5 *Maybe what it means is he actually got the spot*

Physical and cognitive viewpoint in a gesture study

Shweta Narayan

5.1 Introduction

Participants in a gesture study create meaning interactively. In the example analyzed in this chapter, one participant (Speaker), who can see an image, fails to understand it. She comes up with successively less wrong interpretations of the scene, but it is the other participant (Listener), who cannot see the image, who finally figures out the standard canonical interpretation of the image. In the course of the interaction, Listener's cognitive viewpoint shifts, as evidenced by his speech and physically viewpointed gestures. His viewpoint goes through four phases:

1) Taking Speaker's viewpoint: this is only clear in the *interaction* of speech and gesture.

   LISTENER: . . . to the left . . . ? [*gestures to his right (Speaker's left)*]

2) Misunderstanding Viewpoint Space: Speaker's viewpoint is of someone looking at the picture. Listener thinks it is of characters within the depicted scene.

   SPEAKER: On the right-hand side of the street, there are five cars . . .
   LISTENER: Hold on, you can't say the right side of the street, because whatever side of the street you're on is gonna be the right side.

3) Switching visual viewpoint: Listener traces the street from Speaker's viewpoint; then pauses and repeats, from his own viewpoint.
4) Mismatch: his gesture contradicts his speech, and Speaker's earlier gesture.

   SPEAKER: He's pulling out of a spot [*gestures car backing out*]
   LISTENER: He's trying to pull out of the spot? [*gestures car moving forward*]

Listener, then, begins by aligning himself to Speaker's physical viewpoint, but through a series of viewpoint conflicts, he shifts away from this alignment. After the switch, he starts asking for elaboration instead of clarification, suggesting that he is actively building a space corresponding to the picture. Only after this does he disagree with Speaker, presenting the correct framing of the image:

"... maybe what it means is he actually got the spot." His shifts in cognitive viewpoint are indexed by changing visual viewpoint.

Gesture is inherently viewpointed, since it is an embodied action that occurs in space, and any body in space has a richly structured frame of reference. It is not possible to gesture iconically without taking a viewpoint on the gestured scenario; often the viewpoint is that of the gesturer's body in some blend. Because of this, gesture provides insight into interlocutors' cognitive viewpoint that speech alone cannot. In the data examined here, the interaction of speech and gesture reveals the constructed, imagined relationship of participants' bodies to a picture, and thereby also profiles the frames of reference that must have been preserved in the ongoing discourse blend for these relationships to exist.

Recent work on gesture and signed languages has shown that both can be profitably understood in terms of Mental Spaces Theory (Fauconnier and Sweetser 1996; Fauconnier 1997; Fauconnier and Turner 2002), and in particular in terms of Real Space blends (Liddell 2003; Dudis 2004; Parrill and Sweetser 2004). When a communicator makes a "grasping" hand motion in the air, for example, and is understood as depicting some character grasping an object, the construal of the gesture is a blend between the communicator's motion in Real Space (her, or her interlocutor's, understanding of the physical space around her) and the imagined character's motion in the described Event Space. As we shall see, interpreting pictures is actually more complex than the gesture just described, but human ability to create and develop Real Space blends is basic to our ability to make and understand meaningful gestures.

5.2 The experiment

5.2.1 The communicative situation

In the study that this chapter's gesture data are taken from,[1] two subjects sat facing each other in a room (see Figure 5.1). One (Speaker) could see a projector screen, and the other (Listener) could not. Nine different comics panels (including the one under discussion) were presented on the projector screen, one at a time. The Speaker's task was to describe each panel so that the Listener could understand it. The Listener was encouraged to ask questions if anything was unclear. After each panel was described to both subjects' satisfaction, the Speaker pressed a key[2] to move on to the next image. Images were presented in random order.

As seen in Figure 5.2, subjects were mostly facing each other, though Speaker could swivel to look at the projector screen (direction indicated by the arrow).

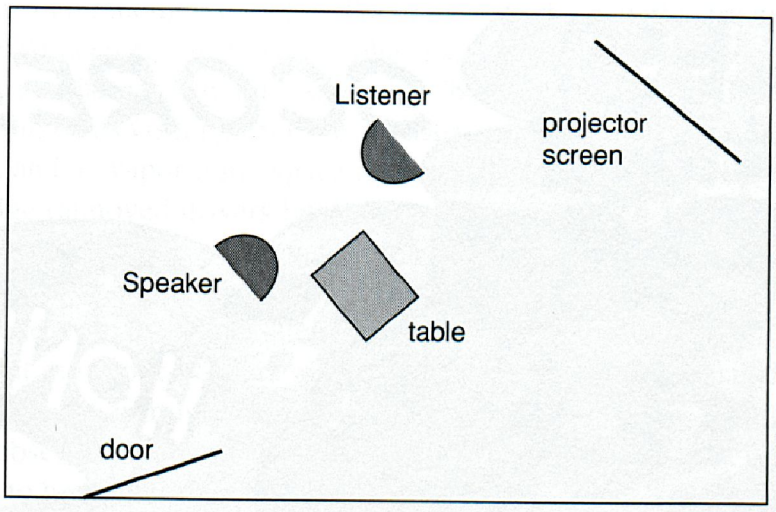

Figure 5.1 The room

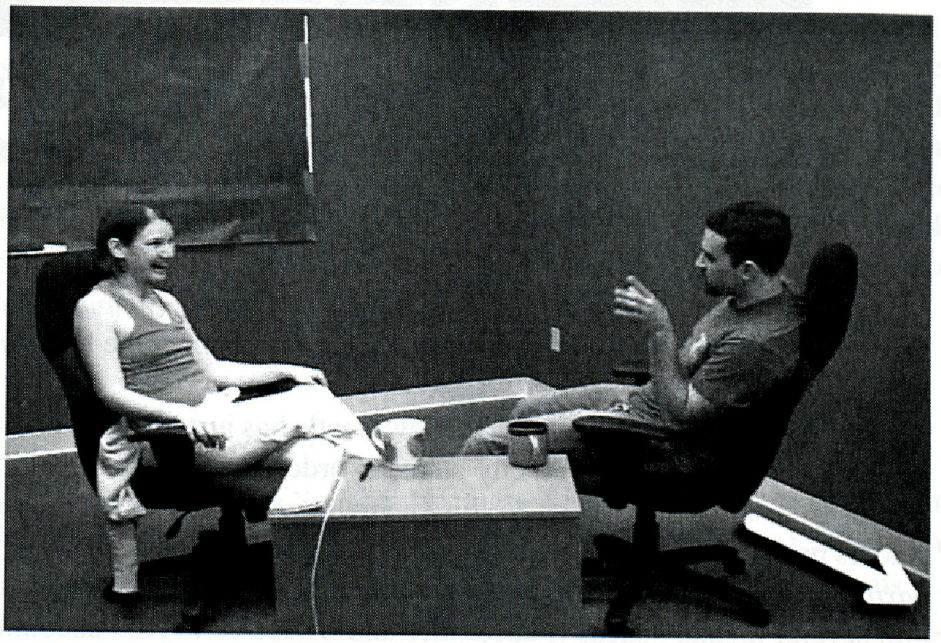

Figure 5.2 The communicative situation

5.2.2 *The stimulus*

Given the task, Speakers could potentially have done most of the talking, with Listeners giving a more limited contribution, perhaps only a few clarification questions. However, this was often not what happened; the process of meaning construction was more collaborative, despite the asymmetry in information availability. This is especially evident in the data from the session analyzed

Figure 5.3 Stimulus image (from Patrick Farley [2003], *Barracuda: The Scotty Zaccharine Story* www.electric sheepcomix.com)

here, where the Speaker became extremely confused about the image shown in Figure 5.3.

In this image, a yellow-green VW New Beetle has just made an illegal U-turn against oncoming traffic on a city street in order to pull into a parking space. To get from the static image to this mentally constructed dynamic scenario, readers need to do some fairly sophisticated cognitive processing. They need to *partition* (Dudis 2004) their understanding of the physical space of the comic into several mental spaces (Fauconnier and Sweetser 1996; Fauconnier 1997; Fauconnier and Turner 2002). Notice that the various words and image pieces are not depicting simultaneous events: the U-turn indicated by the screech and the vapor trail precedes (and partially overlaps with) the *honk* and *hey* events, and *score* presumably is the driver's final mental comment as the parking process is complete. So the image represents a succession of temporally and causally related spaces, which the viewer is invited to construct by relating the frames prompted in the scene. Making the "correct" mappings, to reach the canonical interpretation, requires frame-based knowledge (Fillmore 1985) of real-world situations, image-schematic (Hampe and Grady 2005) and force-dynamic (Talmy 2000) structure, and knowledge of comics conventions such as

vapor trails and thought bubbles. These provide the generic structure. Viewers thus at least need an image space, plus a space for each linguistic item (in this case, each of the single words "screech," "honk," "hey," and "score," which are not interpreted as visual parts of the constructed scene, but as separate auditory stimuli), and a "vapor trail" space indicating motion. Using their frame-based knowledge (annoyed drivers honk or shout when someone makes a U-turn in front of them; seekers of scarce parking spaces may exult when they finally find one), viewers can then integrate these mental spaces to create a dynamic, noisy scene from the static, silent stimulus.

Without this work, readers would draw some rather implausible inferences. For example, they might infer that there are bright yellow letters floating in the air, obscuring parts of the road and a car, or a curved area of tarmac that happens to be lighter than the rest of the street.

Speakers in this study had no trouble inferring that the letters spelling out *Honk* represent a sound, and that the sound was made not by the green car, but by the white car behind it. In addition, almost all Speakers understood from the yellow letters that the green car made a screeching sound, that the "vapor trail" indicated the path that the car followed, and that the co-location of the yellow letters and the path of the vapor trail indicated that the car screeched *as* it traversed this path. Speakers could generally infer from the combination of trail and letters that the motion was fast. Theoretically, they might also infer from the placement of letters *where* along the path the screeching sound occurred, and conclude that when the car gets to where it is drawn, it is no longer screeching; however, none of them mentioned that specifically.

5.2.3 *The Speaker's confusion*

In the clip analyzed here, the Speaker has a great deal more trouble than most subjects. In twenty-four subject pairs, she is the only one who had this much trouble; she seems to have missed the motion trail – she never mentions it. Possibly as a result, she never links the SCREECH with the green Beetle at all.

Her initial understanding of the picture is both very far from the interpretation I proposed, and extremely unsatisfactory to herself. She changes and modifies her interpretation several times during the course of the discourse. Initially, she tries to structure the image according to a video game frame.

(1) *[0:01]* **Speaker**
The biggest thing that stands out is this huge thing that says SCORE!
Oh, maybe it came from a video game.
Looks like it coulda come directly from a video game.

She never mentions a comics frame, but does implicitly evoke one and restructure the scene accordingly.

(2) *[0:55]* **Speaker**
There's honk, screech
and some guy saying
hey
or the, a bubble
that says hey coming out of a car

But as soon as she has done this, she mistakenly evokes an accident frame, and only a little later does she realize that the image is of a street scene but not of an accident.

(3) *[1:06]* **Speaker**
y- there's no visual . . . thing, of
honk . . . or of a, of a accident . . .
but like, the screech and the word honk make you think that there may have been an accident.

(4) *[1:16]* **Speaker**
Oh, I g- okay, I'm understanding what's happening here.
Nervous laughter
So, it's a, it's a street scene . . .

She then gets confused about the car's direction of motion, concluding that it is pulling *out* of the parking space.

(5) *[3:19]* **Speaker**
RH hold diagonally in front of face
[One guy . . .][is actually
[*LH B, palm facing R, fingers touching R arm mid-forearm*]
[*LH moves down away from R arm*[
[Either[
[*LH moves down slightly more*
He must be pulling out of a spot.
LH returns to R forearm, then pulls downward away from RH
LH returns to R forearm
He's pulling out of a spot. And it is a . . .
LH pulls downward away from RH

Speaker's right forearm represents the upper edge of the road; it is held at an angle that mirrors the road's angle from a viewer's perspective, and its length is mapped to the visible segment of road. Her left hand represents the green

Beetle; it maps iconically to the car's position and orientation relative to the road, and then, when she moves it, to the car's inferred motion.

Because of Speaker's continued and changing confusion, this subject pair provides a rather spectacular example of collaborative construction of meaning. The comics artist provides cues to meaning construction that Speaker does not pick up on; but as we shall see, Listener, who cannot see those cues, nevertheless picks up on them from Speaker's description, and from the mental model he constructs. His comments feed back into Speaker's meaning construction, thereby allowing them both to reach an understanding of the image that is close to the canonical interpretation I have proposed.

5.3 Speaker and Listener viewpoints

So how does Listener break away from the interpretation that Speaker has constructed? How does he move from nodding and asking for clarification to proposing an alternative solution? This shift in interaction seems to be tied to his shifting cognitive viewpoint, which goes through four sequential phases, indexed by his viewpointed speech and gesture.

In the first phase, he takes Speaker's viewpoint, by moving his hand to his right (her left), while saying "left" (much as a dance teacher might when demonstrating for students). In the second, he misunderstands her Viewpoint Space – she is describing the picture as a visual object seen from her physical viewpoint, while he thinks she is describing the depicted scene from the imagined viewpoint of someone within that scene. In Mental Spaces terms, he assumes that her Viewpoint Space is the nested space of a driver-participant in the depicted traffic scene, when her Viewpoint Space is actually her own Base Space – that of a person looking at a picture. So when she says "the right hand side of the street," meaning the part of the street further to the right of the depicted image as she sees it, he objects that "whatever side of the street you're on, that's gonna be the right side."

In the third, he stops taking Speaker's viewpoint, and changes the direction of his gestures so that they would be coherent if *he* were looking at the picture (though the only picture he's "looking at" is imagined). In Mental Spaces terms, his speech and gesture indicate that his Viewpoint Space is that of his (reconstructed) understanding of the depicted scene, a blend in which some of the physical space to the front of his actual body is construed as being part of the imagined scene, such that the concrete gestures centered on his own body correspond to structure in his imagined space. Such construals of actual space are known as Real Space Blends (Real Space is a speaker's conceptual model of the physical space around her [Liddell 2003; Dudis 2004]; recall that in Mental Spaces Theory, speakers/thinkers never have objective knowledge, but only their own perceptual and cognitive models, at their disposal).

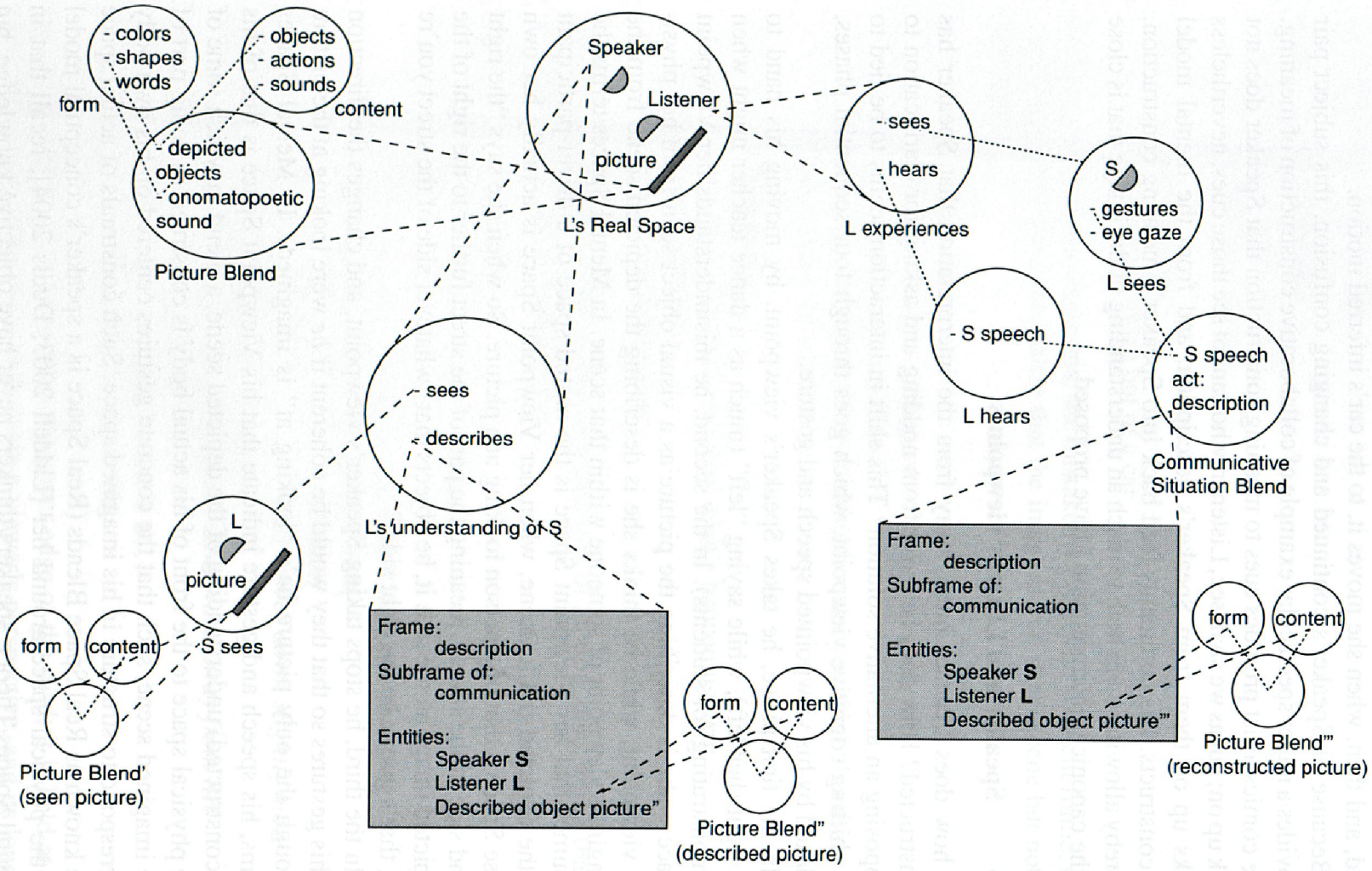

Figure 5.4 Listener's accessible space network

In the fourth phase, he shows a gesture mismatch with Speaker's gesture, repeating her words but with a gesture that moves in the opposite direction from hers. Only after this does he disagree explicitly with Speaker, presenting the correct framing of the image.

Listener's shifts in cognitive viewpoint are indexed by changing visual viewpoint, which are detailed in the sections below. To properly understand them, however, we must spend a little time on his constructed understanding of the communicative situation, before we move on to his understanding of the depicted scene.

This is more complex than it may seem at first, as can be seen by Figure 5.4, a minimal representation of Mental Spaces accessible to Listener and informing his understanding of the discourse. Listener needs to keep track of both his Real Space (*L's Real* Space in Figure 5.4, including Speaker and a screen) and his understanding of Speaker's experience and actions (*L's understanding of S* in Figure 5.4 – what is he interpreting her as describing, and reconstructing her as seeing?). He must also be tracking what he experiences (*L experiences*), which includes the nested spaces of what he sees and what he hears, some parts of which (saliently Speaker's speech and gestures) are blended into a *Communicative Situation Blend*.

In Listener's Real Space is also the visually inaccessible but saliently present projector screen behind him, which shows a comics image. He knows, then, that there is a picture on the screen with formal and semantic elements (expanded into his *Picture Blend* in Figure 5.4). The dotted lines between the spaces indicate that the *Picture Blend* Space is an elaboration of the Real Space, a partitioned part of it, rather than a nested space. The dots connect to the Blended Space rather than the Form and Content Spaces because the blend is primary. This blend is generic – Listener does not see the picture and cannot form a full understanding of the blend; he only knows it is there and possibly uses some of the generic blending conventions he has for comics images once he knows that the image is part of a comic.

In addition, he must be aware of what Speaker sees – especially that she has (presumably veridical) experience of that picture. This is not the same blend as the picture itself; it is a part of Speaker's experience as Listener constructs it, shown in Figure 5.4 as *Picture Blend' (seen picture)*. Her task is to describe this blend, thus structuring her experience with a description frame. The act of describing creates another blend – *Picture Blend'' (described picture)*. Note that this is all Listener's understanding of Speaker's experience and action – her own mental space network would be entirely separate, in *her* head rather than his, and presumably would contain analogs of many of these spaces.

Because the task is to listen to her description, Listener's *Communicative Situation Blend* must be structured by a description frame. This description lets him create an elaborated understanding of the stimulus, shown

in Figure 5.4 as the nested *Picture Blend''' (reconstructed picture)*. This is a different instantiation of the description frame from the one which structures his understanding of *her* experience, because this one structures *his* Communicative Situation Blend. They do share identity mappings, however, since the situation is intersubjective, and the description she gives is the same one he hears.

In these terms, Listener's task is to bring *Picture Blend'''* into alignment with the actual picture, via his experience of *Picture Blend''*, and the assumption that image-schematic structure is preserved from the picture to Speaker's seen picture, to *Picture Blend'*. (In order to keep the diagram possibly comprehensible, I have not drawn in any of these identity mappings.) As we shall see, in order to do so, he starts off with speech and gesture that suggest his cognitive viewpoint is *S Sees*, and shifts his viewpoint to *Picture Blend''' (reconstructed picture)* over the course of the interaction. This shift, this pulling away from alignment with Speaker's viewpoint, is what lets him make the cognitive leap that she does not make.

5.3.1 Listener takes Speaker's viewpoint

Later in the exchange, Listener makes the following comments:

(6) *[1:50]* **Listener**
so the larger
B hands, palms facing each other, go to top right corner
the larger angle of the road is
Both-hand beat
on – going off – to the . . .
B hands move further up and right
left-hand side?
B hands stay up and right

(7) *[1:56]* **Listener**
And then, kinda shrinks down as it goes to the right-hand side?
B hands trace a line down and left, ending in bottom left corner

These examples show an apparent mismatch between Listener's speech and gesture. He says "right" while gesturing to his left, and vice versa. Listener is not, however, confused. He is facing Speaker, and when he gestures to his right, he is gesturing to her left; he does this, and then hesitates before saying "left," which indicates some difficulty rather than a mistake. It is worth noting that he is mirroring Speaker's earlier gestures at this point; the hand shape and movement of his gesture reflect her previous one, as shown in Figure 5.5.

I have called this "taking Speaker's viewpoint." Obviously, however, we do not have direct access to other people's mental spaces, and Listener's Viewpoint

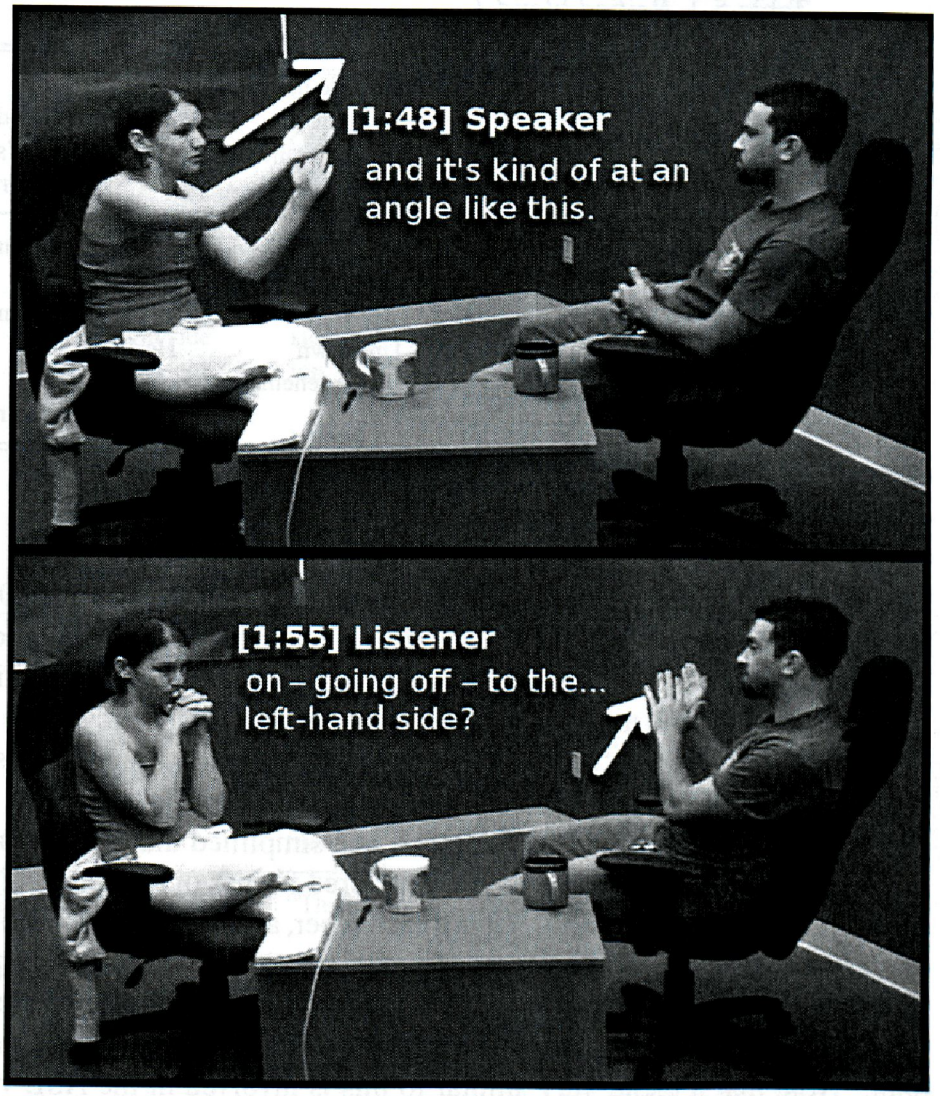

Figure 5.5 Mirroring gestures

Space is not actually that of Speaker's visual experience. It is Listener's partial, minimal reconstruction of that visual experience. This is the space labeled *S Sees* in Figure 5.4.

In terms of the discourse, what this means is that the spatial frame of reference of *S Sees*, and its image-schematic structure, are preserved and taken as canonical in the blend that Listener is currently running. This blend takes two reified, standard blends as input.

5.3.1.1 Reified blend 1: shifting to an interlocutor's frame of reference

We shift frames of reference all the time, by integrating our Base Space with a space created by some image-schema transformation (Lakoff 1987). One very simple, reified version of this happens in face-to-face conversation; we

Table 5.1 *Reified blend 1*

Input 1: A's Frame of Reference	Input 2: B's Frame of Reference	Generic image structure of: Bodies Frames of Reference		Blend: A shifts to B's Frame of Reference
A's left side	B's left side	Right–left image schema		"Your left"
A's right side	B's right side			"Your right"
A's head	B's head	Up–down image schema	Head	Up
A's feet	B's feet		Feet	Down

blend our Real Space with a 180-degree mental rotation (Kosslyn 1980) of Real Space, to create the blend that is our understanding of our interlocutor's Real Space. (*S Sees* in Figure 5.4 is just such a blend.) The process of rotation often takes some work, and may be the reason for Listener's disfluency in the segment at 1:50, before he says "left." But in the blend, his right-directed gesture is a representation of her left-directed gesture.

Reified blend 1, laid out in Table 5.1, is a simplified diagram of this blend, noting only four points of our richly structured frames of reference.[3] It presupposes that A and B are upright, facing each other, as interlocutors prototypically are; however, the cognitive process of mental rotation is versatile and potentially preserves all relevant structure. Since that cannot be completely represented in text form, I have restricted the diagram to elements relevant to the current data.[4] Note that a blend very similar to this is involved in the ASL viewpoint shift mechanism described by Janzen (this volume), where a mental rotation allows a signer to represent a mirror image of a scene without in fact moving his or her body at all.

5.3.1.2 Reified blend 2: imposing a canonical frame of reference

Another common process is of imposing a frame of reference on an object that does not inherently have one (Lakoff 1987), blending our frame of reference with its inherent image schema structure[5]: thus a picture has left–right structure parasitic on the viewer's body (the picture is either blended with the viewer's own left–right structure (as in this case) or with that of an imagined person facing the viewer. Reified blend 2 is laid out in Table 5.2. Like the previous Reified blend, it is a simplication of this process for current purposes.

Table 5.2 *Reified blend 2*

Input 1: Viewer's Frame of Reference	Input 2: Object's spatial orientation	Generic image-schematic structure	Blend: Imposed Frame of Reference
Viewer's left side	side	Right–left image schema	"The/its left"
Viewer's right side	side		"The/its right"
Viewer's head	top	Up–down image schema	"The/its top"
Viewer's feet	bottom		"The/its bottom"

Table 5.3 *Discourse Blend 1*

Input 1: Listener's Frame of Reference	Input 2: L shifts to S's Frame of Reference blend	Input 3: Imposed Frame of Reference blend	Shared image schema structure of: Bodies Frames of Reference	Blend: Listener shifts to Speaker's Frame of Reference as canonical
L's right side	"Your left"	"The left"	Right–left image schema	"The left"
L's left side	"Your right"	"The right"		"The right"

5.3.1.3 Discourse blend 1: "the . . . left-hand side"

Reified blend 1, then, lets Listener construe his right, where he is gesturing, as Speaker's left. By Reified blend 2, Speaker's frame of reference *is* the picture's canonical frame of reference. Therefore, integrating all these spaces into the discourse's current Real Space Blend, Listener's right is *the left*. This is shown in Table 5.3. Figure 5.6 highlights the image-schematic structure being preserved in this blend.

As has been noted elsewhere in this volume, physical viewpoint is a marker of cognitive viewpoint. At this point in the discourse under examination, by preserving Speaker's frame of reference in his discourse blend and taking it as canonical, Listener is subsuming his own physical viewpoint to his understanding of Speaker's physical viewpoint, indicating a close alignment of his cognitive viewpoint to hers. This is natural, given his dependence on her for information about the picture.

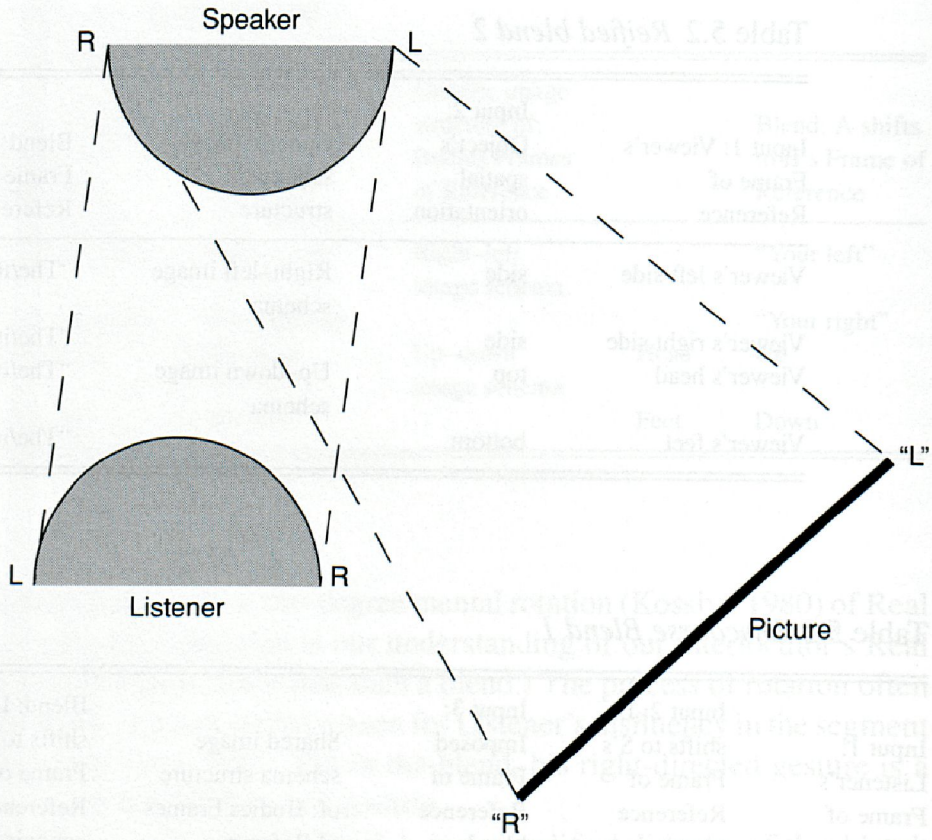

Figure 5.6 Image-schematic structure in Discourse blend 1

In the following sections, that alignment breaks down, until Listener gives up on it completely and begins to speak and gesture about the picture as though he were looking at it himself.

5.3.2 *Listener misunderstands Speaker's viewpoint*

The exchange continues as follows:

(8) *[2:40]* **Speaker**
Um, on the right-hand side of the street
There are . . .
Five cars. One of them is in the middle of the street,
driving.

[2:48] **Listener**
Hold on hold on h'lon
You can't say the right side of the street because whatever side of the street you're on,
that's gonna be the right side, 'cause that's where you're driving.

Here we see two different viewpoints: Speaker's viewpoint is her Base Space – her Real Space, with herself looking at a two-dimensional image projected onto a screen. This allows her to talk about the panel boundary and the objects contained within; it is also a sign that she may not be comprehending the scene yet. The frame of reference that she preserves in her discourse blend is the one she has imposed on the picture.

Listener, however, is confused by her use of the word "right"; his confusion gives us a clue that he understands her as viewing the scene in the depicted image from inside – perhaps he thinks she is taking the viewpoint of someone driving a car along a street. It is only within this space, structured by the driving and street scene frames, that cars have to be on the right side of the street, "'cause that's where you're driving." His constructed understanding of her blend preserves the frame of reference of the inferred driver in the depicted scene.

5.3.2.1 Discourse blend 2: Speaker's construction of the scene

In this analysis, my conception of Real Space differs from Liddell's (Liddell 2003), by which pictures are inherently part of Real Space. In Liddell's analysis, the mental spaces involved in understanding a viewed picture *are* part of Real Space, since visual processing makes depicted spaces accessible as part of our primary experience. Thus, for example, the image of a car in a picture is indeed an object in the physical Real Space of the viewer.

However, this analysis requires significant expansion to allow for the complex blending processes involved in understanding some pictures – no coherent understanding of this particular picture (or of many other comics images) could be reached by construing it just as a static object present in Real Space. Nor, therefore, does treating a picture simply as a Real Space object account for the misunderstanding discussed here – Speaker is construing the picture as a two-dimensional object viewed from outside, while Listener believes she is construing the picture as a three-dimensional space (and perhaps viewing it from inside the constructed scene). Since the misunderstanding arises from the subjects dealing with a picture in versatile ways, which suggest they have access to several spaces, more unpacking is necessary; pictures are not only objects in Real Space. The relationship between Real Space and picture spaces is perhaps better understood as a primary blend network, and this is how I analyze them.

Within this network, then, Speaker's blend at [2:40] preserves the frame of reference of her Real Space, as it has been imposed on the picture. This blend is diagrammed in Table 5.4.

5.3.2.2 Discourse blend 3: Listener's misunderstanding of Speaker's blend

If Listener had understood Speaker perfectly, then his Viewpoint Space, the one whose frame of reference was preserved in *his* blend, could simply be *L's*

Table 5.4 *Discourse blend 2*

Input 1: Picture Space blend		Input 2: Speaker's Real Space	Input 3: Imposed Frame of Reference blend	Generic structure	Blend:
Form	**Constructed meaning**				
		Room			Room
		Speaker			Speaker
					Speaker's frame of reference
		Listener			Listener
		Screen			Screen
Picture shape: square (has horizontal and vertical extent)		Speaker's left	"The left"	Right–left image schema	"The left"
		Speaker's right	"The right"		**"The right side of the street"**
Grey wedge	Street	Depicted scene		Pictures (subtype: comics)	The street
Green ovoid	Car (Beetle)				The Beetle
Other shapes	Other cars, pedestrians, buildings, etc.				Other cars, pedestrians, buildings, etc.
Words	Sound				Sound
Bubbles containing words	Speech				Speech

understanding of S in Figure 5.4. However, *L's understanding of S* is not a salient part of the blend he is constructing at [2:48]; rather, he is elaborating his understanding of *Picture blend'''*, already structured by the street scene frame, by blending in a space structured by the driving frame. As I said, this kind of blended picture interpretation, as a three-dimensional scene with participant viewpoints, has not so far been elaborated by Speaker.

The frame of reference that Listener's blend preserves is that of the Driver, an entity in the Driving Space. It is the Driver's viewpoint that Listener takes when he says that "whatever side of the street you're on, that's gonna be the right side, 'cause *that's where you're driving*."

Table 5.5 *Discourse blend 3*

Input 1: Picture Space blend		Input 2: Driving	Generic Structure	Blend:
Form	**Constructed Meaning**			
Picture shape: square				
Entities	**Entities**	**Entities**		**Entities**
Grey wedge	Street	Street		The street
	Driver (implied)	Driver	Driver: instance of *human*	Driver
				Driver's frame of reference
Green ovoid	Car (Beetle)	Driver's car	cars	Driver's Beetle
Other shapes	Cars	Cars		Cars
	Pedestrians	Pedestrians	Pedestrians	Pedestrians
	Buildings	(optional: buildings)	Buildings	Buildings
Words	Sound	Sound	Sound	Sound
Bubbles containing words	Speech	(optional: speech)	Speech	Speech
Relations	**Relations**	**Relations**	**Relations**	
		Driver sits in car to drive, facing forward		Driver sits in car to drive, facing forward
		Driver's Frame of Reference imposed on road		Driver's Frame of Reference imposed on road
Shapes along edge of grey wedge	Cars along side of street	Cars drive on "the right"	Cars drive on street	Cars driving on **"the right side of the street"**

This suggests that even at this early point in the discourse, Listener is creating a more coherent understanding of the image than Speaker is, as laid out in Table 5.5 as Discourse blend 3.

5.3.3 Switching visual viewpoint

The interchange featured in this section happens just after the one in section 5.3.2; the misunderstanding seen in section 5.3.2 may be cuing a more general realignment of Listener's viewpoint.

Table 5.6 *Listener's Real Space blend*

Input 1: Listener's Real Space	Input 2: Spatial orientation of Picture blend'''	Shared image schema structure of: Frames of Reference	Blend: Listener's Real Space blend of the reconstructed picture
Listener's right side	Vertical edges of picture	Right–left image schema	Vertical edge *is* on L's right side
Listener's left side			Vertical edge *is* on L's left side
Top of L's gesture space	Top of picture	Up–down image schema	Top of picture *is* at top of L's gesture space
Bottom of L's gesture space	Bottom of picture		Bottom of picture *is* at bottom of L's gesture space

(9) *[3:02]* **Listener**
So there's four cars,
RH 5, palm down, sweeps across gesture space, L bottom to R top
there's four cars going up . . .
Going up the, uhh . . .
RH 5, palm down, angles from R bottom to L top

At the beginning of this example, Listener is taking Speaker's viewpoint, as we saw in section 5.3.1. However, he then hesitates and changes his gesture so that it matches *his* left–right orientation, rather than Speaker's.

In terms of the ongoing blend that Listener is constructing, he is no longer making use of the *Shifting frame of reference* blend outlined in section 5.3.1.1, but rather creating a different Real Space Blend, diagrammed in Table 5.6, in which the *Picture Blend'''* Space is integrated with his Real Space in a location in front of him, and *his* frame of reference is imposed on the constructed image.

Creating a Real Space Blend gives a spatial orientation to *Picture Blend'''* – an imaginary, constructed "object." By taking this Real Space Blend as an input to the Imposed Frame of Reference Blend (in a manner analogous to that in section 5.3.1.2), Listener imposes *his* frame of reference on his understanding of the picture. This is laid out in Discourse Blend 4, shown in Table 5.7.

5.3.3.1 Discourse Blend 4: Listener imposes his frame of reference on Real Space Blend of reconstructed picture

Listener's Viewpoint Space, then, has shifted entirely from *S Sees* to his own Real Space, and the reconstructed picture blended with it; even formal

Table 5.7 *Discourse blend 4*

Input 1: Listener's Frame of Reference	Input 2: Spatial orientation of blended "Object"	Generic image-schematic structure	Blend: L's Frame of Reference imposed on Real Space blend
Listener's left side	Vertical edge	Right–left image schema	"The/its left"
Listener's right side	Vertical edge		"The/its right"
Top of L's gesture space	Top	Up–down image schema	"The/its top"
Bottom of L's gesture space	Bottom		"The/its bottom"

elements now take the viewpoint of this blended space. His gesture and language no longer depict *L's understanding of S* (or spaces nested below that space). This suggests that Listener's cognitive viewpoint has become less closely aligned with Speaker's; he is no longer working to understand the image in terms of what Speaker sees, but to imagine his own view of the reconstructed picture.

Before Example 9, Listener has been asking for clarification and information. Once he switches viewpoint, the grammatical form of his questions changes: some are phrased as statements, which Speaker must accept or reject. The next two examples show Listener's last utterance before the [2:40] point, and his first one after the [3:02] point.

(10) *[2:39]* **Listener**
how many cars are in the picture?

(11) *[3:09]* **Listener**
Okay.
They're all moving cars, no parked cars on the side of the street.

This suggests that, having shifted viewpoint to his own reconstructed picture blend, Listener is now talking about that constructed understanding of the image. Since his task is to align this space with that of Speaker's understanding of the picture (and, via Speaker, the picture itself), this makes sense.

5.3.4 Mismatch

At several points through the clip, Speaker has used a right arm hold, elbow down, diagonally across her face, to depict the top *edge* of the road, and has made iconic gestures representing objects *in* the road with her left hand, at points below her right arm. She does so again at [3:31], using her LH to

represent the green Beetle. Her fingers, which start in close proximity to her RH, map to the front of the car (which is close to the curb), and the heel of her hand maps to the back of the car (which is out in the street). Her gesture is unlikely to be ambiguous at this point, as she is elaborating on a pre-established iconic structuring of Real Space.

(12) *[3:31]* **Speaker**
And he's tryna [pull out] into the road.
R arm returns to diagonal hold
[LH B moves downward from R arm]

[3:32] **Listener**
he's tryna pull out into the road?
RH B, palm facing left, fingers forward, hand moves out from body

Speaker's left-hand gesture, co-timed with "pull out," is iconic for a car backing up, since the car could not be moving forwards legally in that orientation to the relevant side of the street, and since fingertips generally map to the "leading edge" of a moving object, most often the canonical front. The structuring of her Real Space blend, and thus the iconicity of the left-hand movement, should be unambiguous.

However, when Listener repeats her words, he makes a small gesture, mismatched to hers – and possibly mismatched to his speech. Assuming that both speakers are mapping the front of the car onto their fingers and the back onto their palms, Speaker's gesture iconically depicts the car backing out,[6] but Listener's gesture depicts the car moving forward. In addition, if he has set up his Real Space Blend of the scene in a way consistent with Speaker's (as might be expected from his earlier, closely aligned viewpoint), then his gesture would map iconically to a car pulling *into* a space at the "top" of the road, not pulling out of a space.

This gesture suggests a viewpoint matching and elaborating that of Discourse Blend 3, preserving the frame of reference of the Beetle's implied Driver. While Speaker's gestures are on a two-dimensional space, suggesting that the picture in her Real Space Blend is still a flat object, Listener's gesture moves away from his body, suggesting a three-dimensional, egocentric representation, with the car at Ego location. His mismatch looks like a classic mismatch (Goldin-Meadow 2003) between gesture and speech, but it is hard to see how he could have made it if he had simply been mirroring Speaker. And indeed, like some of Goldin-Meadow's subjects' gesture–speech mismatches, it presages a new construal that has not yet emerged in speech.

Not long after this, Listener provides the interpretation that they both agree is correct.

(13) *[3:57]* **Speaker**
The car parked behind the bug is so hard to see, could even be a taxicab, it's just a
yellow car, um . . .
It's hard to see because the word honk
Um
Is basically over the top of it.

[4:06] **Listener**
It's interesting that it says SCORE, maybe what it means is he actually got the spot,
as opposed to . . .
looks at Speaker

[4:11] **Speaker**
Ohhh! Yeah!

[4:12] **Listener**
He scored the [spot

[4:13] **Speaker**
[He's like [Score! I got a spot!]
[R arm held vertically; S hand, palm facing face, moves diagonally across face]

In the moments before he suggests this correct interpretation of the picture he cannot see, Listener stops responding to Speaker's speech and stares into space. His focus is clearly not on any space that pertains directly to Speaker; it seems to be on his reconstructed Picture Space Blend.

The breakdown we saw in section 5.3.2, where Speaker and Listener's iterative process of collaborative meaning construction failed due to a viewpoint misunderstanding, might actually be *helping* Listener here. His resulting disalignment of viewpoint from Speaker's, seen in sections 5.3.3 and 5.3.4, allows him to conceptualize the image in a way that she is not doing, and thus to create meaning that she is not constructing.

5.4 Discussion

In section 5.3, putting together speech and gesture data allowed us to trace the stages of Listener's conceptualization, which would only be very partially available from the speech track alone. We saw that Listener's visual viewpoint indexed a cognitive viewpoint that started closely aligned with Speaker's, then shifted; his attempts to build meaning were apparent in the interaction of his speech and gesture, as was a sense of the changing frame of reference he was working with as his cognitive viewpoint shifted.

An analysis of these data that focused merely on the speech would prove outright misleading; in only one of these crucial sections of the interaction could we have inferred much about viewpoint from speech alone (5.3.2: Listener misunderstands Speaker's viewpoint). Gesture data, therefore, are crucial to our understanding of Listener's viewpoint shifts, and therefore to our understanding of why his speech patterns shifted at the [3:02] mark, from confirmatory questions to declarative statements, leading to the construal that resolved Speaker's confusion.

Why might gesture prove crucial to the study of cognitive viewpoint? Certainly, one answer is that gesture, as a less conscious communicative track than speech, is crucial to the study of language in general; also, being visuospatial, it has much richer scope for iconicity than the speech stream (Taub 2001). But in addition to these general qualities – as mentioned at the beginning of this chapter, gesture is an embodied action, and thus inherently viewpointed. The iconic and deictic information coded in gesture can only *be* iconic or deictic with regard to some frame of reference (as well as some Focus Space, disambiguated by speech [Parrill and Sweetser 2004]). By imposing a frame of reference on the ongoing blend, then, the Viewpoint Space constrains possible gestures to those that can meaningfully occur within that frame of reference. Cognitive viewpoint imposes the frame of reference of a metaphorical "space" rather than a literal one; language is inherently viewpointed too, but rarely as unambiguously as gesture.

Gesture takes place in a physical space, with observable physical viewpoint – we can see what the subject is looking at, or in what direction he is placing his arm to iconically represent an object. Language is often hugely ambiguous between reference in different spaces; it is actually useful for interlocutors to consider that they are talking about the "same" thing in referring to S's imagined picture and L's imagined picture, or to the physical picture on the screen and the imagined contents of the picture. But for the analyst, trying to unpack these spaces that language conflates, gestural viewpoint provides invaluable cues to distinguish parts of the network from each other.

And by imposing a frame of reference on the blend, the Viewpoint Space constrains cognition; it is the space *in whose terms* cognizers conceptualize the rest. Viewpoint, then, is indexed physically by gesture and provides a constraint on the meaning that can be constructed; in turn, gesture provides a channel through which viewpoint can be analyzed. This tight link between speech, gesture, and cognition is unsurprising; as Sweetser (1998) notes, language is a manifestation of a cognitive system "concerned with interaction and the situation of the person." With regard to aspects of cognition such as viewpoint, which depend crucially on our embodied understanding of space, the surprise would be if gesture analyses were *not* crucial to our understanding of communication.

Yet another general lesson reinforced by this example is that cases where communication and reasoning processes break down are often more revealing to the analyst than cases where all goes without a hitch. Smooth interaction is generally transparent to participants – and often transparently understandable to the analyst as well. On the other hand, breakdown and renegotiation force presupposed and transparent mechanisms to the fore, and profile the cognitive underpinnings of framing and viewpoint structure.

References

Coulson, Seana. 2001. *Semantic Leaps: Frame-shifting and Conceptual Blending in Meaning Construction*. Cambridge University Press.

Dudis, Paul G. 2004. Body partitioning and real-space blends. *Cognitive Linguistics* 15:2, 223–38.

Fauconnier, Gilles. 1997. *Mappings in Thought and Language*. Cambridge University Press.

Fauconnier, Gilles and Eve Sweetser. 1996. *Spaces, Worlds, and Grammar*. University of Chicago Press.

Fauconnier, Gilles and Mark Turner. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.

Fillmore, Charles J. 1985. Frames and the semantics of understanding. *Quaderni di semantica* 6:2, 222–54.

Goldin-Meadow, Susan. 2003. *Hearing Gesture: How Our Hands Help Us Think*. Cambridge MA: Harvard University Press.

Hampe, Beate and Joseph E. Grady (eds.). 2005. *From Perception to Meaning: Image Schemas in Cognitive Linguistics*. Berlin/New York: Mouton de Gruyter.

Kosslyn, Stephen M. 1980. *Image and Mind*. Cambridge MA: Harvard University Press.

Lakoff, George. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press.

Liddell, Scott K. 2003. *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge University Press.

Parrill, Fey and Eve Sweetser. 2004. What we mean by meaning: conceptual integration in gesture analysis and transcription. *Gesture* 4:2, 197–219.

Sweetser, Eve. 1998. Regular metaphoricity in gesture: bodily-based models of speech interaction. In *Actes du 16e Congrès International des Linguistes* (CD-ROM). Elsevier.

Talmy, Leonard. 2000. *Toward a Cognitive Semantics*. Cambridge MA: MIT Press.

Taub, Sarah F. 2001. *Language from the Body: Iconicity and Metaphor in American Sign Language*. Cambridge University Press.