Cooperative Meaning Construction in an Interactional Task

Zachary Weinberg Shweta Narayan Rafael Núñez

September 2, 2007

Abstract

We present data from a gesture experiment in which one person describes comics panels to another. The task of the speaker, who can see the image, is to describe it so that the listener (who cannot see it) can understand it. In this case study, feedback from the listener is critical to both participants' understanding of the picture. Over the course of several minutes, they incrementally and interactively construct an understanding of the picture that takes into account all available data.

Introduction

Consider the Comics panel shown in figure 1. For the moment, imagine everything visible to be physically present in the depicted scene. This is no ordinary photograph of a city street. Enormous words are floating in the air, obscuring street, cars, and buildings. Other words, surrounded by white areas, are overlaid on the scene, as if we were looking through a window with stickers on it. There is a gray stripe on the road, tracing a smooth curve that ends at the green car.

This interpretation is not the first to come to mind, even when explicitly raised. Rather, viewers understand the words, "bubbles," and stripes to be conventional Comics markers of sound, speech, and motion. These are drawn, by necessity, on the same page as the picture, but not meant to be seen as part of it.[1] These markers cue the construction of a dynamic scene within viewers' minds. The green car has just made a sharp turn at speed into a parking space; it skidded a bit early in the turn and its tires screeched. In so doing, the green car cut off the white car, angering its driver, who honked and yelled "Hey!" Someone else yelled "Score!" It is likely to be the driver of the green car, since parking spaces in a large city are hard to find.[2]

All the inferences required to reach this "canonical" interpretation of the image are immediate and obvious. Someone familiar with Comics will find it difficult to interpret the panel any other way, even though, as McCloud (1999) notes, no one has said a word. The question for cognitive

[1] Comics artists do sometimes make speech bubbles or other such markers into physical objects within the scene, but this is a rare act, analogous to "breaking the fourth wall" in theater. It suspends the conceit that the comic is a world unto itself.

[2] The scene in figure 1 takes place in downtown San Francisco, where parking spaces may be worth more than the cars parked in them.

Figure 1: The Comics panel in this case study (Farley, 2003). Used with permission.

science is, how do readers make these inferences? How can readers understand so easily that the screech was a sound made *by* the green bug, *before* it was at its depicted location, *while* it was making a U-turn? How do they know that the honk came from the white car, not the green one, and that this happened after the green car's tires screeched? More complex yet: how do readers infer that the green car's driver has angered the white car's driver by stealing a parking space that the white car was waiting for?

We can gain insight into the mental processes involved by looking at situations where the inferences are neither immediate nor facile. In this case study, two experimental participants collaboratively work out the canonical interpretation of figure 1 over the course of several minutes. Only one participant (referred to throughout as the *viewer*) could see the picture. She initially misinterpreted it, probably because she did not notice some of the cues provided by the artist. The other participant (the *listener*) could not see the picture at all, but inferred some of those cues from the viewer's description and his own mental model. His comments fed back into the viewer's meaning construction, allowing them both to reach an interpretation close to the canonical one.

Theoretical Background

Our analysis follows the principles of Parrill and Sweetser (2004). We will be discussing participants' background knowledge in terms of frame semantics (Fillmore, 1985) and their online reasoning in terms of conceptual integration theory (Fauconnier and Turner, 2002).

Frame semantics describes background knowledge as structured into *frames*, each of which encapsulates knowledge about some domain of experience. Perceptions—words, gestures, images—can *evoke* one of these frames, bringing to mind all of the relevant information. Thinking about one frame can, in turn, evoke other frames with related information.

Conceptual integration theory describes online reasoning as occurring within and relative to *mental spaces*. A space is a context for some collection of facts in working memory, referred to as *elements* of the space. When people are reasoning relative to a particular space, they are said to have that space for their *viewpoint*.

People construct mental spaces on the fly, as needed. They can hold several spaces in working memory at once, and can construct new *blended* spaces from two or more *input* spaces. Blending is not a set operation: what each input provides to the blend depends on what the blend is for, and on the details of the relationship between the input spaces. In the process of blending, *mappings* are established between elements of the input spaces. Mapping is a generic term; it means only that there is some relationship between those elements that is currently relevant. Mappings usually produce elements of the blend. Semantic frames are not mental spaces, but the background knowledge and structure contained in a frame can be used to construct a space for online reasoning. From the perspective of conceptual integration theory, when viewers use the **Comics** frame to identify elements of an image, they do this by blending a mental space containing elements from the image with a mental space containing elements from the frame.

In the scenario we are presenting, the participants make continuous use of five mental spaces related to the content of the image. The *picture* space is the most basic; its elements describe the image *as a picture*, a visible two-dimensional object with a complex appearance. In this space there is no motion or sound. Viewer-participants tend to refer directly to this space when they are having trouble interpreting the image. The *pictorial* space is a mental space with elements from a semantic frame (e.g. **Comic**). When blended with the picture space, the chosen frame imposes its structure on the cues in the picture space, allowing them to be interpreted. Details of the image in the picture space and frame-based information in the pictorial space are mapped onto aspects of an *event* in the resulting *depicted world* space. Unlike the picture space, the depicted world space can contain motion, sound, and anything else one might find in a real event.

The depicted world space is itself the input to a second blend. The other input is the *scenario* space, with elements from another semantic frame (e.g. **Street Scene**). This frame imposes its structure on the event taking place in the depicted world, providing context and motivation for the action taking place. The result of that blend is the ultimate *interpretation* of the image.

For example, when the picture space constructed from figure 1 is structured by the **Comics** frame, which has elements such as motion lines and speech bubbles, the floating words and lines in the picture space can be interpreted as event cues. In the depicted-world space, a car is moving, and screeching, and someone is saying "Hey!" The **Street Scene** frame, which structures the scenario space, allows the cognizer to understand *why* this is happening.

While most of the discussion happens within the context of these five spaces, a sixth occasionally becomes relevant: the *discourse* space, whose elements record the conversation between the participants. It comes to the foreground when there are communication difficulties that require participants to talk about their own conversation in order to clarify matters.

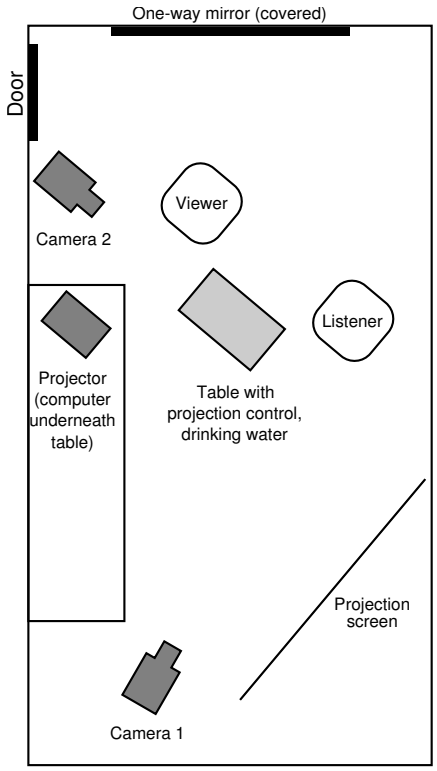

Figure 2: Layout of experiment

Figure 3: View from camera 1

Case Study

We present data from the pilot run of an experiment in language and gesture. Participants were paired and seated in a room organized as shown in figure 2. The participant on the left is the *viewer* referred to above. She could see the projection screen; the participant on the right (the *listener*) couldn't. The projection screen displayed a sequence of nine still images, one of which was the picture shown in figure 1. All of the images were taken from various online comics. The viewer controlled how long each image was displayed. She was instructed to describe each image in detail to the listener, taking as long as necessary, and to advance to the next image only when the listener agreed that he understood the picture. The listener was instructed to pay attention to the descriptions, ask questions, and indicate when he understood the picture. He was warned that he would have to recognize the images in a "quiz" after the experiment.

Both participants were videotaped continuously throughout the experiment. Camera 1 was the primary source of data; camera 2 was a back-up in case the view from camera 1 proved to be inadequate (for instance, if the listener had turned his swivel chair away from camera 1). Figure 3 is a still frame taken by camera 1 during the discussion of the image in figure 1.

In a setup like this, a "classical" position on meaning and discourse would predict the viewer to understand the images easily, and do almost all the work of describing them. The listener's role would be passive, perhaps injecting the occasional question or request for more detail. If the viewer has any serious difficulty understanding the picture, she might pause for some time before

***V:** viewer. **L:** listener.*

V: 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.
both hands rise to in front of face, move outward a bit, return to resting at sides

L: *nods*

V: Um. The word Score is in black, and it's outlined by white, er, [it's white behind it]
2 hands flat, one behind the other

...

V: Score. Just score, [with an exclamation point.

L: [Is someone saying score?

V: *shakes head* There's nobody saying it, it's just [in the top of the screen], and it says Score[] um there's no bo dy say ing it.
RH rises to forehead level, arm extended then retreats to near forehead and holds
R forearm on head with thumb on forehead

L: Is there anyone in the picture?

V: There, okay, so in the picture, []that was just kinda the highlighting thing, there's other words. There's [Honk], [Screech], and some guy saying [Hey], or [a bubble that says Hey] coming out of a car.
R arm slips behind head and then down to ready position with hand next to shoulder
4 successive index-finger points to different areas of gesture space

Transcript 1: First reactions to the picture.

V: ... hard to see because the word honk is basically on top of it

L: It's interesting that it says score. Maybe what it means is he actually got the spot, as opposed to—

V: Oh yeah!

L: He scored [the spot

V: [He's like [Score], I got a spot!
R forearm vertical, RH in fist
the guy behind him's like, [fucking dick!

L: [You asshole.

V: *laughs* He's like, hey, [I wanted that spot!

L: [I wanted that spot.

V: Yeah, that's what's happening there.

Transcript 2: Three minutes later...

beginning to describe it.[3] She would *not* be expected to start her description using a perspective that can describe the picture only poorly. For the most part, the behavior of participants in this experiment is consistent with those expectations. However, in this case study we will present a pair of participants who do not conform.

Incremental Refinement of Meaning

As you can see from transcripts 1 and 2, the viewer's initial interpretation of the image in figure 1 bears little resemblance to the artist's intended reading. Over the next three minutes of conversation, she gets closer to the intended reading, but she persistently fails to understand how the cars are meant to be moving. The listener, using only the ambiguous and partially inaccurate information she provides him, manages to figure out the significance of SCORE, deduce the intended action from that, and suggest it to her, provoking an "aha" experience.

In transcript 1 the viewer is taking one of the picture's most prominent features, the large spiky

[3]This does happen in our data—another picture in the study caused almost every viewer-participant to hesitate for a while, then burst out laughing.

Frame Name		Video Game
Entities	Real world	Player
		Video Game system (computer, console, etc)
	Game	Score
		Game scenario
	Game Scenario	Character
		Action
		Scene
		Goal
Scenario	**General**	**Player** has identity mapping to **Character**
		Player controls **Character** in **Scene**
	T1	**Character** performs **Action**
	T2	**Action** affects **Scene**
		Action may or may not achieve **Goal**
	T3	Achieving **Goal** increments **Score**

Frame Name		Comic
Entities	Real world	Reader
		Comics Panel
		Panel picture
		Convention: Motion trails
		Convention: Onomatopoetic words
		Convention: Speech bubbles
		Convention: Narrative box
	Depicted world	Character(s)
		Scene
		Cued: Actions
		Cued: Sounds
		Cued: Speech
	Cued: Narrative	Words
	Cued: Speech	Words
Scenario	**Conventions** map to **Cued** elements	
	Other details of the picture may also map to **Cued** elements (e.g. motion cued by character posture)	
	This frame provides no explanation of *why* the scene is as it is.	

Table 1: Comparison of pictorial frames. **T1**, **T2**, ... indicate a time sequence of events.

speech bubble with the word SCORE in it, and considering it evidence for the image being a screen capture from a video game. The listener is active almost immediately, asking a question which presupposes a more Comics-like interpretation ("Is someone saying SCORE?") Viewer says no, but signals uncertainty: she slows down, reiterates the observation after having moved on to other detail, and leaves her arm on top of her head for an extended period. Later, she implicitly raising the possibility of its being a Comics panel, by using Comics conventions to describe some aspects of the image. Notice how she initially describes the "Hey!" speech bubble: *some guy saying Hey*, going straight to the conventional interpretation of speech bubbles. She then backtracks and redescribes it in more literal terms: *a bubble that says Hey coming out of a car*. This may be because video games generally do not use speech bubbles, and she is not overtly construing the image as a Comics panel at this stage.

The participants have raised two possibilities, **Video Game** and **Comic**, for a semantic frame to interpret the image on the *pictorial* level: explaining its form, not its content. These frames are compared in table 1. A key difference is that the **Video Game** alternative offers elements for the scenario space as well as the pictorial space. It can *explain* the SCORE bubble: it is not part of the scene, but an overlay to praise the player for a particularly good move that increased their score. However, it cannot even identify the speech bubbles and sound effects, because those elements are not normally used in video games. The **Comic** frame *can* identify the speech bubbles, sound effects, and motion trail as conventional entities which cue their counterparts in the depicted world, but it does not provide any explanation for the SCORE bubble—or anything else in the picture.

V: It's the [scene] of someone it looks like they've
2 hands make a horizontal plane
been, uh, []looks like there's been a, uh, uh
leans forward, staring at image; 3sec pause
There's no visual []thing of, of [honk],
each hand held in grasping-invisible-object pose; still leaning forward
both hands make slight offering gesture, return to hold
or of of a, []accident
LH now held flat vertically at shoulder level with fingers spread
L: Uh huh
begins twiddling his thumbs
V: but, like, the [screech] and the word [honk]
LH gesture to the left; RH to the right
make you think that there [may have been]
scale-balancing gesture
an accident

Transcript 3: The accident interpretation

Frame Name		Road Accident
Subframe of		Vehicular Accident, Road Scene
Entities		Trajector (Vehicle 1)
		Driver (of Trajector)
		Landmark (Vehicle 2 / pedestrian / object)
		Location
		Road
		Optional: other vehicles / objects
		Optional: witnesses
Scenario	**General**	**Trajector** and **Landmark** exist in **Location**
		Driver is located within and in control of **Trajector**
		Witnesses may also exist in **Location**
	T1	**Driver** loses control of **Trajector**
		Optional: **Trajector** may screech or make other loud noises
	T2	**Trajector** runs into **Landmark**
		Optional: Other vehicles may honk or screech as a result
	T3	Optional: **Driver** and other people may be injured

Table 2: The **Road Accident** frame

At this point in the discourse, the participants do not overtly decide which of these frames is correct. However, the video game possibility is never mentioned again, and both participants go on to discuss the depicted world using the identifications that the **Comic** frame would make. For instance, SCREECH and HONK are henceforth treated as sound effects, with the participants talking about cars that *are* screeching or honking. Thus, we may conclude that the comics interpretation has been chosen. However, this means there is no explanation for the SCORE; it is identified as a speech bubble but there is no indication of who's saying it or why it's there. SCORE is not mentioned again until the listener brings it up in transcript 2.

The viewer now moves on to propose that the scene depicts an auto accident. In frame terms, she is proposing a **Road Accident** frame as the basis of the scenario space. Transcript 3 shows the discourse, and table 2 the frame. Note that the viewer is basing this interpretation entirely on the sound effects: *The SCREECH and the word HONK make you think that there may have been an accident*. She signals lots of uncertainty in her interpretation, by posture (leaning forward and staring at the picture), disfluencies, and lengthy pauses in her speech. Note also that unlike the previous passage, the listener doesn't ask any questions; in fact, he starts twiddling his thumbs at one point, suggesting that he isn't paying a lot of attention to this explanation. This may be

V: Oh I [okay], I'm understanding what's happening here.
interlaces fingers, arms return to lap
V: *laughs* So it's a it's a street scene. There's a street. And, uh, there's cars parked on both sides of the street. Um
L: Is it just [one portion of the street]
hands make 2 parallel vertical planes
or is there an [intersection]
hands make a T-shape, still vertical planes
V: It's just one portion of the street.

Transcript 4: Street scene

V: Okay, so there's five cars. Three of 'em are parked. [Three of them], there's [two at the top] and those two are parked stationary.
R arm returns to horizontal at eye level.
LH pinch below R hand.
[One guy is] actually, either, he must be pulling out of a spot.
LH moves to middle of R forearm, hand flat vertical
He's pulling out of a spot, and it is a yellow, um, Bug.
LH moves up and down a few times
L: He's driving a yellow Bug.
V: He's driving a bran, like a new one [...
L: [And he's trying to pull out into the road.
V: And he's trying to pull out into the road.
Repeat above gesture, LH at middle of R forearm and then pulled away (down).
L: And someone behind him is going SCREECH.
V: Someone behind him is going [HONK] and
R fist slamming car horn gesture
there's, yeah, there's a big [SCREECH]
index finger in short arc, mid-space
So the car—
L: That's honking and screeching
V: That's honking and screeching is white...

Transcript 5: Pulling out of a parking space?

Frame Name		Street Scene
Subframe of		Urban Scene, Road Scene
Entities		Location:Road
		Location: Buildings
		Location: Sidewalk
		Optional: vehicles
		Optional: pedestrians
		Optional: bicyclists, etc
		Optional: traffic lights
Relations	**General**	Moving **cars** exist on the **Road** (driving on the right side)
		Parked **cars** exist on the side of the **Road**
		Pedestrians exist on **Sidewalk**

Table 3: **Street Scene** frame

a response to the viewer's uncertainty, perhaps a signal that he's heard enough about the words already and would like her to move on to other elements of the picture.

The auto accident frame tells us that an accident happens when a moving car collides with something else, either another car or some stationary object. We also know that the sound effects presented are those for a car horn and car tires slipping on pavement; these noises are both likely to occur in an accident or near-accident situation. Thus, all the information so far available to both participants is consistent with the accident interpretation. However, the accident frame also tells us

that the cars involved are likely to be damaged. Looking back at figure 1, there is no indication that any of the cars has been damaged, nor in fact any indication that there has been a collision. There are plenty of other situations where cars might generate SCREECH and HONK noises. Furthermore, we cannot explain the SCORE, which is an implausible thing for someone to shout during or after a car accident. These absences may motivate the viewer's uncertainty.

The participants move on, again, without definitely accepting or discarding the frame they have discussed. Transcript 4 follows immediately after transcript 3; notice the viewer's return to fluency, and the listener reengaging the discourse. At this point the viewer begins to describe elements of the image other than the words. She abandons the attempt to describe the event, instead saying that the image shows a street scene and laying out the particular elements that are present. The **Street Scene** frame is outlined in table 3; it partially fills in the scenario space by setting expectations for what one might find in the scene and what sorts of events typically occur. It does not, at this point, supersede **Road Accident** in providing an explanation for events.

After extensive description of the scene in a static fashion, the viewer elaborates on the **Street Scene** scenario by describing another possibility for the event taking place there. Her proposal (shown in transcript 5) is now that the green car[4] is pulling *out* of the parking space. This is quite close to the intended reading, except for the direction of motion. The blended interpretation accounts for most of the image as it has been described. The green car is leaving the parking space and entering the roadway. The SCREECH and HONK noises both come from the white car; perhaps it has had to stop abruptly to avoid a collision with the green car. Note that the listener has deduced this last from the mere existence of SCREECH and HONK as depicted sound effects, plus the presence of a car that (by hypothesis) is pulling out into the road, and a car on that side of the road that is neither parked nor pulling out (the fifth from her count).

This scenario cannot explain the SCORE any more than the accident scenario could. Also, in the picture, the SCREECH is in the wrong place to be coming from the white car. Of course, the listener does not know this when he suggests that *someone behind [the green car] is going SCREECH*, and the viewer does not overtly contradict his interpretation. She does restate it in a way that decouples the SCREECH from the white car: *someone behind him is going HONK, and there is a big SCREECH*. He persists in his interpretation; it is illogical for the SCREECH to be coming from anywhere but the white car in the pulling-out scenario, so this is unsurprising. We cannot say to what extent her restatement affected his reasoning later.

At this point (the end of transcript 5) the viewer talks for about thirty seconds about the shapes of the white car and the hard-to-see yellow car behind the HONK sound effect. The listener stops asking questions, and toward the end of this period, also stops giving indications of engagement (nods, "uh huh" noises). When he speaks again, it is to suggest the alternative "scored the spot" interpretation shown in transcript 2. Clearly, during this time he has deduced this near-canonical interpretation from the information he has available. To reiterate, he knows that this is a Comics panel depicting a street scene; there is a car maneuvering relative to a parking space, and another car honking; there is a SCORE speech bubble that is unexplained. He may also take into account the uncertain source of the SCREECH sound effect, and the viewer's uncertainty when she says the car "must be" pulling out into the road. In addition to hesitating, her gesture at that point is ambiguous between pulling out and pulling in, and the "must be" phrasing is uncertain relative to "is."

His proposal easily accounts for the SCORE: as discussed above, parking space in a city is

[4]Like several other viewer-participants, she describes the green car as yellow.

valuable. It also resolves the uncertainty about the car's motion direction. With a further deduction, it explains why the driver of the white car is so annoyed as to honk *and* shout: that car was preparing to enter the space itself. (Both participants immediately make this deduction, as you can see in transcript 2.) It suggests an explanation for the SCREECH: the green car made some sort of rapid maneuver in order to cheat the white car out of the space, and its tires, not the white car's tires, are squealing. This explanation is backed up by other picture elements—the gray motion trail, and the position of the SCREECH relative to that trail—but neither element is discussed in the video, and we can't be sure the participants reach that conclusion. However, with that caveat, the "scored the spot" interpretation neatly accounts for all the information available to the listener.

Viewpoint, Discourse, and Meaning

The co-construction of meaning is a social as well as a cognitive process. How does the listener break away from the meaning the viewer has constructed? How does he move from nodding and asking for clarification, to proposing an alternative solution? His changing role seems to be tied to his physical and cognitive viewpoint, which goes through four sequential phases, as revealed in his speech and gesture. In the first phase, he takes the same viewpoint as the viewer. In the second, he takes a viewpoint within the depicted world, while the viewer is still taking an external, picture-space viewpoint. In the third, he switches visual viewpoint, from the viewer's to his own, and in the fourth, he shows a gesture mismatch with the viewer's gesture.

L: So, the larger angle of the road is going [off to the] ... left-hand side?
*Two hands on **right** side of body, about six inches apart, parallel*
L: And then kinda [shrinks down as it goes toward] the right-hand side?
*Both hands move across his body to his **left** side, getting closer together as they do*

Transcript 6: Mirrored gestures

V: On the right hand side of the street, there are
1s pause with lips moving—counting?
[five cars.] [One] of them is [in the middle] of the street driving.
R forearm held horizontally at eye level throughout.
LH four fingers splayed, near R hand
LH folds into index finger pointing back-and-forth below R forearm with L index finger
L: Holdonholdon. 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. [. . . (inaudible)
V: [Okay. So [back to this one]
Both arms held in V-shape across torso
the [long side] of the street
waves upper arm back and forth a bit
L: Okay, so the [upper side] of the street
flat RH in air above, in front of head
V: Yeah.
L: [So there's four cars
V: [Looking at it your way
V rotates 90° in chair, swings the 2-arm hold around another 90°

Transcript 7: Viewpoint clash

L:	how many cars are in the picture?
	...
L:	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
	...
L:	Okay.
	They're all moving cars, no parked cars
	on the side of the street.

Transcript 8: Switching visual viewpoint

V:	And he's tryna [pull out] into the road.
	R arm returns to diagonal hold
	[*LH B moves down from R arm,*
	fingers pointing up]
L:	he's tryna pull out into the road?
	RH B, palm facing left, fingers forward,
	hand moves out from body

Transcript 9: Gesture–speech mismatch

In transcript 6, there is an apparent mismatch between the listener's speech and his gesture. He is mimicking the two-arm gesture she used earlier in the discourse while describing the road, but he speaks of things on the "left" and gestures to his own right, and vice versa. He is not confused; he is facing the viewer, so his right-moving gestures appear to be left-moving from her perspective. He hesitates slightly before saying "left," which suggests additional cognitive load rather than a mistake. These details suggest that he is taking her visual viewpoint. Sweetser (TBD) observes that visual viewpoint can indicate cognitive viewpoint; in this case, the listener is using the viewer's visual viewpoint in preference to his own, indicating a close alignment with her cognitive viewpoint.

In transcript 7, the listener raises an objection to the viewer's mention of cars on the "right-hand side of the street." This demonstrates a viewpoint mismatch between the two participants. The viewer is thinking of the picture as a picture; she defines "right-hand side" of the street relative to the vantage point where the picture was taken. She clarifies her thinking by bringing back a gestural representation of the street in the picture that she used earlier, thus explicitly tying her description to the picture space. She seems to think he's confused about the orientation, so she turns around in her swivel chair so her gestures' "left" and "right" will align with his. This last suggests that she didn't notice that he was using mirrored gestures earlier.

The listener, on the other hand, objects because he is taking a viewpoint within the depicted world, imagining a dynamic scene with cars moving along a street. Within this space, "right-hand side" is ambiguous because a car is always on the right-hand side of the street *from the perspective of its driver*.[5] His ability to do so, and her confusion at his query, suggest that the listener's understanding of the image is already coherent enough to take the depicted-world viewpoint and hers isn't, even though he can't see it and she can.

Before the interchange in transcript 7, the listener has been asking for clarifications and additional information. His requests are all phrased as questions. In transcript 8 we see him continue the viewpoint realignment that he began in transcript 7. At the beginning of that transcript he is again taking a picture-space viewpoint. However, he then hesitates and changes his gesture so that it matches *his* left-right orientation, and not the viewer's. Furthermore, once he switches gesture

[5]Provided that the driver is following the traffic laws of the USA—the natural assumption for participants from the UCSD community.

viewpoint, his questions change grammatical form: some are phrased as statements. Rather than continue to ask for clarifications about the viewer's understanding of the picture, he is now presenting his own understanding for her to accept or reject. The last sentence in transcript 7 is the first of these.

In transcript 9 When the viewer says "he's tryna pull out into the road," her gesture is iconic for a car backing up. Assuming that both speakers are mapping the front of the car onto their fingers and the back onto their palms, the viewer's gesture depicts the car backing out, but the listener's gesture depicts the car moving forward.[6] In addition, if he has set up his depicted world space in a way consistent with the viewer's (as might be expected from his earlier, closely aligned viewpoint) then his gesture is that of a car pulling into a space at the "top" of the road, not pulling out of a space. It is at this point that he provides the interpretation that they both agree is correct (transcript 2).

The listener's visual viewpoint indicates his cognitive viewpoint, which starts closely aligned with the viewer's, and then shifts as he gains independent understanding of the image. The interaction of his speech and gesture shows that he is actively trying to build meaning right from the beginning. Once he gains cues (frame structure and details of spatial layout) with which to build that meaning, he can reason in terms of his own model and find out how closely it aligns with the viewer's.

Discussion

Figure 4 summarizes the final mental representation that the participants reach, using a simplified, informal version of the blending diagrams used in Fauconnier and Turner (2002). Each box represents a mental space. At the top left corner of the diagram is the picture space, structured by the raw percept. Directly below that is the depicted-world space, and below that is the final interpretation. On the right, we have the two frame-based spaces that are ultimately identified as providing the best explanation of the image.

The text within each box lists the elements of the space that are relevant to the conversation we have presented. (Some elements, such as the motion trail, did not appear within the conversation, but we present them for completeness.) The large gray arrows indicate blending operations. The color-coding and horizontal multicolored arrows show mappings between the input spaces. For simplicity, generic spaces are not shown.

With this representation in hand, let us consider the process by which the participants reached it. It is important to keep in mind that the diagram shows the final result, and that the blending steps are *not* sequential in time. The participants build the structure of five spaces in working memory when they begin to discuss the picture; during the discourse they fill in elements and cross-space relations. Even in the very first transcript the viewer was asking questions that could only be answered by reasoning relative to the scene interpretation (e.g. *Is someone saying SCORE?*) The viewer proposes several different frames to structure the right-hand input of one or other of the blends. Each frame considered generates a space that offers mappings for more elements of the left-hand input than the previous candidate did. The listener attends to cues in all modes: speech, gesture, fluency, timing, and posture. He attends selectively to frame-relevant information, and

[6] Both possibilities are consistent with the orientation of the car in the picture, if not with the motion trail and the SCREECH.

asks frame-challenging questions. His questions often target elements marked with uncertainty cues given by the viewer. Finally, he offers a refinement of the interpretation that integrates all information available to him. By putting together data on linguistic and gestural viewpoints, we have been able to trace the stages of both participants' conceptualizations in a way that would have been difficult or impossible using only the linguistic track.

The participants are in a laboratory setting; one might object that the constraints on their behavior (especially, the listener's not being allowed to turn around and look at the picture) are not realistic. While the precise setup here is contrived, we would not have been able to provoke the

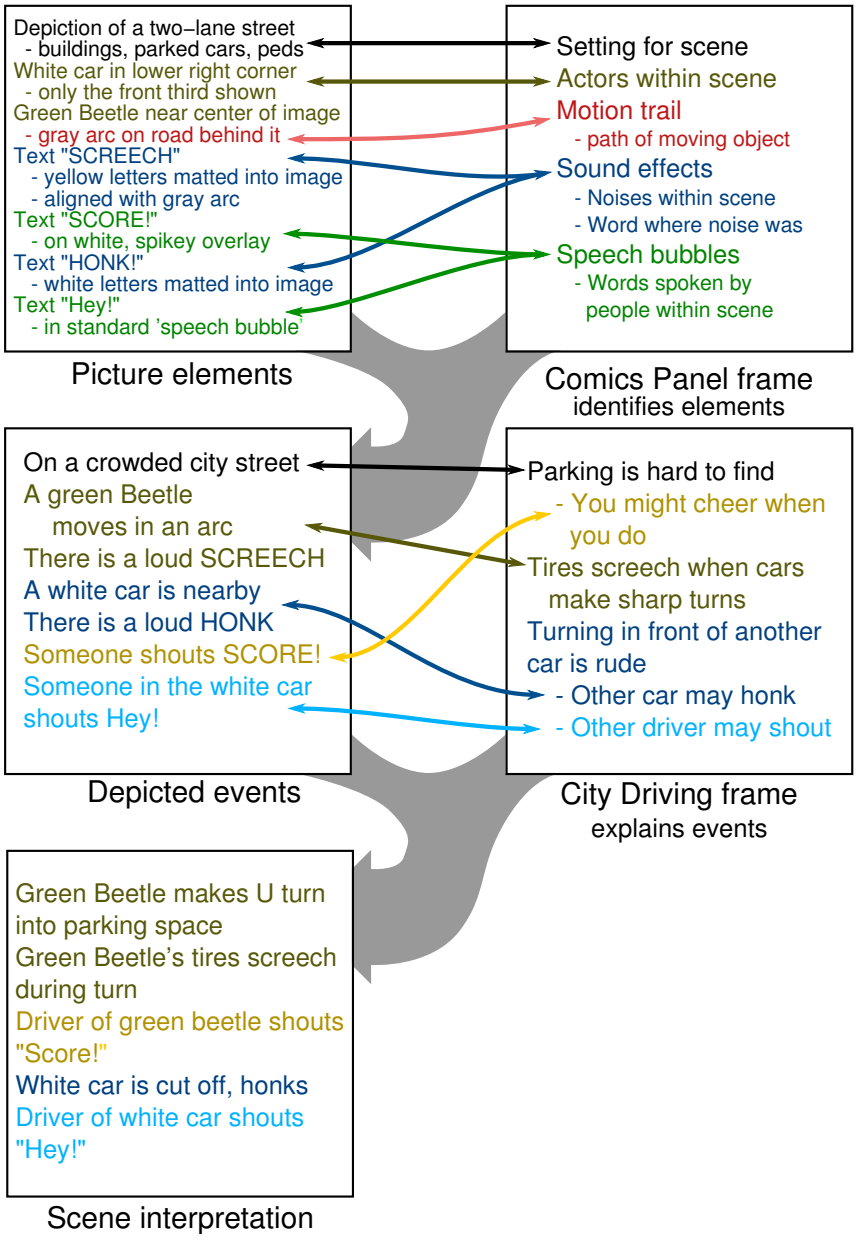

Figure 4: Blending analysis of figure 1.

behavior we have described if both participants could see the image. Furthermore, the behavior we see from these participants is not different in kind from behavior observed in more naturalistic settings. For instance, the viewer in this case study used a repeated arm gesture as a "material anchor" (Hutchins, 2005) for a shared representation of the road. We see similar gestural anchors from many participants in the main run of the study. We also see some participants using the water bottles we provided (replacing the mugs seen in figure 3) and the surface of the small table to illustrate aspects of picture layout. Smith (2003) reports similar phenomena in his observations of students collaboratively solving math homework; they continually make use of ad hoc material anchors drawn from their environment, ranging from hand gestures and writing on air to complex use of whiteboard diagrams together with gesture.

Situations involving asymmetric access to discourse-relevant information are common. Practically any time two people have a conversation, each of them knows things the other doesn't. This is especially evident in settings such as negotiation and interrogation, where some or all participants do not wish to reveal everything they know; instruction, where one participant is actively trying to convey new information to the others; or collaborative problem solving, where the goal is to combine all available knowledge and insight. It can also appear in more casual conversation, for instance one person telling another about the events of their day. This sort of dialogue often turns into collaborative "puzzling out" of confusing or awkward situations, so the listener can take an active role here too. All of these interactions can involve successive refinement of a complex blend shared among all participants. Thus, we believe that our experimental setting is a good model for some categories of naturalistic human discourse.

We close by pointing out that despite the initial confusion, our participants manage to communicate, and do so without major effort or adopting marked strategies. In fact, we could argue that what is most remarkable about the dialogues we presented is how very ordinary they are, up till the listener's spectacular leap of insight.

Acknowledgments

This paper is based on a presentation given at CSDL 2006, with additional material from a presentation at ICLC 2007 (Narayan, 2007). We would like to thank Amaya Becvar, Seana Coulson, Edwin Hutchins, Christine Johnson, Jim Hollan, Nathaniel Smith, Eve Sweetser, and the members of the DCOG and ECL groups for their helpful comments.

Shweta Narayan was supported by an NSF dissertation improvement grant during the research presented in this paper.

References

Farley, P. (2003). Barracuda: The Scotty Zaccharine Story. Webcomic. `http://www.e-sheep.com/zaccharine/`.

Fauconnier, G. and Turner, M. (2002). *The Way We Think: Conceptual blending and the mind's hidden complexities*. Basic Books, New York.

Fillmore, C. (1985). Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254.

Hutchins, E. (2005). Material anchors for conceptual blends. *Journal of Pragmatics*, 37(10):1555–1577.

McCloud, S. (1999). *Understanding Comics: The Invisible Art*. Paradox Press, New York.

Narayan, S. (2007). Visual and cognitive viewpoint in an interactional task. In *International Cognitive Linguistics Conference*.

Parrill, F. and Sweetser, E. (2004). What we mean by meaning: Conceptual integration in gesture analysis and transcription. *Gesture*, 4(2):197–219.

Smith, N. (2003). Gesture and beyond. Undergraduate thesis, Program in Cognitive Science, University of California at Berkeley.

Sweetser, E. (TBD). *Citation to be determined*.