SPEECH, ACTION AND GESTURES AS CONTEXT FOR ONGOING TASK-ORIENTED TALK

Justine Cassell
MIT Media Lab
20 Ames Street
Cambridge, MA 02139-4307
justine@media.mit.edu

The other project was a scheme for entirely abolishing all words whatsoever; and this was urged as a great advantage in point of health as well as brevity. . . An expedient was therefore offered, that since words are only names for things, it would be more convenient for all men to carry about them such things as were necessary to express the particular business they are to discourse on. . . I have often beheld two of those sages almost sinking under the weight of their packs, like peddlers among us; who, when they met in the streets, would lay down their loads, open their sacks, and hold conversation for an hour together; then put up their implements, help each other to resume their burdens, and take their leave.

Jonathan Swift Gulliver's Travels Chapter V, Part III

In this paper, I address the question of why communicating autonomous agents need not only language but also bodies and a world in which to interact or, otherwise put, what the relationship is among actions, gestures and words. In order to place the issue at hand -- theoretical considerations in designing a semantics for artificial agents -- in perspective, I start by discussing the development of situated language in children, then turn to the ways in which language can be situated for adults, arriving at the idea that the gesture - speech relationship partially determines the context for a given communicative act. These notes are very much a work in progress, the beginnings of an attempt to take seriously Malinowski's (1923) claim that the study of language in use reveals "the dependence of the meaning of each word upon practical experience, and of the structure of each utterance upon the momentary situation in which it is spoken."

Background

A growing body of evidence shows that people unwittingly produce gestures along with speech in many different communicative situations. These gestures have been shown to elaborate upon and enhance the content of accompanying speech (McNeill, 1992; Kendon, 1972), often giving clues to the underlying thematic organization of the discourse or the speaker's perspective on events. Gestures have also been shown to identify underlying reasoning processes that the speaker did not or could not articulate (Church and Goldin-Meadow, 1986).

We know that gestures are still produced in situations where there is no listener, or the listener cannot see the speaker's hands , although more gestures may be produced when an addressee is present . In addition, we know that information appears to be just about as effectively communicated in the absence of gesture -- on the telephone, or from behind a screen, and thus gesture is not **essential** to the interpretation of speech. But when speech is ambiguous or in a speech situation with some noise, listeners do rely on gestural cues (and, the higher the noise-to-signal ratio, the more facilitation by gesture). And, when adults are asked to assess a child's knowledge, they are able to use information that is conveyed in the child's gesture (and not conveyed by the child's speech) to make that assessment. Finally, it has been shown that when gestures are present in the communicative situation, listeners heed the information they convey, ultimately seeming to build one single representation out information conveyed in the two modalities (see Cassell & McNeill, forthcoming, for all references).

Gesture Change in Ontogenesis

Imagine the following child:

- Thirteen month old Lucinda wants to play with the telephone, sitting on a table. She points at her mother, at the telephone, and cries.
- Two year old Lucinda is playing at talking on the phone: she holds the toy receiver to her ear, and babbles.
- Three year old Lucinda is telling her mother that she wants to dial the phone to talk to Grandma: she says "I wanna [gesture]" and makes concentric circles with her forefinger to represent the unknown word "dial".
- Four year old Lucinda is playing at talking on the phone: she grips an invisible receiver to her ear, makes dialing motions with her other forefinger, and says "I'm going to call my Gramma on the telephone".

Pointing gestures precede language in children. The earliest pointing gestures appear to function as bids for the attention of an adult. Later pre-verbal pointing gestures, however, seem to have a referential function -- that is, pointing serves to isolate an object of reference from its

context and this indicative function distinguishes pointing from reaching, and has been argued to be an early example of symbol use (Werner & Kaplan, 1963). From reference by pointing, the child turns to reference by iconic representation of some aspect of the object or event to which reference is intended. For example, Bates, et al. (1983) describe the use in communication of non-indexical manual gestures (that is gestures with an object that pantomime the object's function such as picking up a brush and making brushing motions in the air). Piaget (1952) argued that the child's first representations were built from sensorimotor schemata -- that is, the schemata that underlie movements by the child's body that resemble objects or actions. In fact, in the infant, action schemata may constitute the matrix of a true propositional language (McNeill, 1975). Once some language is acquired, gestures are used by the child to replace words that the child doesn't know. Thus iconic gestures -- usually enactments of an object or action -- replace words in mother-child interaction (Acredolo & Goodwyn, 1991).

In sum, the path followed by hand movements in the child's development is two-fold. On the one hand, once language has been at least partially acquired, gesture leaves off its role as sole referential medium, and then can be reintegrated into communicative activity as an integral but non-redundant accompaniment to language. On the other hand, hand movements that once interacted with objects to represent activities begin to represent both the objects and the activities, thus detaching those movements from their reliance on the external environment -- distinguishing action from gesture.

Gesture Change in Microgenesis

Children are not the only ones progressing from actions to representations of perception and actions, and from thence to the propositional representations underlying language, including non-redundant contributions by gesture. Adult speakers may perform the same task, in a much shorter time span, when they *learn* a new concept or semantic frame. This point is demonstrated in the following section.

Imagine the following scenario. Lucinda (now an adult homeowner) is watching "This Old House", a television show about do-it-yourself home renovation[1]. During the credits we see the hero of the show repairing windows in a beautiful old home. During the show, the narrator, talking about weatherproofing a Victorian home, is describing the caulking gun. The narrator picks up a caulking gun from the table in front of him and introduces its use:

"This is a caulking gun, which one fills with tubes of caulk, and which is used to fill and waterproof exposed wood."

As the narrator speaks, he lifts the handle to show where the caulk tube is inserted, and lowers the handle to show how to extrude the caulk. He also points to a tube of caulk on the workbench. He then replaces the tool on the workbench, and continues his discussion of exposed wood by explaining how to waterproof window ledges. In this second part of the discussion, the narrator describes how to use the caulking gun. The narrator starts, however, by framing the relevance of his talk[2]:

"In this [next part]$_A$ I'm going to tell you how to use a caulking gun to [prevent leakage]$_B$ through [storm windows]$_C$ and [wooden window ledges]$_D$ (. . .) Press the [handle of the caulking gun slowly as you move the nozzle across the window ledge that needs caulk]$_E$" .

The narrator makes the following gestures during this segment of talk:

A. The narrator opens his hands, with the palms facing one another and the fingers facing away from his body, and then moves his wrists so that his fingers are facing down -- as if he is delineating a box.

B, C, D. The narrator's right hand is lax, but he flips it over and back so that the palm is facing first upward and then downward.

E. The narrator forms his left hand into a fist to represent the body of the caulking gun, holding it diagonally in front of his body, and uses his right hand to represent the hand of the user, pumping an imaginary handle up and down.

Lucinda goes to Home Depot to pick up materials to repair her leaking windows. She says to the salesperson "where do I find that . . . [gesture 'D'] to fill the cracks in my window ledges?"

Lucinda has learned the concept of 'caulking gun' from a linguistic-action-gestural performance[3]. Note that while the narrator is first

[1] This example -- as with the earlier examples from Lucinda's life -- is fictive, but constructed on the basis of an intimate and long-term association with how-to gestures, instructions concerning old home renovation, and caulk guns.

[2] Square brackets indicate the extent of co-temporaneous gestures.

[3] The following terminological distinction is useful here: I will refer to speech as linguistic, manipulations of objects

defining the caulking gun in speech, he is adding non-redundant features to that definition with his actions. Only in his actions does he show the relationship between the caulking gun and the tube of caulk (e.g. that the tube is fit into the cradle of the caulking gun), and the manner of using the caulking gun (e.g. that pushing down the handle extrudes the caulk). Likewise, in the second part of the narrator's description, the speech and gesture are non-redundant: only in gesture is the manner of movement demonstrated. By this second part of the performance, the hands have become a *symbol* of the object spoken of: an iconic representation of the tool and the home owner's interaction with it[4]. And not only representational gestures are used here; three other qualitatively different gestures are also a part of this communicative act.

Thus, in the time of one communicative event, the context of talk has gone from being the world itself -- as populated by do-it-yourself-show hosts, caulking guns, and tubes of caulk -- to being a representation of the world. In other words, the background against which talk is being interpreted by the listener has gone from being the Victorian home in the country, to the studio of "This Old House", to an imaginary event of home renovation. Although the *focal event* (Goodwin & Duranti, 1992) has remained the same -- the use of a caulking gun -- the *background* has changed.

Representations of gesture and speech

My discussion here arises from the basic claim that our conceptual representations are never entirely propositional or linguistic. That is, as described in earlier work, instead they have linguistic aspects and imagistic aspects (Cassell & McNeill, forthcoming; Cassell, 1995). Here, I additionally suggest a motor schematic (or *action,* or even *kinesthetic*) aspect. When we communicate with others about objects and events, our communicative contributions arise from those complex conceptual representations. Thus, the output is linguistic and gestural, propositional and imagistic. This is not, by any means, a novel idea (see, inter alia, Fillmore, 1985). In this paper, however, I would like to concentrate on the ramifications of such a theory of conceptual representation for the generation and understanding of task-oriented language. In particular, using it to question our design of a discourse semantics, and our notion of the *context* of language.

With this in mind, several parts of the fix-it performance are worth examining. I'm only going to examine one here, but I point out the others as needing discussion.

- One should consider here the role of the actions without speech -- the engagement of the narrator in a physical world where, before the camera, during the credits, he caulks the windows of a lovely Victorian home. This action contextualizes the episode to follow, much as a cover photo tells us about the topic of a book. Does it have ongoing consequences for the interpretation of the communicative acts.

- One should also address the passage from tool use to representation of tool use in speech and action. What aspects of the tool must be represented for the action to still make sense? What aspects of the tool can *more easily* be represented without the tool present? Is there a kinesthetic component to representations of tool use that is composed of the *feel* of the tool? Must this then be re-enacted in the demonstration of that tool?

- The topic that I will address is the passage from action to gesture, where gesture represents action and object, and also indexes other aspects of the communicative event, in particular, what is taken to be focal and what is context at any given moment in the discourse.

In what follows, I will describe four types of gestures that represent different kinds of concepts, contextualize discourse in different ways, and I will describe how these contextualization cues act as resources for interpreting ongoing communicative acts.

- *Iconics* depict by the form of the gesture some feature of the action or event being described; such as the gesture that accompanied "Press the [handle of the caulking gun slowly as you move the nozzle across the window ledge that needs caulk]".

Iconic gestures may specify the manner in which an action is carried out, even if this information is not given in accompanying speech. For example, only in gesture does the narrator specify the essential information of how the handle of the caulk gun is to be manipulated.

Iconic gestures may also specify the viewpoint from which an action is narrated. That is, gesture can demonstrate who narrators imagine themselves to be, and where they imagine themselves to stand

as action, and hand movements that do not interact with objects in the real world as gestures.

[4] The interpretation of passage from action to symbol (as well as the idea of looking for examples of this movement in home improvement TV shows) is from Streeck, 1995.

at various points in the narration, when this is rarely conveyed in speech. For example, the narrator in his gestures takes the role of the homeowner repairing her window. In another TV crafts show, describing blowing glass, the narrator might puff out his mouth, and make a balloon with his hands to represent the glass expanding at the edge of the pipe. Here he is enacting the part of the glass. Alternatively, he might have blown out with his lips and mimed holding a glass-blowing pipe, to represent the glass blower.

- *Metaphoric gestures* are also representational, but the concept being depicted has no physical form. An example is "the meeting went on and on" accompanied by a hand indicating rolling motion. There need not be a productive metaphor in the speech accompanying metaphoric gestures; sometimes the "metaphors" that are represented in gesture have become entirely conventionalized in the language. There does need to be a recognizable vehicle that mediates between the form of the gesture and the meaning of the speech it accompanies.

Some common metaphoric gestures are the 'process metaphoric' just illustrated, and the 'conduit metaphoric' which objectifies the information being conveyed, representing it as a concrete object that can be held between the hands and given to the listener. Conduit metaphorics commonly accompany new segments in communicative acts; an example is the box gesture that accompanied "In this [next part] I'm going to tell you how to use a caulking gun". Metaphoric gestures of this sort contextualize communication; for example, placing it in the larger context of social interaction. In this example, the speaker has prepared to give the next segment of discourse to the television audience.

- *Deictics* spatialize, or locate in the physical space in front of the narrator, aspects of the discourse; these can be discourse entities that have a physical existence, such as the tube of caulk that the narrator pointed to on the workbench, or non-physical discourse entities. An example of the latter might be pointing over one's shoulder while saying "in last week's show. . .".

Deictic gestures populate the space in between the speaker and listener with the discourse entities introduced and referred to.

- *Beat gestures* are small baton like movements that do not change in form with the content of the accompanying speech. They serve a pragmatic function, occurring with comments on one's own linguistic contribution, speech repairs and reported speech. An example is "she talked first, I mean second" accompanied by a hand flicking down and then up.

Beat gestures may signal that information conveyed in accompanying speech does not advance the "plot" of the discourse, but rather is an evaluative or orienting comment. For example, the narrator described the content of the next part of the TV episode by saying "I'm going to tell you how to use a caulking gun to [prevent leakage] through [storm windows] and [wooden window ledges]. . ." and accompanied this speech with several beat gestures to indicate that the role of this part of the discourse was to indicate the relevance of what came next, as opposed to imparting new information in and of itself.

Gestures as context for ongoing talk

Importantly, the gestures produced by the speaker become a part of the shared context, and the speaker assumes that those gestures are accessible to the listener. First of all, they are available for later co-reference. For example, a visual artist describing a non-representational free-form sculpture[5] (in the sculpture's absence) said

> It has several [wavy parts].
> *makes 3 quick wave gestures at 3 different places in space*
>
> Now, one of them [folds in on itself]
> *repeats wave in the last location & adds a convoluted gestures*

Only the gesture tells us which wave of the sculpture folds in on itself.

Gestures are also available in context to disambiguate anaphoric reference. That is, speakers use deictic gestures to indicate which discourse referent they were referring to, as in the following example. A speaker who was retelling a Hitchcock film said "[Frank] looks at the [artist] before [he leaves] the restaurant". In speech, the "he" is ambiguous between Frank and the artist. The speaker, however, points at a space to his left when he says "Frank", to his right when he says "artist", and then performs a leaving gesture from Frank's location in space, thus disambiguating the reference. This is, of course, identical to anaphora in American Sign Language, and we know that signers have the ability to keep track of several anaphoric chains simultaneously, indexed by different locations in space.

And gestures interact with information structure in discourse, occurring more frequently with new

[5] These data were collected by Wendy Plesniak. My thanks for the permission to describe them here.

information than with given information, with rhematic material than with thematic material (Cassell *et al*, 1994). This distribution is available to signal to listeners when information is to be particularly attended to.

I would claim, further, that gestures have consequences for later speech; that gestures serve as a context of interpretation for the speaker as well as the listener. In the following example, the gesture determines the properties of a simile being established, and the simile depends on those properties, available only in gesture. The speaker was describing another free-form sculpture, and said "It's like a helmet". He first held both hands in front of his chin with the palms open and fingers outstretched upward as if he was lifting a helmet off his head, and then he moved his hands down to the table in front of him, as if he was placing the helmet on the table. Now, note that "helmet" most commonly today means bicycle or motorcycle helmet, which covers mostly only the top of the head. But this speaker's representation of a helmet resembles a medieval full-face helmet, as shown in the full-face gesture and the speaker goes on to refer to the properties of that type of object: "and the eyes are sunken" and he gestures in at roughly where the eye holes would be in the helmet he has placed on the table.

In addition, gestures serve not only as a context for semantic interpretation, of referents etc., but also as a frame for the situating of whole segments of discourse. Thus, the part of the caulking gun explanation accompanied by gesture 'A' is not about caulking guns, but about the explanation of caulking guns: it is intended to orient the listener to what is about to be conveyed. In this sense, it is *metatextual* -- framing the coming speech and providing resources for its appropriate interpretation. The metaphoric gesture that accompanies this segment indexes the metatextual nature of the talk.

In this sense, the gesture-speech relationship resembles the interaction of words and graphics in the generation of multimodal text (Feiner & McKeown, 1991; Wahlster et al., 1991).

Designing a semantics for speech and gesture

This section has the status of a call for help. Several issues concerning the design of a discourse semantics are raised by the preceding remarks:

In a first attempt to generate gestures, speech and intonation together from one underlying representation, we (Cassell *et al*, 1994) implemented a discourse planner that relied on three levels of discourse representation and a level of semantic representation. The discourse levels were a representation of the intentional structure , a representation of the attentional state, and a representation of discourse purposes (as first proposed in Grosz & Sidner, 1986). These discourse levels allowed the assignment of information structure to discourse entities in such a way as to associate the occurrence of a new gesture with new information. The semantic level assigned a notion of 'spatializability' to concepts, such that concepts with a physical existence in the world might be represented by iconic gestures, concepts that could be located in space but not represented could be indexed by deictic gestures, and so forth. But, we were forced to generate the actual form of the gestures from a dictionary of gesture forms. This was a provisional solution; a richer semantics could include the *features* relevant for gesture generation, so that the form of the gesture could be generated algorithmically from the semantics.

What might such a semantics look like? The challenge is to find a representation for gesture that interacts with the representation for speech, that allows the agent of an action to be represented in speech, and the manner of the action to represented in gesture, for example. That is, the information conveyed in gesture and speech should not have to be overlapping. And this non-redundancy must be available at the semantic and at the discourse level. That is, gestures should be able to convey the discourse status of a segment, as well as being able to convey the anaphoric status of a discourse referent..

This is particularly difficult because gestures are polysemic -- they don't carry meaning outside of particular events of language use. Thus, the metaphoric gesture resembling a box, which accompanied "In this [next part]. . . " could equally well have been an iconic gesture, representing a box, and could have accompanied talk about how to make a waterproof box for children's bath toys. In sum, a kind of encyclopedic representation is called for, with a complex of semantic constituents, imagistic features, and a kinesthetic component.

References

Acredolo, L., & Goodwyn, S. (1991). Sign language in babies: The significance of symbolic gesturing for understanding language development. In R. Vasta (Ed.), Annals of child development (Vol. 7). Greenwich, CT: JAI Press

Bates, E., Bretherton, I., Shore, C., & McNew, S. (1983). Names, gestures, and objects: Symbolization in infancy and aphasia. In K. E. Nelson (Ed.), Children's language (Vol. 4). Hillsdale, NJ: Erlbaum.

Cassell, J. (1995). "The role of gesture in stories as multiple participant frameworks". AAAI Spring Symposium Working Notes.

Cassell, J., McNeill, D. & McCullough, K.E. *(forthcoming)* "Speech-Gesture Mismatches: Evidence for One Underlying Representation of Linguistic and Non-Linguistic Information". Cognition.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M. (1994) "ANIMATED CONVERSATION: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents". SIGGRAPH '94.

Cassell, J. and McNeill, D. (1991). "Non-verbal Imagery and the Poetics of Prose". Poetics Today Vol. 12:3. pp. 375-404.

Church, R.B. & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. Cognition, *23*, 43-71.

Feiner, S., & McKeown, K. (1991). "Automating the generation of coordinated multimedia explanations". IEEE Computer, 24:10.

Fillmore, C. (1985). "Frames and the Semantics of Understanding", In Quaderni di Semantica, VI:2.

Goodwin, C. & Duranti, A. (1991). "Rethinking Context: An introduction" In Rethinking Context. Cambridge: Cambridge University Press.

Grosz, B., & Sidner, C. (1986). "Attention, intentions, and the structure of discourse". Computational Linguistics, 12:3.

Kendon, A. (1972). "Some relationships between body motion and speech". In A. W. Siegman & B. Pope (Eds.), Studies in dyadic communication. New York: Pergamon Press.

McNeill, D. (1975). "Semiotic Extension". In L.E. Solso (Ed.), Information Processing and Cognition. Hillsdale, NJ: Lawrence Erlbaum.

McNeill, D. (1992) Hand and Mind: What gestures reveal about thought. University of Chicago Press.

Piaget, J. (1952) The Origins of Intelligence in Children. (M. Cook, Trans.) New York: Intnl Universities Press.

Streeck, J. (1995). "Experiential roots of gestures: From instrumental to symbolic action". Paper presented at *Conference on Gestures Compared Cross-Linguistically.* July 7-10, 1995.

Thorisson, K. (1995). "Computational characteristics of multimodal dialogue". In this volume.

Wahlster, W., Andre, E., Graf, W. & Rist, T. (1991). "Designing illustrated texts". In Proceedings of the 5th EACL: 8-14.

Werner, H., & Kaplan, B. (1963) Symbol Formation. New York: Wiley & Sons.