Cognition 147 (2016) 75–84

Contents lists available at ScienceDirect

Cognition

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

Listeners use speaker identity to access representations of spatial perspective during online language comprehension

Rachel A. Ryskin [a,*], Ranxiao Frances Wang [a,b], Sarah Brown-Schmidt [a,b]

[a] Department of Psychology, University of Illinois at Urbana-Champaign, United States
[b] Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, United States

ARTICLE INFO

Article history:
Received 8 November 2014
Revised 5 November 2015
Accepted 20 November 2015

Keywords:
Spatial perspective-taking
Partner-specific encoding
Language comprehension
Eye-tracking

ABSTRACT

Little is known about how listeners represent another person's spatial perspective during language processing (e.g., two people looking at a map from different angles). Can listeners use contextual cues such as speaker identity to access a representation of the interlocutor's spatial perspective? In two eye-tracking experiments, participants received auditory instructions to move objects around a screen from two randomly alternating spatial perspectives (45° vs. 315° or 135° vs. 225° rotations from the participant's viewpoint). Instructions were spoken either by one voice, where the speaker's perspective switched at random, or by two voices, where each speaker maintained one perspective. Analysis of participant eye-gaze showed that interpretation of the instructions improved when each viewpoint was associated with a different voice. These findings demonstrate that listeners can learn mappings between individual talkers and viewpoints, and use these mappings to guide online language processing.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Much of human communication requires keeping track of what another person knows. For example, when a coworker says, "How was the talk?", taking the sentence at face-value you might begin to think about every talk that you have ever attended (or even heard of), which would lead you to an uninformative response—at best a clarification question, at worst telling your coworker about something irrelevant. In the case of something that occurs more frequently than "talks", it could even lead to an interminable memory search process. However, a more effective strategy, and one that successful communicators must employ, takes into account what your coworker knows and narrows the search space to only talks that she knows occurred, that she knows she did not attend, and that she knows you did attend. Tracking others' knowledge places constraints on the set of possible intended referents and eases the burden of comprehension, allowing conversation to proceed smoothly. The nature of this constraining knowledge can take many forms, from what topics have been previously discussed between two interlocutors, to an individual's viewpoint on the physical environment.

In order to understand a speaker, listeners must consider that speaker's perspective (Clark, 1992) and how it may differ from their own. Indeed, listeners are sensitive to differences in perspectives between themselves and an interlocutor and bring this information to bear in the early moments of processing a sentence (Brown-Schmidt, 2009, 2012; Brown-Schmidt, Gunlogson, & Tanenhaus, 2008; Hanna, Tanenhaus, & Trueswell, 2003; Heller, Grodner, & Tanenhaus, 2008; Nadig & Sedivy, 2002). The bulk of this evidence comes from paradigms in which a difference in perspectives between the speaker and listener is created by occluding an item from the speaker's view. Much of this research shows that listeners (at least partially) discount occluded objects as potential referents, on the assumption that the speaker is unlikely to speak about something they have no knowledge of. This successful use of perspective corresponds to what has been referred to as Level 1 knowledge—mental simulation that involves distinguishing what is visible to oneself from what is visible to others, as in occlusion situations. Level 1 knowledge emerges early in development and is thought to require little cognitive effort, even by age three (Flavell, Everett, Croft, & Flavell, 1981; Masangkay et al., 1974).

Differences in perspective can arise from situations other than occlusion, as well. In particular, differing spatial viewpoints, which are the focus of the present manuscript, require interlocutors to take this into account in order to understand each other (Schober, 1993). It has been argued that this Level 2 knowledge—the ability to appreciate not only that another person sees something, but how they see it—emerges later in development and is more cognitively effortful (Apperly & Butterfill, 2009; Flavell et al., 1981; Salatas & Flavell, 1976).

* Corresponding author at: Department of Psychology, University of Illinois at Urbana-Champaign, 603 E. Daniel St., Champaign, IL 61820, United States.
E-mail address: ryskin2@illinois.edu (R.A. Ryskin).

http://dx.doi.org/10.1016/j.cognition.2015.11.011
0010-0277/© 2015 Elsevier B.V. All rights reserved.

In daily life, the spatial viewpoints of conversation partners are often misaligned. In fact, in some sense, they are never truly aligned because interlocutors can never inhabit the exact same location at the same time (Schober, 2009). Speakers frequently and spontaneously take into account such differences in perspectives when communicating (Tversky & Hard, 2009). For instance, when giving walking directions to a friend, you might say, "From Sixth Street you'll take a *left* on Daniel Street, and I'll be standing halfway up the block." From your own perspective, your location is actually to the *right* of Sixth Street but, as a courtesy to your friend who is unfamiliar with the area, you take their perspective in order to avoid confusion. Indeed, Schober (2009) found that when participants with high spatial perspective-taking ability are matched with participants with low abilities, they adopt spatial language consistent with that partner's perspective more often when giving directions compared to when they are paired with someone who is equally capable at spatial perspective-taking. Similarly, speakers use their egocentric perspective less when directing a person who is unable to provide immediate feedback about whether they understood the spatial instruction (Schober, 1993). The fact that speakers often choose to start with their own perspective when they know that any confusion can be easily resolved (i.e., when the person receiving the instructions can ask for clarification) points to the inherent difficulties of performing a spatial perspective transformation.

In this paper, we briefly review what is known about how spatial perspectives are represented in memory and the mental computations required to imagine another perspective. We then discuss how these spatial memory representations might be called upon during language processing when speaker and listener perspectives differ. We hypothesize that spatial perspectives of interlocutors can be tied to the speaker's identity in memory and accessed on-line to constrain interpretation of what has been said. Finally, we provide empirical evidence that supports our hypothesis and discuss the implications of our findings for theories of perspective-taking and language comprehension more broadly.

1.1. Spatial perspective-taking and memory

Studies of memory for spatial layouts of objects indicate that a mental change in viewpoint renders information about object-to-object relations more difficult to retrieve compared to when the viewpoint remains stable (e.g., Shelton & McNamara, 1997, 2001; Simons & Wang, 1998). Rieser (1989) asked participants to memorize an array of objects and then tested their ability to retrieve the relative spatial location of an object from a novel point of view. Participants had more difficulty doing so when the novel location was reached by a rotation than by a simple translation. One potential explanation for this processing cost associated with viewpoint rotation comes from evidence that participants most often encode the environment, and objects within it, using an egocentric reference frame (e.g., Wang, 2007, 2012). As a result, taking another perspective requires the effortful transformation of the original (egocentric) reference frame to fit a new orientation (Easton & Sholl, 1995; Kessler & Thomson, 2010; Mou, McNamara, Valiquette, & Rump, 2004). Others have argued that the processing cost results primarily from the sensorimotor interference created between the coordinates in the person's own perspective and those in the imagined perspective (Brockmole & Wang, 2003; May, 2004; Wang, 2005).

Furthermore, the difficulty of spatial perspective-taking increases with the angular disparity between the participant's viewpoint and the novel viewpoint presented at test (e.g., Huttenlocher & Presson, 1973; Kessler & Rutherford, 2010; Kessler & Thomson, 2010; Levine, Jankovic, & Palij, 1982; Rieser, 1989; Surtees, Apperly, & Samson, 2013). The detrimental effects of greater angular disparity suggest that spatial perspective-taking is an embodied cognitive process (Kessler & Thomson, 2010). Thus, a listener taking into account the perspective of her conversation partner will mentally rotate her egocentric perspective to align it with the partner's.

One way to reduce the cognitive burden of spatial perspective-taking is by providing advance information about a viewpoint. Studies asking participants to imagine a perspective before seeing an array from the new viewpoint show that the representation of a perspective can be maintained in memory in the absence of the visual array that it applies to (Avraamides, Ioannidou, & Kyranidou, 2007; Avraamides, Theodorou, Agathokleous, & Nicolaou, 2013; c.f. Wang, 2005). Further, Galati, Michael, Mello, Greenauer, and Avraamides (2013) provide evidence that speakers do learn and store representations of their future conversation partner's spatial viewpoint, when it is made available to them in advance. However, the nature of these representations and how they are stored and accessed may differ substantially between speakers and listeners. The task of the speakers is to put into words the spatial perspective that they have chosen to adopt whereas the listeners must remain flexible enough in their representations to adapt to whichever unknown perspective they are about to hear an instruction from.

1.2. Spatial perspective-taking during language processing

The challenges involved in representing others' spatial perspectives are well documented. Yet, the comprehension processes involved in interpreting spatial language from a perspective that differs from one's own are less well understood. It is clear that speakers often produce spatial language from the intended recipient's perspective and listeners (or readers) can come to understand spatial directions that are given from a different perspective (Schober, 1993; Tversky & Hard, 2009; Taylor & Tversky, 1992). Yet, little is know about the mechanisms involved in, or the time-course of, adopting a different spatial perspective during language comprehension.

The integration of an occlusion-based difference in perspectives occurs rapidly during sentence interpretation (e.g., Heller et al., 2008). However, the processes involved in computing a differing perspective are not the same when that difference is the result of occlusion compared to when it stems from an alternative spatial orientation (Michelon & Zacks, 2006). Occlusion prompts participants to use a simple line-of-sight tracing strategy to compute the differences between their perspective and that of their partner. By contrast, when spatial perspectives are misaligned, participants must undergo an imagined transformation of their perspective and remapping of reference frames, which may lead to a conflict between the imagined and egocentric reference frames. Thus, spatial perspective-taking and occlusion-based perspective-taking may differently guide the on-line comprehension of utterances.

Nonetheless, some evidence suggests that listeners are able to use information about a speaker's spatial viewpoint to constrain the interpretation of a sentence as it unfolds. Ryskin, Brown-Schmidt, Canseco-Gonzalez, Yiu, and Nguyen (2014) monitored the eye movements of listeners as they processed sentences with potentially ambiguous spatial language. Participants heard instructions to move objects around a complex display of animals with accessories (e.g., a hat, a purse). The instructions, such as "*Go left to the pig with the hat*," were given either from the participant's egocentric perspective (i.e., "left" = participant's left) or the opposite perspective (a 180° rotation; "left" = participant's right). The displays were designed such that instructions were temporarily ambiguous between two potential referents. For example, "*the pig with the...*" was temporarily consistent with two different pigs

on the screen, one of which was located to the left of the starting position (e.g., a pig with a hat), and the other was located to the right of the starting position (e.g., a pig with a purse). Critically, this temporary ambiguity could be resolved early by integrating the speaker's perspective on-line during comprehension. Analysis of eye-gaze to the potential referents revealed that instructions that were generated from the opposite spatial perspective posed challenges and delayed processing. However, despite these challenges, participants showed a clear target preference well before the onset of the disambiguating word (e.g., *hat*), demonstrating that even when spatial perspectives are misaligned, listeners are able to use knowledge about the speaker's spatial viewpoint to interpret their utterances as they unfold.

1.3. Present research

Though we know that listeners can remember one speaker's spatial perspective and use it to interpret language online, little else is known about how comprehension processes and spatial perspective-taking processes interact. How might a listener represent multiple perspectives and switch between them, as one often has to do when conversing with multiple people who all have varying perspectives on the visual array in question? Additionally, is the task of tracking these perspectives made more difficult when the speakers' perspectives are more dissimilar from the listener's (e.g., a 135° rotation vs. a 45° rotation)?

Because a given speaker's viewpoint is likely to be relatively stable over the course of a conversation, listeners can predict, with some certainty, that their conversational partner will continue using the same perspective throughout the dialogue. It may then be computationally efficient to store memories of an interlocutor's perspective along with other cues tied to speaker identity. Previous findings of angular disparity effects (e.g., Kessler & Rutherford, 2010; Kessler & Thomson, 2010; Surtees et al., 2013) indicate that storing or accessing this representation will be more challenging when it requires a larger rotation.

In the present research, we examine whether listeners encode spatial perspectives that differ from their own in a partner-specific way. We test this in a spatial perspective-taking paradigm where listeners hear instructions that alternate between two perspectives and move items onto a target location. In one case, each perspective is uniquely tied to one individual. This is analogous to when you are sitting at a table and listening to two people on either side of you—their viewpoints remain consistent throughout the conversation. In the other case, participants switch between two perspectives that are tied to one individual—imagine you are sitting at a table and one person is moving from one side to another while holding the conversation. If listeners do use speaker identity as a cue to their representation of a speaker's spatial perspective, the case in which there is a one-to-one mapping between speakers and perspectives should facilitate perspective-taking for the listener. The one-to-one mapping would support distinct representations of each spatial perspective, making them easier to access during on-line sentence comprehension. On the other hand, if spatial perspectives are not stored partner-specifically, the speaker's voice should not be a helpful cue to accessing the relevant spatial perspective, and as a result, there should be no processing benefit in the two-speaker case. In our experiments, this facilitation will be reflected in more fixations to the perspective-appropriate target in the two-speaker case as compared to the one-speaker case.

In the first experiment, we also examine the effect of angular disparity on online comprehension of the spatial term. We predict that, if participants make use of the unique speaker-perspective mappings, they may be particularly helpful in situations where the angular disparity is greatest. When the angular disparity is small, the advantage provided by storing perspectives partner-specifically and accessing them online compared to simply applying the mental rotation *de novo* may be minimal.

2. Experiment 1

2.1. Method

2.1.1. Participants

Forty-nine undergraduate students at the University of Illinois at Urbana-Champaign participated in this experiment in exchange for partial course credit. Participants had normal or corrected-to-normal vision and spoke English fluently.

2.1.2. Materials

Participants completed a spatial perspective-taking task on a desktop computer while their eye-movements were monitored using an Eyelink-1000 desktop-mounted eye-tracker at 1000 Hz. Stimulus presentation was controlled using Matlab's Psychophysics Toolbox 3 (PTB-3, Brainard, 1997). On each trial, participants saw a display with a variety of circles and triangles[1] (Fig. 1) and listened to pre-recorded instructions about which object to move around the screen. They were instructed to imagine that the display was actually laid out on a table in front of them and that they and the person giving them the instructions was looking at this table as well. Participants were also told that there might be one person speaking or two people alternating giving instructions.

At the start of each trial, an arrow appeared in one of the four corners of the screen. This arrow indicated which viewpoint the audio instruction would be given from. Relative to the listener's viewpoint, the arrow could be at an angle of 45°, 135°, 225° (as in Fig. 1), or 315°. The audio instruction began playing at the same time as the display, including the arrow, appeared. Instructions were of the following form: "Move the [shape] with the [color][pattern] to the [DIRECTION TERM] onto the [shape] with the [color][pattern]".[2] For example, a participant might hear "Move the circle with the green crosses to the right onto the circle with the purple dots." The destination of the instruction—in the preceding example, the circle with the purple dots—was coded as the "target" object.

The displays and audio instructions were designed such that unless the listener correctly interpreted the direction term, the instruction was temporarily ambiguous between multiple potential objects. On all trials, this potential ambiguity was lexically resolved at the final word in the instruction, e.g., *purple dots*. The temporary ambiguity was created by placing a competitor of the same shape and color in the opposite direction of the target. For example, in Fig. 1, a circle with *purple lines* is placed to the *left* of the circle with the green crosses (from the perspective indicated by the arrow). Thus if a listener did not correctly interpret the spatial term (*left*), the instruction would be ambiguous until the final word "dots". The onsets of the direction term (*left*), the target shape word (*circle [with the purple dots]*), and the target pattern (*dots*) were identified by hand using Praat software (Boersma & Weenink, 2012), which allowed eye-movement analyses to be time-locked to the audio instructions.

Participants saw each specific array of objects (e.g., the layout in Fig. 1) for 8 trials in a row. After completing the instruction on each trial, the dragged object popped back into its original position for the beginning of the next trial. After 8 trials, the array of objects

[1] These shapes were chosen because they do not have an intrinsic up, down, left, or right.

[2] Shapes: circle or triangle; colors: blue, green, red, orange, purple, or yellow; patterns: dots, crosses, stars, or lines; direction terms: left, right, forward, backward.

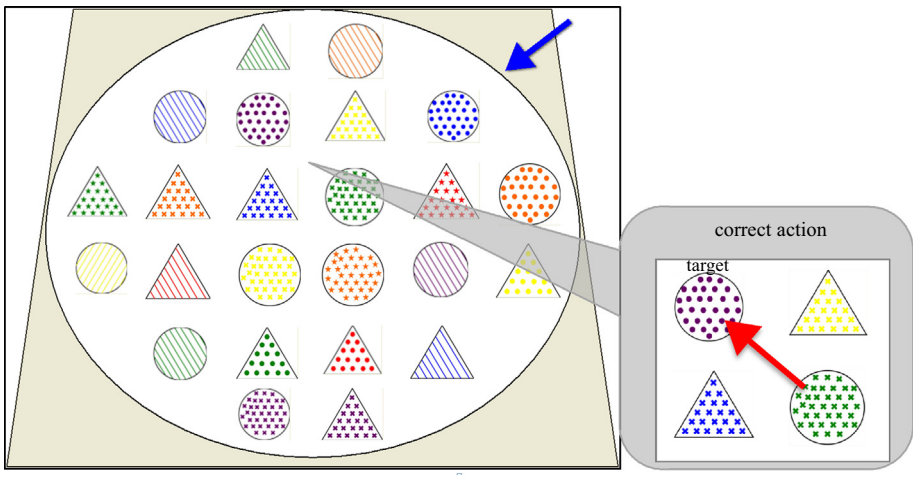

Fig. 1. Example array seen by participants and the correct action they should execute after hearing the instruction "*Move the circle with the green crosses to the right onto the circle with the purple dots*." The beige-colored "tabletop" converges toward the top of the computer display to give the appearance of depth. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

changed. The layout of the objects in each array was random except for two constraints: (1) Only one exemplar of each object type (e.g., circle with green crosses) appeared in a given array. (2) Each array contained eight target–competitor pairs (e.g., two circles with a purple pattern separated by one to-be-dragged shape with a different color) allowing for eight trials in a row to occur without changing the overall display. Participants completed two blocks of trials. Each block contained 11 unique arrays, each of which contained 8 trials, for a total of 176 trials (88 trials per block) per participant.

2.1.3. Experimental design

Within each test block, participants followed instructions from one of two spatial perspectives. The critical manipulation in this study was whether those spatial perspectives were yoked to a single talker or to different talkers. Thus, between blocks we manipulated whether participants always heard one speaker (One Speaker condition) giving instructions from each perspective (i.e., the same voice gave instructions regardless of where the arrow was), or two speakers (Two Speaker condition) gave instructions and each speaker was associated with a particular perspective (e.g., one voice gave instructions when the arrow was at 135° and the other voice gave instructions when the arrow was at 225°). Participants were informed before the start of each block if they would be hearing one voice or two.

In order to create the two conditions, instructions from three different speakers were recorded—two female voices (A and B) and one male voice (C). In the One Speaker condition participants heard either voice A or B, and in the Two Speaker condition participants heard B and C, or A and C. The male–female voice contrast in the Two Speaker condition was used so that participants had no doubts that there were two different voices. The speaker manipulation was within-subjects, and each participant heard all three different voices over the course of the experiment (for a given subject a particular voice was heard in only one of the conditions).

Finally, in order to reduce interference across blocks, in one test block the instructions were given from the two top angles (135° and 225°), and in the other block, instructions were given from the two bottom angles (45° and 315°). Two versions of each array were created so that each array (and trial) appeared both in the Top Angles condition and the Bottom Angles condition between subjects. This was done by rotating each array by 180° and keeping the auditory stimulus exactly the same. Number of Speakers and Angle pair were fully crossed with the version of the array, and the order of conditions and voices was counterbalanced, resulting in sixteen experimental lists (Appendix A).

Within a test-block, instructions alternated between the two angles pseudo-randomly. Based on previous findings that switching between spatial viewpoints poses challenges (Ryskin et al., 2014), for each test trial, we coded whether the previous trial was given from the same perspective (No Switch, e.g., 135° then 135°), or from the alternative perspective (Switch, e.g., 135° then 225°). Block order (one speaker vs. two speaker; top angles vs. bottom angles) was counterbalanced across participants. Each participant was tested on a single list.

2.2. Analysis and results

Participants were successful at the spatial perspective-taking task. Accuracy for dragging and dropping the objects was 95%. The dependent measure used to index spatial perspective-taking was the eye-movements that participants made as they interpreted the potentially ambiguous instruction (e.g., *Move the circle with the green crosses to the right onto the circle with the purple dots*). Eye movements associated with the interpretation of spatial perspective were analyzed in terms of a binary measure: whether the participant fixated the target during the specified time window or not.[3] A fixation was coded as a target-fixation if the *x*, *y* fixation-coordinates landed on the target object (e.g., *the circle with the purple dots*), or on a small portion of the white space surrounding it (this buffer space did not overlap with any other object). See the Online Supplement for summary figures describing fixations to the other objects on the screen.

In order to examine both early and late processing effects, target fixations were measured in three consecutive time windows. The first time window (average duration 1550 ms) began at the onset of the direction term (e.g., *right*) and ended at the onset of the target shape (e.g., *circle*). The second time window (average duration 1700 ms) began at the onset of the target shape term and ended at the onset of the pattern term (e.g., *dots*). The third

[3] Note that eye-gaze analyses in language tasks sometimes use a proportion measure (proportion of target fixations on each trial in a specified time-window). We used a different approach because inspection of the proportion-based measure revealed that the data distribution was highly zero-inflated and would violate the linear model assumption of normally distributed residuals.

time window began at the onset of the pattern term and ended 1500 ms after the onset of the pattern term. The first two windows captured interpretation of the potentially ambiguous portion of the critical instruction, with the first window focusing on interpretation of the spatial term, and the second focusing on the ambiguous noun. The third time window captures any processing that may occur post-lexical disambiguation. The time windows were all offset by 200 ms due to the time needed to program and launch an eye movement (Hallett, 1986). The proportion of trials with a target fixation in each time-window are plotted by Speaker condition (1 Speaker vs. 2 Speakers), as well as Angle pair (Top angles vs. Bottom angles), and Switching condition (no Switch vs. Switch) in Figs. 2–4.

For each time window, the proportion of trials with a target fixation were analyzed in a multilevel logistic regression, using the *lme4* software package in R (Bates, Maechler, Bolker, & Walker, 2014). Speaker, Angle pair, and Switching, along with their three-way interaction and all two-way interactions, were entered as fixed effects (Table 1) with subjects and trials as random effects. All fixed effects were coded with mean-centered contrast codes. When the maximal random effects structure justified by the design did not converge, random slopes with the least variance were removed until the model converged (see Appendix C). Model comparison was used to assess the significance of effects.

In the first time window (*Move the circle with the green crosses to the right onto the circle with the purple dots*), there was a main effect

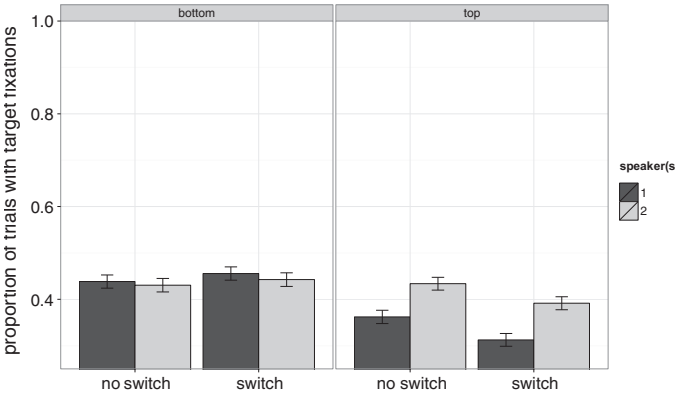

Fig. 2. Proportion of trials with target fixations in the first time window (e.g., *right onto the*). Error bars represent standard error of the mean, calculated using the method from Morey (2008).

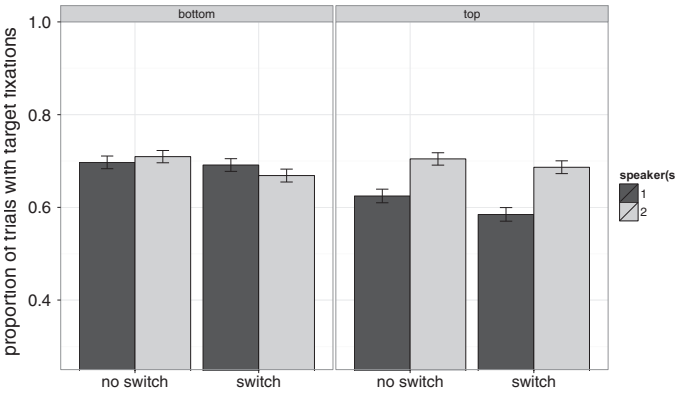

Fig. 3. Proportion of trials with target fixations in the second time window (e.g., *circle with the purple*). Error bars represent standard error of the mean, calculated using the method from Morey (2008).

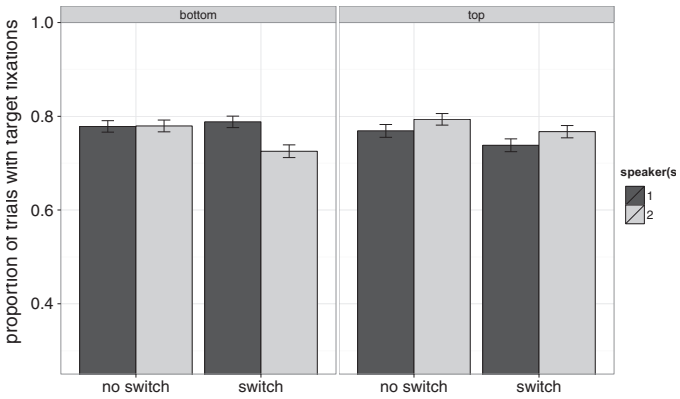

Fig. 4. Proportion of trials with target fixations in the third time window (e.g., *dots + 1500 ms*). Error bars represent standard error of the mean, calculated using the method from Morey (2008).

Table 1
Experiment 1: Results of the logistic mixed-effects model of target fixations across three time windows. See Appendix C for random effects. Reported *z*-values based on Laplace approximation estimates; χ^2 values and corresponding *p*-values are based on model comparison. *Note:* * indicates effects that are significant at an alpha level of 0.05.

Fixed effects	β	SE	*z*-Value	χ^2 (1)	*p*-Value
First Time Window					
(Intercept)	−0.464	0.170	−2.725		
Speaker	0.165	0.159	1.039	1.088	0.297
Angle pair	−0.330	0.162	−2.033	3.961	0.046*
Switching	−0.093	0.083	−1.115	1.073	0.300
Speaker × Angle pair	0.471	0.668	0.705	0.499	0.481
Speaker × Switching	0.013	0.104	0.124	0.017	0.897
Angle pair × Switching	−0.402	0.125	−3.229	14.212	1.63e−4*
Speaker × Angle pair × Switching	0.048	0.208	0.231	0.053	0.819
Second Time Window					
(Intercept)	0.994	0.161	6.163		
Speaker	0.304	0.139	2.191	4.617	0.032*
Angle pair	−0.215	0.139	−1.546	2.310	0.129
Switching	−0.178	0.082	−2.182	4.410	0.036*
Speaker × Angle pair	0.497	0.637	0.781	0.620	0.431
Speaker × Switching	−0.050	0.107	−0.473	0.220	0.639
Angle pair × Switching	−0.062	0.107	−0.579	0.331	0.565
Speaker × Angle pair × Switching	0.253	0.253	1.000	0.970	0.325
Third Time Window					
(Intercept)	1.619	0.121	13.333		
Speaker	0.008	0.059	0.129	0.016	0.898
Angle pair	−0.072	0.059	−1.224	1.475	0.225
Switching	−0.170	0.076	−2.230	4.882	0.027*
Speaker × Angle pair	0.519	0.475	1.092	1.175	0.278
Speaker × Switching	−0.237	0.118	−2.014	3.995	0.046*
Angle pair × Switching	−0.068	0.118	−0.572	0.323	0.570
Speaker × Angle pair × Switching	0.297	0.236	1.259	1.559	0.212

of Angle pair, such that participants made fewer target fixations in the Top Angles condition. A significant interaction of Angle pair and Switching was due to a significant Switching effect in the Top Angles condition ($\beta = -0.29$, $p < 0.005$), but no effect of Switching in the Bottom Angles condition ($\beta = 0.10$, $p = 0.33$). In the Top Angles condition, participants made fewer fixations to the target when the previous trial had been in a different perspective.

In the second time window (*Move the circle with the green crosses to the right onto the circle with the purple dots*), there was a main effect of Speaker such that participants in the Two Speaker

condition made more fixations to the target than participants in the One Speaker condition. Participants also made fewer target fixations when they had to switch perspectives.

In the third time window (*Move the circle with the green crosses to the right onto the circle with the purple dots + 1500 ms*), participants made fewer target fixations when they had to switch perspectives. This Switching effect was primarily driven by the Two Speaker condition ($\beta = -0.31$, $p < 0.001$). There was no significant effect of switching in the One Speaker condition ($\beta = -0.05$, $p = 0.60$).

2.3. Discussion

We hypothesized that listeners could encode spatial perspectives and bind those representations in memory to a particular speaker. If so, we predicted that listeners should then use speaker identity as a cue to efficiently access these stored perspective representations during language processing. Consistent with this hypothesis, we found that on-line interpretation was facilitated when each perspective was mapped to a particular speaker, compared to when two perspectives were mapped to the same speaker. In addition, this experiment replicated previous findings of a cost associated with switching between spatial perspectives (Ryskin et al., 2014). The fact that switching between spatial perspectives incurs costs is generally consistent with the broader conclusion that listeners maintained stored representations of spatial perspective and used these stored perspectives to guide online processing.

It is also worth noting that the effects of Switching and Speaker seem to be driven primarily by the Top angles. Despite the fact that the means are suggestive of a Speaker by Angle Pair interaction (see Figs. 2–4), the statistical analyses[4] do not lend support to our prediction that the benefits of speaker-to-perspective binding would be larger in the Top angles. On the other hand, Angle Pair does interact significantly with Switching such that the Switching effect is absent for the Bottom angles. This may be the result of a ceiling effect for the Bottom angles, consistent with previous findings of greater ease of perspective-taking when the angular disparity is small (e.g., Kessler & Thomson, 2010). However, given that the targets in the Bottom angle conditions are closer to the egocentric competitors (i.e., the cells most consistent with an egocentric interpretation of the direction term), it is difficult to rule out the possibility that this ceiling effect might be due to partial overlap between the speaker perspective and the egocentric perspective (e.g., "right" from 45° is "right and down" from 0° vs. "right" from 135° is "left and down" from 0°). As a result, in our second experiment, we only test the Top angles (135° and 225°).

Finally, while the results of Experiment 1 provide key evidence that listeners can use remembered spatial perspectives to guide online language processing, there exists an alternate explanation of the critical two-speaker advantage. Recall that the arrow cue appeared on-screen at the same time as the spoken instruction. As a result, in the One Speaker condition, participants needed to locate the arrow first, in order to interpret the instruction. By contrast, participants in the Two Speaker condition could simply use the voice cue. On this interpretation, participants did remember which spatial perspective is associated with each talker in the Two Speaker condition, but this association only served the same function as the arrow cues with no spatial representations attached. The difference between One and Two Speakers is caused by the need to fixate the arrow in the One Speaker condition to know which perspective to take, which led to a downstream delay in processing the sentence online and locating the target.

[4] Note the large variability for the Speaker × Angle Pair interaction in Table 1 and Appendix C.

While participants did not spend a lot of time looking at the arrows while they listened to the instruction (Supplementary Figs. 1 and 2), the design of the experiment was such that recording of the eye movements began at the same time as the onset of the audio instructions. This did not allow us to capture eye-movements following the appearance of the new display and before the onset of the audio. It could be that participants gazed at the arrow following scene onset more in the One Speaker condition, limiting inspection of the scene. This could have led to the downstream comprehension slow-down for participants in the One Speaker condition. To address this possibility, Experiment 2 was designed as a replication of Experiment 1, but it included a delay between the presentation of the arrow (and beginning of eye-movement recording) and the onset of the auditory stimulus. This delay would allow participants in both Speaker conditions enough time to look at the arrow before any sentence comprehension processes need to be engaged.

3. Experiment 2

The aim of Experiment 2 was to replicate the two-speaker advantage observed in Experiment 1, while allowing plenty of time for participants to encode the spatial perspective cue prior to interpretation of the critical instruction.

3.1. Method

3.1.1. Participants

Forty-eight undergraduate students at the University of Illinois at Urbana-Champaign participated in this experiment in exchange for partial course credit. Participants had normal or corrected-to-normal vision and spoke English fluently.

3.1.2. Materials and design

The design of this experiment was identical to Experiment 1 except that the Angle Pair factor was removed. Only the top angles (135°, 225°) were used. As a result, the number of experimental lists was reduced to eight (Appendix B). Critically, a delay of 1500 ms was introduced between the appearance on the screen of the display with the arrow and the onset of the auditory stimulus for every trial.

3.2. Analysis and results

Participants were successful at the spatial perspective-taking task. Accuracy for dragging and dropping the objects was 97%. As in the first experiment, the dependent measure was the eye-movements that participants made as they interpreted the potentially ambiguous instruction. Target fixations were analyzed during the same three time windows used in Experiment 1 (region 1: direction term; region 2: target shape; region 3: pattern term). The average proportions of trials with a target fixation in each of the three time windows is plotted by Speaker (1 Speaker vs. 2 Speakers) and Switching condition (no Switch vs. Switch) in Figs. 5–7. Plots of fixations to other objects on the screen are presented in the Online Supplement.

For each time window, the proportion of trials with a target fixation was analyzed in a multilevel logistic regression as before (see Table 2). Speaker, Switching, and their interaction were entered as fixed effects coded with mean-centered contrast codes. Random intercepts were entered for subjects and trials, with random by-subjects slopes for Speaker, Switching, and their interaction, and by-trials slopes for Speaker condition. Model comparison was used to assess the significance of effects.

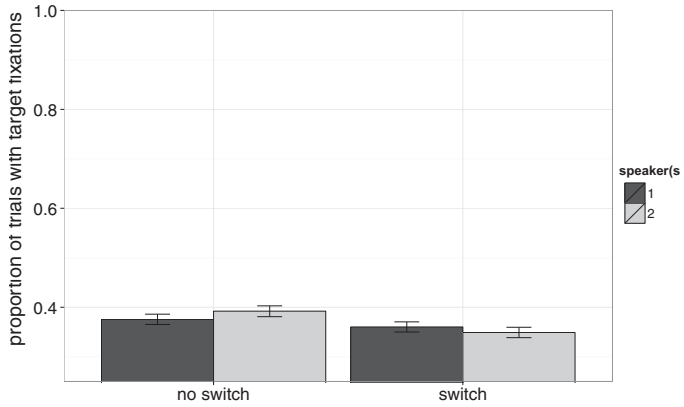

Fig. 5. Proportion of trials with target fixations in the first time window (e.g., *right onto the*). Error bars represent standard error of the mean, calculated using the method from Morey (2008).

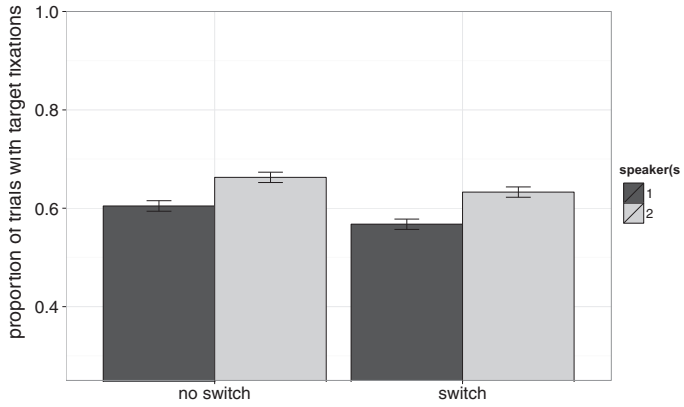

Fig. 6. Proportion of trials with target fixations in the second time window (e.g., *circle with the purple*). Error bars represent standard error of the mean, calculated using the method from Morey (2008).

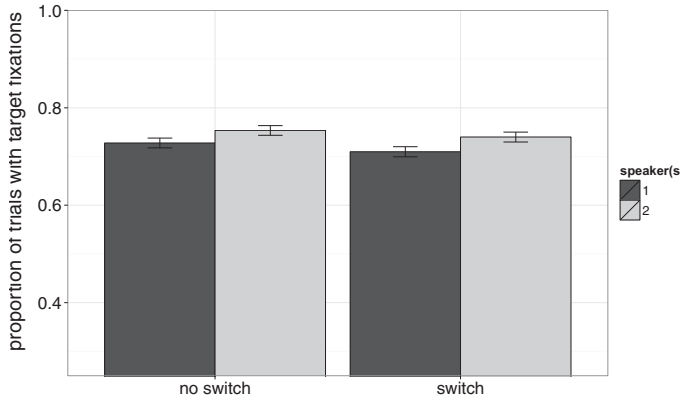

Fig. 7. Proportion of trials with target fixations in the third time window (e.g., *dots + 1500 ms*). Error bars represent standard error of the mean, calculated using the method from Morey (2008).

In the first time window (*Move the circle with the green crosses to the right onto the circle with the purple dots*), there was a marginal effect of Switching, such that participants were less likely to fixate the target when they had just switched perspectives.

Table 2
Experiment 2: Results of the logistic mixed-effects model of target fixations across three time windows (see Appendix D for random effects). Reported *z*-values come from Laplace approximation estimates; the χ^2 values and corresponding *p*-values come from model comparison. *Note:* * indicates effects that are significant at an alpha level of 0.05 and † indicates effect that are marginally significant.

Fixed effects	β	SE	*z*-Value	χ^2 (1)	*p*-Value
First Time Window					
(Intercept)	−0.694	0.204	−3.408		
Speaker	−0.038	0.176	−0.218	0.048	0.827
Switching	−0.184	0.095	−1.927	3.607	0.058†
Speaker × Switching	−0.181	0.117	−1.550	2.359	0.125
Second Time Window					
(Intercept)	0.721	0.198	3.639		
Speaker	0.314	0.150	2.092	4.147	0.042*
Switching	−0.163	0.093	−1.755	2.977	0.084†
Speaker × Switching	0.060	0.135	0.444	0.193	0.660
Third Time Window					
(Intercept)	1.542	0.170	9.063		
Speaker	0.125	0.119	1.057	1.076	0.300
Switching	−0.089	0.086	−1.031	1.034	0.309
Speaker × Switching	0.086	0.137	0.627	0.383	0.536

In the second time window (*Move the circle with the green crosses to the right onto the circle with the purple dots*), there was a main effect of Speaker such that participants in the Two Speaker condition made more fixations to the target than participants in the One Speaker condition. There was also marginal effect of Switching such that participants were less likely to fixate the target after switching.

There were no significant effects in the third time window (*Move the circle with the green crosses to the right onto the circle with the purple dots + 1500 ms*); target fixations were uniformly high following the interpretation of the disambiguating pattern word.

3.3. Discussion

The results of Experiment 2 largely replicate the key findings of Experiment 1. Even with the added delay between the visual and auditory stimuli, perspective-taking was facilitated when each perspective was mapped to a distinct speaker, compared to when two perspectives are mapped to the same speaker. This finding suggests that listeners' representations of perspective can be tied to the individual speaker and that encoding representations in this way makes accessing them on-line more efficient. Experiment 2 also replicates the effect of switching such that switching from one perspective to another is accompanied by an additional cost to processing perspective—laden language.

4. General discussion

Taking into account the spatial perspective of an interlocutor is an essential skill necessary for successful communication. In the present work, we examined how this process unfolds in real time. We hypothesized that listeners would store a representation of each speaker's perspective bound to the identity of that speaker. This partner-specific encoding would allow listeners to flexibly retrieve the appropriate representations and use spatial perspective information to constrain their interpretation during online processing of a sentence.

Across two eye-tracking experiments, we find support for our hypotheses. Participants in a spatial perspective-taking task make more predictive target fixations when each perspective is associated with a specific speaker than when one speaker alternates between two perspectives. We conclude from this that representations of a speaker's spatial perspective on a visual array can be

bound to the speaker's identity in memory and that listeners access these partner-specific representations online during comprehension.

These results point to interesting future avenues of research. The facilitation provided by these partner-specific representations relies on the listener's capacity to bind representations in memory. However, there are limitations on the number of bindings that can be learned by an individual and kept distinct in memory (e.g., Anderson & Reder, 1999). Presumably, there exists a number of speakers for which there is no longer any advantage, and perhaps even a disadvantage, to storing partner-specific perspectives in memory. On the other hand, voices are most likely not the only way to cue partner-specific storage of spatial perspectives. The physical location and orientation of a speaker may provide even stronger cues. However, they might elicit renewed computation of a perspective, rather than retrieval from memory. Further work is needed to address the relative contributions of auditory and visuo-spatial cues to the on-line retrieval of stored perspectives.

Our findings of a partner-specific encoding of perspective have important implications for theories of how information about interlocutors is stored in memory and accessed on-line in the service of efficient communication. The results presented here provide support for a model of language comprehension in which the speaker's identity provides a contextual constraint on the set of possible memory representations that can be brought to be bear on the interpretation of an utterance (Brown-Schmidt, Yoon, & Ryskin, 2015). Moreover, in our studies, participants could have completely ignored the fact that there were different voices and still completed the task successfully, because the arrows provide all the necessary perspective information to interpret each instruction. The fact that participants did not ignore the voice information suggests that encoding of new speaker-specific representations occurs spontaneously, as a dialog unfolds; No explicit suggestion about paying attention to the different voices was needed to elicit this behavior and it is unlikely that the participants assumed that paying attention to the different voices would confer a processing advantage. Though we certainly do not claim that storing such spatial perspective representations is something that listeners *always* do in conversation (our data cannot speak to that), it seems plausible that they would be all the more inclined to do so outside the laboratory setting, where cues to spatial perspective differences are more salient and sentences are not always eventually disambiguated. Indeed, Samson, Apperly, Braithwaite, Andrews, and Bodley Scott (2010) show that listeners compute an agent's perspective spontaneously, even when it is irrelevant to the task at hand. However it is important to note that the form of perspective-taking examined in Samson et al. represents Level 1 perspective-taking, which is thought to be less cognitively taxing and thus may lend itself more readily to automatic processing. Our data suggest that, when the agent's perspective is stable, listeners take advantage of being able to ease the computational burden by tying the perspective to the agent in memory. An interesting avenue for future research might be to identify the constraints that modulate when listeners do or do not store spatial perspective information speaker-specifically.

Further, our results are consistent with prior findings of partner-specific stored information (Creel, 2014; Creel, Aslin, & Tanenhaus, 2008; Creel & Tumlin, 2011 Dahan, Drucker, & Scarborough, 2008; Kamide, 2012; Trude & Brown-Schmidt, 2012; Van Berkum, van den Brink, Tesink, Kos, & Hagoort, 2008) and contribute to the existing knowledge about the forms that partner-specific representations can take. Encoding a speaker-bound spatial perspective likely requires the computation of a new reference frame (Sohn & Carlson, 2003; Avraamides et al., 2007). This indicates that partner-specific representations can range in complexity from low-level, perceptual information (e.g., the acoustic features of an individual's pronunciation; Trude & Brown-Schmidt, 2012) to high-level, relational concepts (e.g., the relationships between objects in the physical environment and another person). A more precise understanding of how these representations are formed and structured will help to answer questions about the attentional and memorial limitations on perspective-taking and bring us closer to implementable models of how perspective-taking occurs.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. NSF 12-57029 to S. Brown-Schmidt. Thanks to Ariel N. James and Daniel H. Katz for recording auditory stimuli and Phoebe Bauer for help with collecting data.

Appendix A. Experiment 1 Design

List	Array	Block 1			Block 2		
		Number of speakers	Female voice	Angle Pair	Number of speakers	Female voice	Angle Pair
1	v1	1T	A	Top	2T	B	Bottom
2	v1	1T	A	Bottom	2T	B	Top
3	v1	1T	B	Top	2T	A	Bottom
4	v1	1T	B	Bottom	2T	A	Top
5	v1	2T	A	Top	1T	B	Bottom
6	v1	2T	A	Bottom	1T	B	Top
7	v1	2T	B	Top	1T	A	Bottom
8	v1	2T	B	Bottom	1T	A	Top
9	v2	1T	A	Top	2T	B	Bottom
10	v2	1T	A	Bottom	2T	B	Top
11	v2	1T	B	Top	2T	A	Bottom
12	v2	1T	B	Bottom	2T	A	Top
13	v2	2T	A	Top	1T	B	Bottom
14	v2	2T	A	Bottom	1T	B	Top
15	v2	2T	B	Top	1T	A	Bottom
16	v2	2T	B	Bottom	1T	A	Top

Appendix B. Experiment 2 Design

List	Array	Block 1			Block 2		
		Number of speakers	Female voice	Angle pair	Number of speakers	Female voice	Angle pair
1	v1	1T	A	Top	2T	B	Top
2	v1	1T	B	Top	2T	A	Top
3	v1	2T	A	Top	1T	B	Top
4	v1	2T	B	Top	1T	A	Top
5	v2	1T	A	Top	2T	B	Top
6	v2	1T	B	Top	2T	A	Top
7	v2	2T	A	Top	1T	B	Top
8	v2	2T	B	Top	1T	A	Top

Appendix C. Experiment 1: Random effects for generalized linear model analyses in three time windows.

Random effects		Variance
First Time Window		
Participants	(Intercept)	1.294
	Speaker	0.445
	Angle pair	0.638
	Speaker × Angle pair	0.662
Items	(Intercept)	0.180
	Angle pair	0.198
Second Time Window		
Participants	(Intercept)	0.999
	Speaker	0.768
	Speaker × Angle pair	3.186
	Speaker × Angle pair × Switching	0.664
Items	(Intercept)	0.114
Third Time Window		
Participants	(Intercept)	0.638
Items	(Intercept)	0.100

Note: Maximal random effects structure justified by the design was used in all models. When the maximal model did not converge, the random component with the least variance was removed and the model was refit.

Appendix D. Experiment 2: Random effects for generalized linear model analyses in three time windows.

Random effects		Variance
First Time Window		
Participants	(Intercept)	1.883
	Speaker	1.293
	Switching	0.042
	Speaker × Switching	0.007
Items	(Intercept)	0.214
	Speaker	0.031
Second Time Window		
Participants	(Intercept)	1.791
	Speaker	0.888
	Switching	0.055
	Speaker × Switching	0.174
Items	(Intercept)	0.183
	Speaker	0.064
Third Time Window		
Participants	(Intercept)	1.287
	Speaker	0.426
	Switching	0.009
	Speaker × Switching	0.063
Items	(Intercept)	0.125
	Speaker	$3.1e^{-5}$

Note: Maximal random effects structure justified by the design was used in all models.

Appendix E. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2015.11.011.

References

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General, 128*, 186–197. http://dx.doi.org/10.1037/0096-3445.128.2.186.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*, 953–970. http://dx.doi.org/10.1037/a0016923.

Avraamides, M. N., Ioannidou, L. M., & Kyranidou, M. N. (2007). Locating targets from imagined perspectives: Comparing labelling with pointing responses. *Quarterly Journal of Experimental Psychology, 60*, 1660–1679. http://dx.doi.org/10.1080/17470210601121833.

Avraamides, M. N., Theodorou, M., Agathokleous, A., & Nicolaou, A. (2013). Revisiting perspective-taking: Can people maintain imagined perspectives? *Spatial Cognition & Computation, 13*, 50–78. http://dx.doi.org/10.1080/13875868.2011.639915.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7. <http://CRAN.R-project.org/package=lme4>.

Boersma, P., & Weenink, D. (2012). *Praat: Doing phonetics by computer [Computer program]*. Version 5.3.23. <http://www.praat.org/> Retrieved 07.08.12.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Brockmole, J., & Wang, R. (2003). Changing perspective within and across environments. *Cognition, 87*, 59–67. http://dx.doi.org/10.1016/S0.

Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychonomic Bulletin & Review, 16*, 893–900. http://dx.doi.org/10.3758/PBR.16.5.893.

Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes, 27*, 62–89.

Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition, 107*, 1122–1134. http://dx.doi.org/10.1016/j.cognition.2007.11.005.

Brown-Schmidt, S., Yoon, S. O., & Ryskin, R. A. (2015). People as contexts in conversation. In B. H. Ross (Ed.). *The psychology of learning and motivation* (Vol. 62, pp. 59–99). Academic Press. http://dx.doi.org/10.1016/bs.plm.2014.09.003.

Clark, H. H. (1992). *Arenas of language use*. University of Chicago Press.

Creel, S. C. (2014). Preschoolers' flexible use of talker information during word learning. *Journal of Memory and Language, 73*, 81–98. http://dx.doi.org/10.1016/j.jml.2014.03.001.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition, 106*, 633–664. http://dx.doi.org/10.1016/j.cognition.2007.03.013.

Creel, S. C., & Tumlin, M. A. (2011). On-line acoustic and semantic interpretation of talker information. *Journal of Memory and Language, 65*, 264–285. http://dx.doi.org/10.1016/j.jml.2011.06.005.

Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition, 108*, 710–718. http://dx.doi.org/10.1016/j.cognition.2008.06.003.

Easton, R., & Sholl, M. (1995). Object-array structure, frames of reference, and retrieval of spatial knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 483–500.

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. *Developmental Psychology, 17*, 99–103. http://dx.doi.org/10.1037/0012-1649.17.1.99.

Galati, A., Michael, C., Mello, C., Greenauer, N. M., & Avraamides, M. N. (2013). The conversational partner's perspective affects spatial memory and descriptions. *Journal of Memory and Language, 68*, 140–159. http://dx.doi.org/10.1016/j.jml.2012.10.001.

Hallett, P. E. (1986). Eye movements. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.). *Handbook of perception and human performance* (Vol. 1, pp. 10.1–10.112). New York: Wiley.

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. http://dx.doi.org/10.1016/S0749-596X(03)00022-6.

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition, 108*, 831–836. http://dx.doi.org/10.1016/j.cognition.2008.04.008.

Huttenlocher, J., & Presson, C. C. (1973). Mental rotation and the perspective problem. *Cognitive Psychology, 4*, 277–299. http://dx.doi.org/10.1016/0010-0285(73)90015-7.

Kamide, Y. (2012). Learning individual talkers' structural preferences. *Cognition, 124*, 66–71. http://dx.doi.org/10.1016/j.cognition.2012.03.001.

Kessler, K., & Rutherford, H. (2010). The two forms of visuo-spatial perspective taking are differently embodied and subserve different spatial prepositions. *Frontiers in Psychology, 1*, 1–12. http://dx.doi.org/10.3389/fpsyg.2010.00213.

Kessler, K., & Thomson, L. A. (2010). The embodied nature of spatial perspective taking: Embodied transformation versus sensorimotor interference. *Cognition, 114*, 72–88. http://dx.doi.org/10.1016/j.cognition.2009.08.015.

Levine, M., Jankovic, I. N., & Palij, M. (1982). Principles of spatial problem solving. *Journal of Experimental Psychology: General, 111*(2), 157–175. http://dx.doi.org/10.1037/0096-3445.111.2.157.

Masangkay, Z. S., McCluskey, K. A., McIntyre, C. W., Sims-Knight, J., Vaughn, B. E., & Flavell, J. H. (1974). The early development of inferences about the visual percepts of others. *Child Development, 45*, 357–366. http://dx.doi.org/10.1111/1467-8624.ep12154629.

May, M. (2004). Imaginal perspective switches in remembered environments: Transformation versus interference accounts. *Cognitive Psychology, 48*, 163–206. http://dx.doi.org/10.1016/S0010-0285(03)00127-0.

Michelon, P., & Zacks, J. M. (2006). Two kinds of visual perspective taking. *Perception & Psychophysics, 68*, 327–337.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*, 61–64. http://dx.doi.org/10.3758/s13414-012-0291-2.

Mou, W., McNamara, T. P., Valiquette, C. M., & Rump, B. (2004). Allocentric and egocentric updating of spatial memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 142–157. http://dx.doi.org/10.1037/0278-7393.30.1.142.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science, 13*, 329–336. http://dx.doi.org/10.1111/1467-9280.00460.

Rieser, J. (1989). Access to knowledge of spatial structure at novel points of observation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 1157–1165.

Ryskin, R. A., Brown-Schmidt, S., Canseco-Gonzalez, E., Yiu, L. K., & Nguyen, E. T. (2014). Visuospatial perspective-taking in conversation and the role of bilingual experience. *Journal of Memory and Language, 74*, 46–76. http://dx.doi.org/10.1016/j.jml.2014.04.003.

Salatas, H., & Flavell, J. H. (1976). Perspective taking: The development of two components of knowledge. *Child Development, 47*, 103–109.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance, 36*, 1255–1266. http://dx.doi.org/10.1037/a0018729.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition, 47*, 1–24.

Schober, M. (2009). Spatial dialogue between partners with mismatched abilities. In K. R. Coventry, T. Tenbrink, & J. A. Bateman (Eds.), *Spatial language and dialogue* (pp. 23–39). Oxford: Oxford University Press.

Shelton, A. L., & McNamara, T. P. (1997). Multiple views of spatial memory. *Psychonomic Bulletin & Review, 4*, 102–106. http://dx.doi.org/10.3758/BF03210780.

Shelton, A. L., & McNamara, T. P. (2001). Systems of spatial reference in human memory. *Cognitive Psychology, 43*, 274–310.

Simons, D., & Wang, R. (1998). Perceiving real-world viewpoint changes. *Psychological Science, 9*, 315–320.

Sohn, M.-H., & Carlson, R. A. (2003). Viewpoint alignment and response conflict during spatial judgment. *Psychonomic Bulletin & Review, 10*, 907–916.

Surtees, A., Apperly, I., & Samson, D. (2013). Similarities and differences in visual and spatial perspective-taking processes. *Cognition, 129*, 426–438.

Taylor, H. A., & Tversky, B. (1992). Spatial mental models derived from survey and route descriptions. *Journal of Memory and Language, 31*, 261–292. http://dx.doi.org/10.1016/0749-596X(92)90014-O.

Trude, A. M., & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes, 27*, 979–1001. http://dx.doi.org/10.1080/01690965.2011.597153.

Tversky, B., & Hard, B. M. (2009). Embodied and disembodied cognition: Spatial perspective-taking. *Cognition, 110*, 124–129. http://dx.doi.org/10.1016/j.cognition.2008.10.008.

Van Berkum, J. J. A., van den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience, 20*, 580–591. http://dx.doi.org/10.1162/jocn.2008.20054.

Wang, R. (2005). Beyond imagination: Perspective change problems revisited. *Psicólogica, 26*, 25–38.

Wang, R. F. (2007). Spatial processing and view-dependent representations. In *Spatial processing in navigation, imagery and perception* (pp. 49–65). New York: Springer.

Wang, R. F. (2012). Theories of spatial representations and reference frames: What can configuration errors tell us? *Psychonomic Bulletin & Review, 19*, 575–587. http://dx.doi.org/10.3758/s13423-012-0258-2.