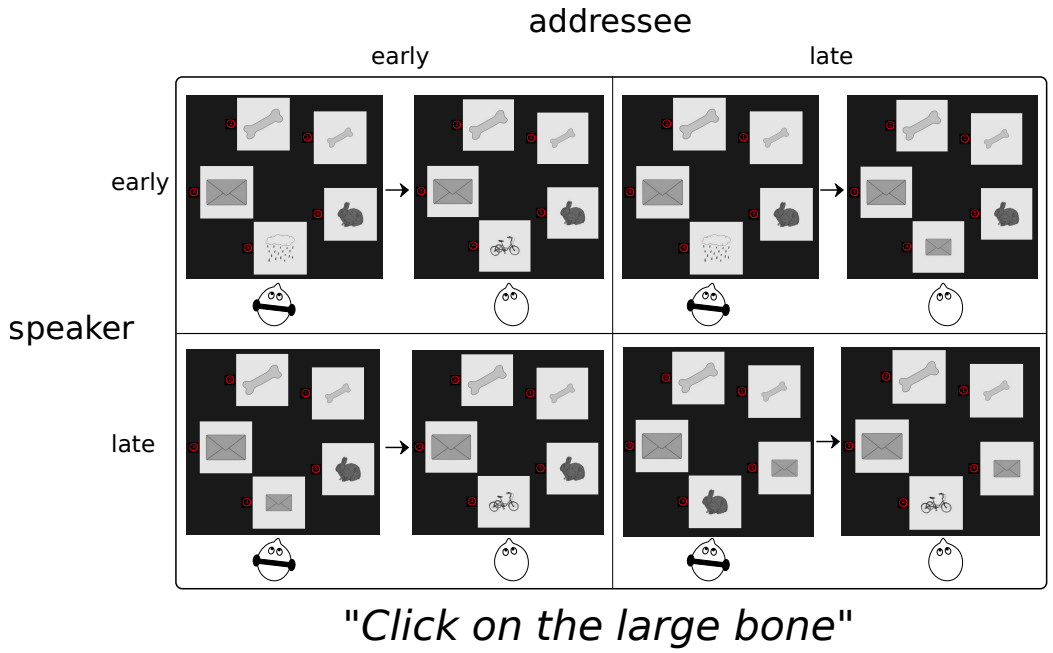

Figure 5

A single item shown across all four conditions in Experiment 2. Unlike Experiment 1, the speaker always prepared her description of the target while viewing the pre-change display.

We repeated Experiment 1 but changed the task by eliminating the matching perspectives condition so that perspectives mismatched in all conditions, thus equating cognitive load and joint attention across all conditions. In other words, across all conditions, speakers formulated their descriptions based on the pre-change display, and addressees interpreted them while viewing the post-change display, and while the speaker was simultaneously looking away from it.. What we manipulated instead was *the way in which perspectives mismatched*. We replaced the former matching perspectives conditions with two new conditions in which both speaker and addressee saw the same number of size-contrasting pairs (one or two) among four of the images in the display, but had different perspectives on the identity of a fifth image that did not partake in any size contrast (Figure 5). This extra image could be a target on filler trials, thus encouraging addressees to attend to the difference.

To make the task more engaging as well as to reduce memory interference across trials by making stimuli more distinctive, we used images of everyday objects instead of typographic symbols. We also did not block trials as we did in Experiment 1; instead, trials appeared in a random order. In most other respects, the experiment was the same as the previous experiment, including the data preprocessing, so in this section we focus on describing differences.

Methods

Participants

A total of 32 participants played the role of addressee in the experiment. All participants self-identified as native speakers of English with normal or corrected-to-normal vision. They participated voluntarily, for course credit, or for a payment of £7. No participants met any of the pre-registered criteria for exclusion (staring at the middle or a target selection rate less than 50% on filler trials). The role of speaker was played by a female research assistant who was a native speaker of British English. The experiment was approved by the Ethics Committee of the College of Medical, Veterinary, & Life Sciences at the University of Glasgow.

Apparatus

We used the same EyeLink 1000 eyetracking system as in the previous experiment, except this time the sampling rate was set to 500 Hz (one sample every 2 milliseconds) instead of 250 Hz.

Design

The design was identical to Experiment 1, with the exception (noted above) that the two formerly matching perspectives conditions (speaker-early / addressee-early; speaker-late / addressee-late) were replaced by mismatching perspective conditions where a fifth object (not participating in any size contrast) changed identity. As noted above and in Figure 5, the perspectives matched or mismatched in terms of the number of size-contrasting sets visible to the speaker and addressee, but always 'mismatched' in terms of the identity of a single object. In the two matching perspectives conditions, the object that changed identity neither destroyed any existing size-contrasting pair nor created any new size-contrasting pair; for instance a raincloud changing into a bicycle, such that both the pre- and post-change displays had a single pair contrasting in size (top left of Figure 5); or a rabbit changing into a bicycle, such that both the pre- and post-change displays had two pairs contrasting in size. In the mismatching perspectives, an object changed such that it created an additional contrasting pair in the post-change display (top right of Figure 5, where a raincloud becomes a small envelope), or eliminated an existing contrasting pair (bottom right of Figure 5, where a small envelope becomes a bicycle).

Materials

We used line art images of everyday objects instead of typographic symbols, downloaded from the Multipic database (Duñabeitia et al., 2018) at https://www.bcbl.eu/databases/multipic/. To manipulate the image files, we used the open-source software imagemagick (https://imagemagick.org), implemented in R through the 'magick' package (Ooms, 2023). R scripts to download and transform the images are available through the project OSF archive, along with slides containing all the displays used in the experiment.

We created large and small versions of each picture. The large version was between 80% and 100% of the size of the original downloaded image, with the value chosen randomly from a uniform distribution. The smaller version was created by subtracting 30% from this

value (e.g., if the 'large' version was 87% of the original size, the small version was 57% of the original). We used this random variation in size to discourage 'absolute' and encourage 'relative' interpretation of the size adjectives. All images were padded as needed with a gray background so that the final image had a 200x200 pixel resolution.

From the 750 transformed grayscale images, we selected 240 for use in the 48 critical displays, and 184 for use in the 44 filler displays. Each image was unique to each display so that no image was seen more than once across the experiment. The images on each display were positioned at the corners of an imaginary pentagon, with the pentagon randomly rotated across items so that the physical locations varied. Each of the 48 critical displays were constructed from combinations of five distinct images categories. For targets and competitors in critical displays, we chose what we deemed to be the 96 most easily named objects. The small and large versions of these images were used to form the two size-contrasting sets, and the remaining three "unrelated" singleton images were randomly chosen from the leftover set. We went through the five images for each display and replaced any unrelated images that had an obvious semantic or phonological association with the target or competitor. Figure 5 shows pre- and post-change displays for a single stimulus item rotated across all four conditions. All other displays followed this basic template, with item positions randomized, and half of the critical displays having one of the larger images as target, and half having the smaller image as target.

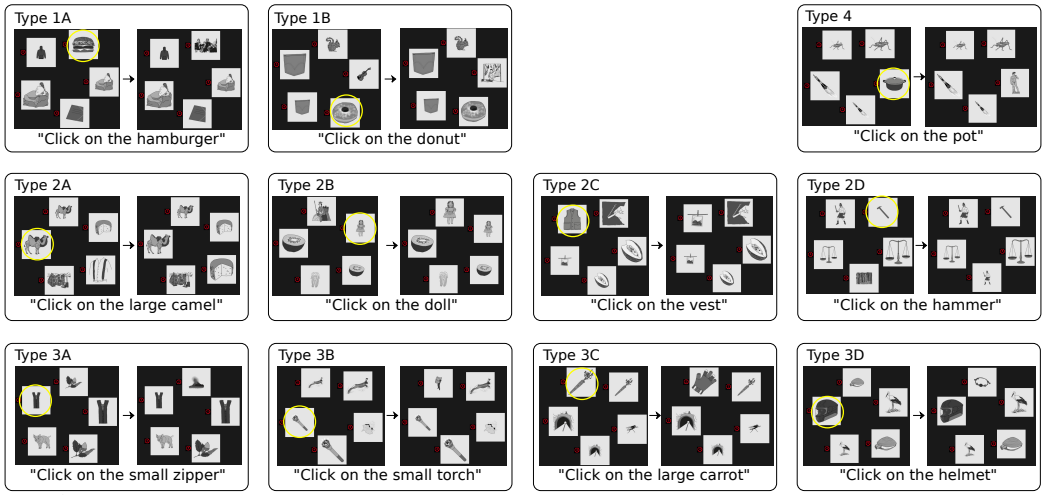

Figure 6

Filler types, Experiment 2.

We included 11 types of fillers (Figure 6) to prevent addressees from being able to guess the identity of the target based on the patterns that repeat across the critical items. Each type was modeled on one of the four conditions of critical items. There were four instances of each type, for a total of 44 fillers.

The first two types fillers (1A and 1B) were modelled on the speaker-early / addressee-early condition. On critical displays in this condition, a single size-contrasting pair was seen along with three singleton objects (one of which changed identity). The target on critical trials in this condition was always one of the size-contrasting pair. For filler type

1A, there was also a single size-contrasting pair, but the target was the singleton whose identity changed. For filler type 1B, the target was one of the singletons whose identity did not change.

The next four filler types (2A, 2B, 2C, 2D) were modelled on the mixed perspective conditions where the speaker saw a single contrast and the addressee saw two. On critical trials in this condition the target was always a member of the stable size-contrasting pair (i.e., whose members did not change identities), and was of the same size as the stable member of the second contrast (the competitor). On fillers of type 2A, target was instead the member of the stable pair that matched the size of the competitor-contrast (which first appeared in the post-change display). On fillers of type 2B and 2C, the target was a member of the "unstable" size-contrasting pair (i.e., the pair where the size contrast was introduced by the change in identity of a singleton object), the "competitor" in the former, and the "competitor-contrast" in the latter. For 2D fillers, the target was the unrelated singleton object.

Filler types 3A, 3B, 3C, and 3D were similarly designed to prevent guessing in mixed perspective conditions, but where the speaker saw two contrasts and the addressee saw one. On critical trials in this condition, the target was always a member of the stable pair, and its size always matched the stable member of the second pair (i.e., the competitor). To counteract this pattern, the target for fillers of type 3A and 3B was the member of the first pair that matched the size of the competitor contrast. For fillers of type 3C, the target was the changed object. For fillers of type 3D, the target was the unrelated object.

Filler type 4 was like the speaker-late / addressee-late condition, but the target was the unrelated object, which changed identity.

We created four stimulus lists to control presentation of stimuli to participants. Each participant received one of the four lists, and each list was structured so that each stimulus item appeared once, and also so that there would be an equal number of stimulus items in each of the four cells of the design. The lists were counterbalanced: the condition each item appeared in was rotated across lists following a Latin Square. We assigned an equal number of participants to each stimulus list.

Procedure

The procedure was the same as in the previous experiment, with one difference: the speaker in this experiment used full imperative sentences, e.g., *click on the small/large X*, instead of unadorned descriptions, e.g., *small/large X*. We split the trials into two balanced blocks to afford a break halfway through the experiment. The trials in each block appeared in a random order.

Results

Accuracy

Overall, on critical trials, addressees were highly accurate at identifying the target image with a mean accuracy rate across participants of 99.9% (SD = 0.4%) and a minimum of 97.9%. They were also highly accurate on filler trials, with an average rate across participants of 98.4% (SD = 2.0%) and a minimum of 90.9%, which meant that none of our participants needed to be excluded according to our pre-registered accuracy

criterion (less than 50% accuracy on filler trials). Addressees were also accurate on the more challenging filler trials where perspectives mismatched and the target or its potentially contrasting member changed identity (types 1A, 2B, 2C, 3C, and 4; see Figure 6), such that the speaker's description misspecified the target from the addressee's perspective. To accurately identify the target on these trials, addressees would at least need to recall the location where the change took place. Their performance indicated that they did, with accuracy rates well above the 20% chance rate, mean 97.7% (SD = 4.4%) and minimum of 80.0%. As in Experiment 1, participants' performance indicates they took the task seriously and were motivated to track the speaker's differing perspective.

Eyetracking

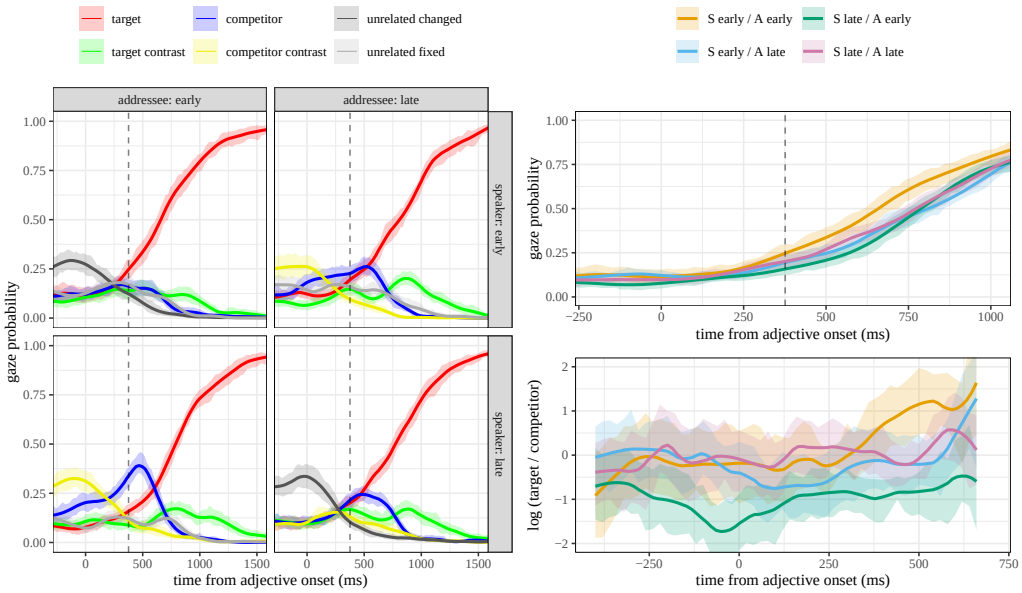

Figure 7

Eyetracking data, Experiment 2. The left half of the plot shows the average probability of gazing at the display images over the full trial broken down by speaker and addressee perspective; the top right image directly compares looks to the target across conditions for the full trial; and the bottom right displays the log-ratio scores indicating preference for the target versus the competitor, up to 200 ms after noun onset. The dashed vertical line represents the mean onset of the noun at 376 ms.

For expository convenience, we describe the results in terms of the example in Figure 5, although in reality the results are averaged across all 48 critical displays. In this example, the target is the large bone, the competitor is the large envelope, and the target contrast is the small bone. The images playing these three roles remained constant across pre- and post- change displays. The "unrelated changed" image changed from one unrelated singleton image into another (e.g., a raincloud into a bicycle). The "unrelated fixed" image was a singleton image that did not change identity during the trial (e.g., the rabbit in the

speaker-early / addressee-early condition).

As an additional expository convenience, we refer to the conditions where both perspectives had the same number of size-contrasting pairs as the "matching perspectives" conditions and to the two conditions where they were different as the "mismatching perspectives" conditions, although technically every condition contained a perspective mismatch.

Before addressing the main question about the effects of perspective on early versus late disambiguation, some tendencies in the data during the preview period should be noted. As expected, prior to the onset of the adjective addressees were most likely to gaze at the image that changed identity, which was the competitor-contrast in the two mismatching perspectives conditions, and an unrelated singleton object in the two matching-perspectives conditions. Looks to this image drop off rapidly 200 ms after adjective onset, as looks to the target start to rise in every condition.

A pattern that was not expected was the elevated tendency of addressees to look at the competitor prior to adjective onset in the two mismatching-perspective conditions. Although we also saw elevated attention to the competitor during preview in Experiment 1, we assumed this was due to the competitor changing identity on half of the trials in each mismatching perspectives condition. This explanation is not viable in the current experiment because the identity of the competitor remained constant. The explanation that seems most likely for this pattern is that addressees have noticed the difference in the number of size-contrasting pairs across conditions, and are perhaps are rehearsing the different perspectives in memory.

The tendency to gaze at the competitor during preview seems especially strong in the speaker-late / addressee-early condition, where the competitor no longer forms part of a size contrast. This is particularly evident in the log-ratio measure (Figure 7, bottom right panel), which directly tracks preference for the target over the competitor, with zero indicating no preference, positive values indicating preference for the target, and negative values indicating preference for the competitor. Although the overall log-ratio values during preview are closer to zero than in the previous experiment, they are distinctly negative in the speaker-late / addressee-early condition.

Pre-registered confirmatory analysis

We performed pre-registered cluster-permutation analyses on the raw frame data. As in the previous experiment, our pre-registered analysis window was from 400 ms prior to adjective onset up to 200 ms after the noun onset. Because the latter value depended upon the duration of the adjective, different trials had time series differing in length, and the analysis would become increasingly noisy. We therefore opted to end the analysis at latest frame at which we had lost data for no more than 20% of trials, which was 578 ms after adjective onset. Any time series ending before that point was padded with `NA` values. The analysis was performed in the same way as Experiment 1, with the exception that the interframe interval was 2 ms instead of 4, because of the higher sampling rate used in data collection (500 Hz instead of 250 Hz).

The log-ratio value is shown in the bottom right panel of Figure 7. As in Experiment 1, there was a significant interaction between speaker and addressee perspective which was significant from 124 ms before adjective onset until 578 ms after onset, $CMS_s = 3061.4$, $CMS_i = 2831.4$, $p_s = 0.001$, $p_i = 0.001$.

The analysis also detected a main effect of speaker's perspective extending from 334 ms until 544, $CMS_s = -1066.9$, $CMS_i = -894.2$, $p_s = 0.001$, $p_i = 0.007$. The window of this main effect was subsumed within, and thus qualified by, the interaction effect noted above. Consideration of the log-ratio plot indicates that this main effect was driven by early disambiguation in the early-matching condition, with the speaker-early / addressee-late condition falling below the matching-late condition.

Follow-up eyetracking analyses

All of the follow-up analyses of the eye-tracking data in this section were not part of the original pre-registration, and were conducted to validate the contrastive effect in the early-matching condition, as well as to aid in interpreting the observed interaction effect.

In the pre-registered analysis, the interaction effect became significant prior to the onset of the adjective, which suggests that it was at least partly driven by differences in anticipatory looks across conditions. Figure 7 does suggest that the only condition in which there is a rise in looks to the target is the early matching condition, but it is nevertheless important to confirm this observation statistically, taking into account the clear anticipatory baseline differences across conditions. We used statistical correction to control for these anticipatory effects (Barr, 2008a; Barr et al., 2011). We defined the main window as all frames more than 180ms after adjective onset until the last frame of data in the pre-registered analysis, at 578 ms. We defined a baseline window of the same size (396 ms) extending backward from 180 ms after onset to -216 ms before onset. We then calculated the log ratio of looks to the target versus competitor over each window in each of the conditions (Figure 8). What is evident from the figure is that the only condition to show an increase in preference for the target from baseline was the early-matching condition. We ran a linear mixed-effects analysis of the log ratio data with the factor of Window dummy coded (0 = baseline, 1 = main), and with the two Speaker and Addressee factors deviation coded (early = $\frac{1}{2}$, late = $-\frac{1}{2}$). The model included interactions between all three factors, as well as maximal random effects, entailing by-subject and by-item random intercepts and random slopes for all main effects and interactions, with covariances fixed to zero to improve convergence (Barr et al., 2013). We used likelihood ratio tests to derive p-values.

There was a significant three-way interaction between Window, Speaker, and Addressee, $\chi^2(1) = 4.69$, $p = 0.030$. When disambiguation was early from both perspectives, the log ratio showed a statistically significant increase from baseline of 0.93 log units, $\chi^2(1) = 9.65$, $p = 0.002$. The increase from baseline was not significant in any of the other conditions; indeed, all showed (nonsignificant) *decreases*: of 0.06 in Speaker-Early / Addressee-Late, $\chi^2(1) = 0.04$, $p = 0.841$; of 0.22 in Speaker-Late / Addressee-Early, $\chi^2(1) = 0.64$, $p = 0.424$; and of 0.05 in Speaker-Late / Addressee-Late, $\chi^2(1) = 0.05$, $p = 0.829$. For completeness, we also report the interaction of the main effects with Window: Speaker-by-Window interaction, $\chi^2(1) = 3.80$, $p = 0.051$; Addressee-By-Window interaction, $\chi^2(1) = 2.53$, $p = 0.112$.

For our next follow-up analysis, we confirmed the main contrastive finding of Sedivy et al. (1999) by comparing the two conditions where perspectives matched (early-early vs late-late). There was a significant interaction of this factor with Window, $\chi^2(1) = 7.00$, $p = 0.008$. In our final follow-up analysis, we found that the increase from baseline in the early-early condition differed significantly from the increase in the three other conditions,

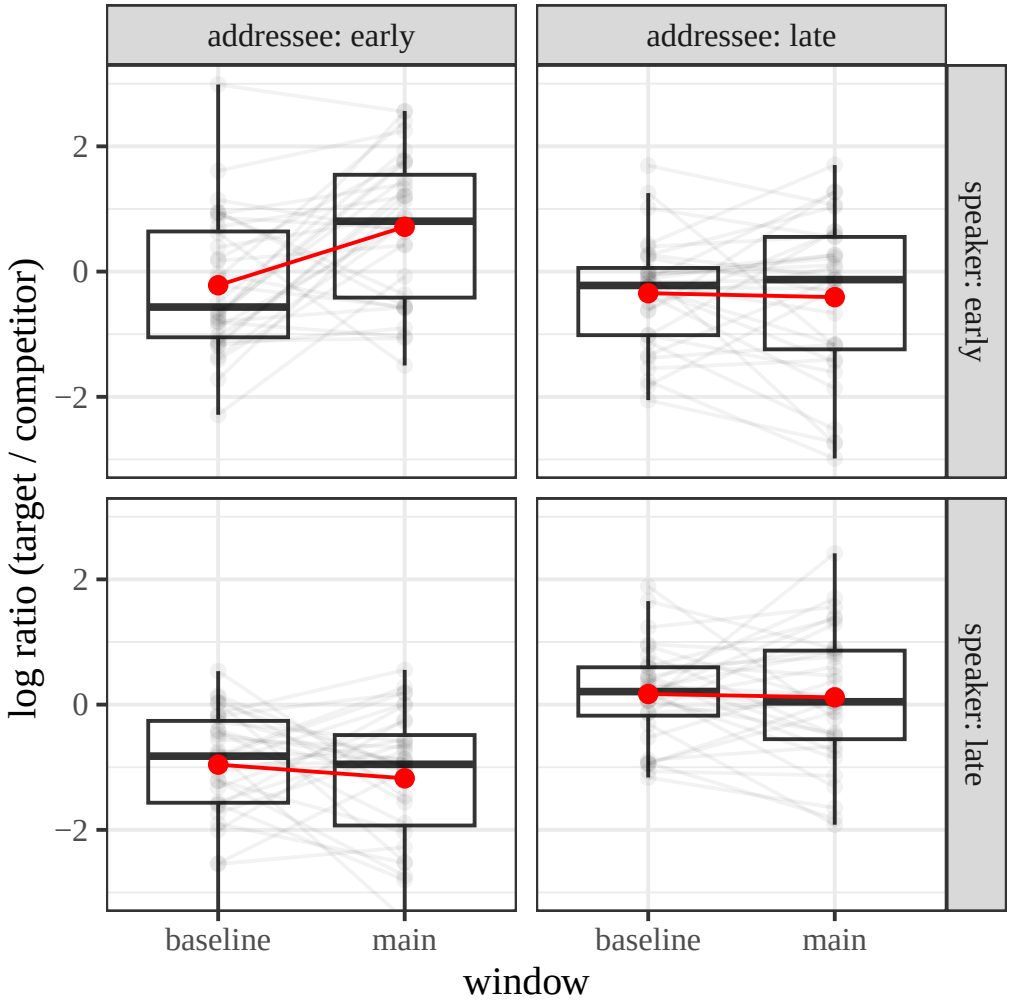

Figure 8

Log ratio (target / competitor) analysis for eyetracking data, Experiment 2, comparing baseline (-216–180ms) to main (182–587ms) analysis windows across factors of Speaker and Addressee Perspective. The larger red dots represent the cell means, and the smaller gray dots in the background represent cell means for individual participants.

supported by an interaction of this contrast with Window, $\chi^2(1) = 11.19$, $p < .001$.

Response time

Table 2

Table 2. Mean response times (and standard deviations), Experiment 2

	Addressee		
	early	late	
Speaker			
early	963 (346)	1075 (392)	1020 (374)
late	1088 (344)	1034 (354)	1061 (350)
	1026 (351)	1055 (374)	

Before calculating mean response times on critical trials, we excluded seven trials where the speaker incorrectly described the target, and one trial where the addressee clicked on the wrong image. Mean response times (and standard deviations) are provided in Table 2. The mixed-effects analysis was conducted in the same way as in the previous experiment. All models converged with the maximal random effects structure. We used likelihood ratio tests to derive p-values.

Cell and marginal means for response time are presented in Table 2. The overall pattern of means was consistent with Experiment 1.

Like in Experiment 1, we detected a significant crossover interaction between speaker and addressee $\chi^2(1) = 19.08$, $p < .001$, with the simple effect of speaker showing facilitation in the addressee-early condition and inhibition in the addressee-late condition. As in the previous experiment, the interaction is driven response time being facilitated in the early-matching condition. An analysis comparing this condition to the mean of the three other conditions detected a significant difference, $\chi^2(1) = 20.30$, $p < .001$. Also, as in the previous experiment, the mean of the two mismatching perspective conditions was slower than the mean of the speaker-late / addressee-late condition, $\chi^2(1) = 4.36$, $p = 0.037$. This result provides unequivocal evidence that addressees had difficulty simultaneously managing the two conflicting perspectives, since the cognitive load was held constant across conditions.

Turning to the main effects, when disambiguation was early for the addressee, they responded about 29 milliseconds faster, but the main effect of addressee was not statistically significant, $\chi^2(1) = 3.82$, $p = 0.051$. In contrast, when disambiguation was early for the speaker, addressees responded about 41 milliseconds faster, a significant main effect of speaker, $\chi^2(1) = 7.38$, $p = 0.007$.

Discussion

Overall, the results of Experiment 2 lend even stronger support for perspective conflict disrupting contrastive inference. In this experiment, the predicted advantage for targets over competitors during the adjective in the early-matching condition was significant in the pre-registered analysis, indicating early resolution of the temporary ambiguity when

it was afforded from *both* perspectives. In contrast, when afforded from just one of the two perspectives, addressees preferentially looked at the target at a lower rate and were slower to select it. Indeed, they were delayed even more than in the late-matching condition. We offer two possible explanations for this intriguing finding in the next section.

General Discussion

Across two visual-world experiments considering two dependent variables (eye gaze and response time), we found no evidence that addressees resolved temporary ambiguity by disambiguating contrastive descriptions from a speaker's differing perspective. Both experiments used a novel 'Display Change' paradigm involving perspective differences across time instead of space. This paradigm made it possible to manipulate the point of disambiguation of size-modified expressions independently for each perspective. In Experiment 1, targets were alphanumeric symbols varying in size, and the 'matching' perspective condition was created by having the speaker prepare the target description after the change had occurred. During the adjective, addressees' preferential looks to the target (versus the competitor) were reduced in both mismatching conditions compared to the matching conditions. Also, addressees were slower to select the target in both mismatching conditions relative to the late-matching condition and fastest in the early-matching condition. However, the pre-registered analysis did not fully replicate the basic contrastive effect of Sedivy et al. when the early-matching and late-matching conditions were compared, although the difference was marginally significant and in the predicted direction. In addition, the observed patterns could be explained by possible confounds with cognitive load and joint attention.

Experiment 2 used images of everyday objects and equated cognitive load and joint attention across matching and mismatching perspectives by always having a perspective mismatch, but manipulating whether the change impacted the number of size-contrasting pairs for the speaker or the addressee. Here we found that the early-matching condition was the only condition under which addressees preferentially looked at the target over the competitor during the adjective. As in Experiment 1 response times were also delayed relative to the late-matching conditions when the number of size-contrasting sets mismatched across perspectives. Although the evidence from Experiment 1 was imperfect, the non-significant trends observed there are fully consistent with the statistically confirmed patterns in Experiment 2.

Consistent with Sedivy et al. (1999), both experiments found that addressees were faster to identify targets when context supported a single contrastive interpretation. Our findings therefore do support the general assumption underlying modern psycholinguistics theories that addressees incrementally combine information from context along with the linguistic input to rapidly identify referents. However, they offer an important qualification: addressees only did so in cases where perspectives matched in terms of the number of contrasting sets. When there was a mismatch, addressees exhibited lower rates of preferential looks toward the target and were delayed in selecting it.

Little evidence was found to support the view that addressees are capable of using a speaker's differing perspective to resolve temporary ambiguities, despite the experiments being designed to be maximally sensitive to such effects. It is not the case that the task was too cognitively demanding relative to studies using visual occlusion, because the perspective conflict was the minimum possible: a disparity in belief about the identity of a single

object. Moreover, the absence of effects of the speaker's perspective on temporary ambiguity resolution must be interpreted against the background of addressees' high levels of accuracy on filler trials where speakers' descriptions did not match any of the on-screen candidates, showing that they did attend to the perspective differences. Finally, the failure to find a speaker-specific effect is unlikely to reflect problems with power. Compared to the closest relevant study that detected positive effects of speaker perspective (Heller et al., 2008), each of our experiments comprised six times the number of observations (32 vs 16 participants and 48 vs 16 stimuli).

Our experiments measured comprehension as it unfolded while participants were being directly addressed during live interaction with a physically co-present speaker who was in the same room with them. Such co-presence may be an important factor for detecting effects of common ground (Brown-Schmidt, 2009). The speaker was not a naïve speaker but a confederate, which can be a source of concern in some circumstances (Kuhlen & Brennan, 2013). In the current experiment, this is unlikely to have affected the results, given that the confederate was truly blind to the number of contrast sets available to the addressee, and always described targets in a manner consistent with what would be expected from a naïve speaker in that situation. Still, it was unavoidable that the speaker would behave in a way that implied greater expertise with the task than what would be expected from a naïve participant. It is unclear how this perception of expertise might have influenced our results. Nevertheless, our experimental task could easily accommodate the use of naïve participants as speakers, and we believe it would be worthwhile to replicate our findings using this more natural approach.

The results from these studies are interesting and unexpected from the point of view of the main theories about perspective taking in language comprehension. All theories would struggle to explain the interaction that was seen across both experiments. If addressees constrained comprehension to information in common ground (Clark & Carlson, 1981), then there should have been a main effect of speaker's perspective alone. If addressees egocentrically anchored comprehension in their own perspective (Keysar et al., 2000), or if they were simply unable to integrate common ground (Barr, 2008b), then there should have been a main effect of Addressee's perspective alone. The "partial constraint" view, along with its recent instantiation in the "multiple perspectives" view (Heller et al., 2016), assumes that multiple perspectives can be entertained simultaneously, and that information from both perspectives are integrated with language in real time. For the conditions in which perspectives mismatched—namely, the early-speaker / late-addressee and late-speaker / early-addressee conditions—this view would predict that the speed of target identification should be intermediate between the early-match and late-match conditions. Instead, both mismatching conditions were delayed relative to the late-match condition.

Experiment 2 showed that it was not conflict *per se* that disrupted pragmatic inference, but only a conflict that led to a divergence in the number of contrasting sets across perspectives. We see two plausible explanations for this disruption: one in terms of interpretation strategies and the other in terms of interference due to cognitive limitations. The first possibility is that when participants noticed a conflict in terms of the number of contrasting sets they engaged in *pragmatic deferral*, adopting a more cautious mode of interpretation and suspending ordinary pragmatic inferences. This interpretation assumes both that (1) language users actively attend to relations between perspectives and pre-compute

pragmatically relevant differences and (2) that pragmatic inferences are optional rather than obligatory. The first assumption is consistent with recent proposals about theory of mind and perspective taking, which emphasize attending to relations between perspectives rather than simply holding multiple perspectives in mind (Deschrijver & Palmer, 2020; Heller & Brown-Schmidt, 2023). But to our knowledge empirical and theoretical support is lacking for the second assumption, that addressees would spontaneously suspend pragmatic inferences in the face of perspective conflict. This would imply a level of metacognitive awareness and flexibility not previously noted in the literature on pragmatic inferencing.

The second interpretation for the disruption also assumes that addressees pay attention to pragmatic relations between perspectives, but attributes the disruption to an inability to simultaneously compute, in real time, diverging pragmatic implications from two incompatible perspectives. Research on perception and attention suggests that people can only entertain one potential interpretation of a stimulus at a time (Chater, 2018; Huang & Pashler, 2007). Similar to a visually ambiguous bistable image, addressees may entertain one of the other perspective at a time, but not both simultaneously. Although accessing the 'egocentric' interpretation based on visually available information at the moment of comprehension should be rapid and effortless, perhaps the delay arises from the need to suppress this egocentric perspective to compute an alternative interpretation from the speaker's view. Under this interpretation, the delay in identifying the target would arise due to the cost of shifting perspectives, rather than the suspension of pragmatic inferences. The current data offers no opportunity to distinguish between these views.

That perspective-taking theories neither predict nor can easily explain the disruption of contrastive inference is likely a consequence of their focus on situations involving embedded rather than conflicting perspectives. However, the picture offered by those few existing studies that do examine conflicting perspectives is not yet cohesive. Of these studies, our results are most consistent with Mozuraitis et al. (2015), who found no reduction in interference from 'knowledge-based' competitors (a candle that looks like a lightbulb) when knowledge of the competitor's functional identity was privileged. That effects of perspective are absent for situations of conflicting knowledge also seems consistent with the theoretical view of Westra & Nagel (2021), who suggest that interlocutors reason more easily about so-called 'factive' differences in perspective—where differences are related to tagging what an interlocutor does and does not know—in contrast with 'nonfactive' differences, which involve appreciating a reality that is distinct to one's own. However, Hanna et al. (2003) used conflicting perspectives in Experiment 2 of their article and found evidence for early disambiguation, in direct contradiction with our results here. This discrepancy seems likely to be related to the difference between inducing perspective conflict by mislabeling a referent (as in Hanna et al., 2003) versus doing so by changing the identity of a referent not previously mentioned. This explanation is supported by Wolter et al. (2011), who found distinct contributions of linguistic mention and visual context on the processing of scalar expression. To settle this question, it would be necessary to directly contrast these different sources of perspective conflict within the same experiment.

Although our failure to replicate Heller et al. (2008) has something to do with the way we instantiated differences in perspective, it may be premature to accept the premise that people are somehow 'better' at perspective taking (e.g., exhibit less egocentric interference) when dealing with embedded versus conflicting perspectives. To date, these two

types of perspective differences have been instantiated in different ways, raising analytical and interpretive problems in drawing comparisons. Specifically, most studies using embedded perspectives involve visually occluded objects or images to which speakers could not plausibly refer, creating anticipatory baseline effects—latent or overt differences in gaze probability that are present in advance of a critical referring expression (Barr et al., 2011). There are interpretive problems in comparing effects that emerge during critical referring expressions across conditions with different baselines (Barr, 2008a), and different researchers seem to deal with such effects in different ways.

The use of the Display Change task may mitigate anticipatory baseline effects, inasmuch as it allows all spatial regions to contain plausible referents. However, it does not wholly prevent them. Objects that changed identity after the speaker turned away attracted attention to themselves as well as to other objects belonging to contrast sets that they created or disrupted. While this is not ideal, it is still possible to statistically correct for such effects, as we showed in Experiment 2. But perhaps such anticipatory effects could be better avoided in future studies if the contrasting dimension was not always size, but varied from trial to trial. Despite these limitations, we believe that the Display Change task opens up exciting new possibilities for studying how addressees deal with perspective differences during real-time spoken language comprehension.

References

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

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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

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

Barr, D. J. (2008b). Pragmatic expectations and linguistic evidence: Listeners anticipate but do not integrate common ground. *Cognition*, *109*(1), 18–40.

Barr, D. J., Britain, H., & Li, Q. (2024). *Perspective taking in the interpretation of prenominal adjectives 2*. OSF. https://doi.org/10.17605/OSF.IO/7EBTJ

Barr, D. J., Gann, T. M., & Pierce, R. S. (2011). Anticipatory baseline effects and information integration in visual world studies. *Acta Psychologica*, *137*, 201–207.

Barr, D. J., Jackson, L., & Phillips, I. (2014). Using a voice to put a name to a face: The psycholinguistics of proper name comprehension. *Journal of Experimental Psychology: General*, *143*, 404–413.

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. https://doi.org/10.1016/j.jml.2012.11.001

Birch, S. A., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*, 382–386.

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language*, *61*, 171–190.

Bullmore, E. T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., & Brammer, M. J. (1999). Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural mr images of the brain. *IEEE Transactions on Medical Imaging*, *18*(1), 32–42.

Chater, N. (2018). *The mind is flat: The illusion of mental depth and the improvised mind*. Penguin UK.

Clark, H. H., & Carlson, T. B. (1981). Context for comprehension. In J. Long & A. Baddeley (Eds.), *Attention and performance ix* (pp. 313–330). Erlbaum.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshe, B. L. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 10–61). Cambridge University Press.

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

Crain, S., & Steedman, M. (1985). On not being led up the garden path. In D. Dowty, L. Kartunnen, & A. Zwicky (Eds.), *Natural language parsing* (pp. 320–358). Cambridge University Press.

Deschrijver, E., & Palmer, C. (2020). Reframing social cognition: Relational versus representational mentalizing. *Psychological Bulletin*, *146*, 941–969.

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). Multipic: A standardized set of 750 drawings with norms for six european languages. *Quarterly Journal of Experimental Psychology*, *71*(4), 808–816.

Eberhard, K., Spivey-Knowlton, M., Sedivy, J., & Tanenhaus, M. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*, 409–436. https://doi.org/10.1007/bf02143160

Flavell, J. H., Flavell, E. R., & Green, F. L. (1983). Development of the appearance-reality distinction. *Cognitive Psychology*, *15*, 95–120.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*, 43–61. https://doi.org/10.1016/s0749-596x(03)00022-6

Heller, D., & Brown-Schmidt, S. (2023). The Multiple Perspectives theory of mental states in communication. *Cognitive Science*, *47*, e13322. https://doi.org/10.1111/cogs.13322

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*, 831–836.

Heller, D., Parisien, C., & Stevenson, S. (2016). Perspective-taking behavior as the probabilistic weighing of multiple domains. *Cognition*, *149*, 104–120.

Huang, L., & Pashler, H. (2007). A boolean map theory of visual attention. *Psychological Review*, *114*(3), 599.

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition*, *57*, 129–191.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*,

11, 32–38.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25–41.

Kuhlen, A. K., & Brennan, S. E. (2013). Language in dialogue: When confederates might be hazardous to your data. *Psychonomic Bulletin & Review*, *20*, 54–72.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of Neuroscience Methods*, *164*(1), 177–190.

Matin, E., Shao, K.-C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics*, *53*, 372–380.

Mozuraitis, M., Chambers, C. G., & Daneman, M. (2015). Privileged versus shared knowledge about object identity in real-time referential processing. *Cognition*, *142*, 148–165.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints on children's on-line reference resolution. *Psychological Science*, *13*, 329–336.

Ooms, J. (2023). *Magick: Advanced graphics and image-processing in r*. https://CRAN.R-project.org/package=magick

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, *5*(2), 125–137.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–190.

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

Shteynberg, G. (2015). Shared attention. *Perspectives on Psychological Science*, *10*(5), 579–590.

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

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth). Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Westra, E., & Nagel, J. (2021). Mindreading in conversation. *Cognition*, *210*, 104618. https://doi.org/10.1016/j.cognition.2021.104618

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*, 103–128.

Wolter, L., Gorman, K. S., & Tanenhaus, M. K. (2011). Scalar reference, contrast, and discourse: Separating effects of linguistic discourse from availability of the referent. *Journal of Memory and Language*, *65*, 299–317. https://doi.org/10.1016/j.jml.2011.04.010

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*, 162–185.