

# Communicating with Hands in Face to Face Conversation

Alex Lascarides  
Joint work with Matthew Stone

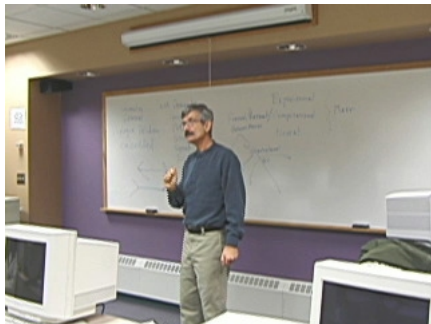
School of Informatics  
University of Edinburgh

Speckled Workshop 2009

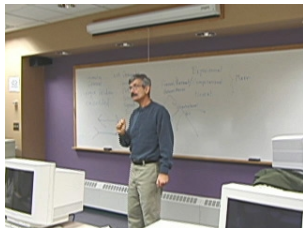
# Outline

- 1 Present some Data involving Gesture
  - Iconic Gesture
  - Deictic Gesture
- 2 Formal Semantic Analysis: Re-use devices from Linguistics
- 3 Challenges in mapping multimodal signal to Syntactic Form
- 4 Conclusion

# An Example of Iconic Gesture



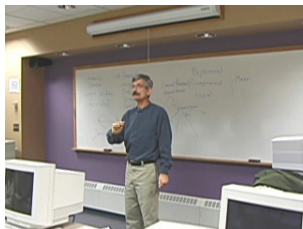
# Paraphrasing the Communicative Act



So that these very low-level phonological errors tend not to get reported. . .

*... because they are being produced continually by an iterative process below our level of awareness.*

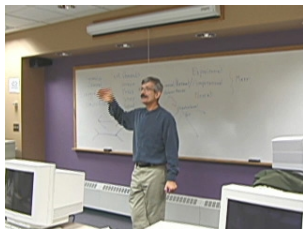
# Paraphrasing the Communicative Act



Now one thing you could do is totally audiotape hours and hours...

*... so that you get a large amount of data that you can think of as laid out on a time line.*

# Paraphrasing the Communicative Act



And exhaustively go through and make sure that you really pick up all the speech errors

*... by individually analysing each unit of analysis along the timeline of your data.*

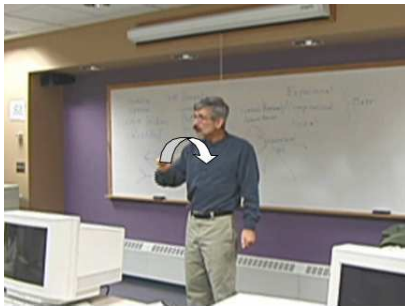
# Paraphrasing the Communicative Act



Allow two different coders to go through it. . .

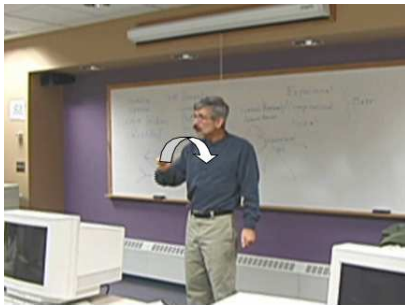
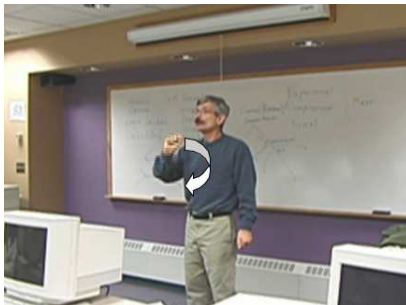
*... and moreover get them to work independently and reconcile their activities.*

## Gestural Form Underspecifies Meaning



- (1) There are these very low-level phonological errors that tend not to get reported.
- (2) The mouse ran on the wheel for a few minutes.

# Gestural Form Underspecifies Meaning



Need **qualitative** values to represent iconic form:

- hand shape is asl-a;  
path is sagittal circle;  
trajectory is iterative, clockwise. . .

# Semantic Scope Across Speech and Gesture

From the AMI corpus

(Carletta, 2007)

- (3) D: And um I thought not too edgy and like a box, more kind of hand-held more um . . . not as uh computery and organic, yeah, more organic shape I think.

When *D* says *computery*, her right hand has fingers and thumb curled downward (in a 5-claw shape), palm also facing down, and she moves the fingers as if to depict typing.

# Deictic Gesture



# Form and Meaning

## Form and Meaning of Deictic Gesture

- **Quantitative** representation of space is necessary.
  - Point to coordinate  $\vec{p} \in \mathbb{R}^4$
- But signal underdetermines form (which  $\vec{p}$  is designated?) and form underdetermines meaning (cf iconic gesture)

(4) That man is wearing a hat (*man physically located at  $\vec{p}$* )  
**equality**

(5) And Norris is exactly across from the library  
(*nothing physically located at  $\vec{p}_1$  and  $\vec{p}_2$* )  
**virtual-counterpart**

# Form and Meaning

## Form and Meaning of Deictic Gesture

- **Quantitative** representation of space is necessary.
  - Point to coordinate  $\vec{p} \in \mathbb{R}^4$
- But signal underdetermines form (which  $\vec{p}$  is designated?) and form underdetermines meaning (cf iconic gesture)

- (6) That man is wearing a hat (*man physically located at  $\vec{p}$* )  
**equality**
- (7) And Norris is exactly across from the library  
(*nothing physically located at  $\vec{p}_1$  and  $\vec{p}_2$* )  
**virtual-counterpart**

# Deictic and Iconic Dimensions



*It's this weird looking building. . .*

- Spatial region designated by *hand trajectory* rather than *hand position* (cf. 'standard' pointing gesture).
- Quantitative spatial dimension crucial to interpretation.
- Denotation of curved region resolved only through semantic connections to speech that exploit its *iconic dimension*.

## Insight (And Challenge)

- Hand movement yields ambiguous analysis of gestural form
- A disambiguated gestural form underspecifies meaning
- Gestural content is resolved through integration with speech, particularly by coherence embodied in rhetorical connections:
  - speech **because** gesture
  - speech **so that** gesture
  - speech **by** gesture
  - speech **and moreover** gesture
- Gesture and speech exhibit co-reference and scope relations, calling for a unified way for inferring and evaluating their meaning.

# Our Aims

- A logical model of gesture interpretation.
- Re-use devices for modelling meaning of language
  - Principles of pragmatic interpretation are general, applying to all communication, in whatever medium it takes place.

# What We Re-use from Linguistics

- 1 *Rhetorical Relations*: knit gesture and synchronous speech into a single thought.
- 2 *Dynamic Semantics*: models constraints on co-reference between speech and gesture and across gestures as the situated discourse proceeds.
- 3 *Syntactic Ambiguity*: Signal doesn't determine a unique representation of form.
- 4 *Underspecification*: The compositional semantics of gesture is highly underspecified, even when form is disambiguated:
  - 1 Form of iconic gesture depicts many alternatives.
  - 2 Underspecified bits must be resolved to yield a specific, coherent interpretation.

## What We Re-use from Linguistics

- 1 *Rhetorical Relations*: knit gesture and synchronous speech into a single thought.
- 2 *Dynamic Semantics*: models constraints on co-reference between speech and gesture and across gestures as the situated discourse proceeds.
- 3 *Syntactic Ambiguity*: Signal doesn't determine a unique representation of form.
- 4 *Underspecification*: The compositional semantics of gesture is highly underspecified, even when form is disambiguated:
  - 1 Form of iconic gesture depicts many alternatives.
  - 2 Underspecified bits must be resolved to yield a specific, coherent interpretation.

## What We Re-use from Linguistics

- 1 *Rhetorical Relations*: knit gesture and synchronous speech into a single thought.
- 2 *Dynamic Semantics*: models constraints on co-reference between speech and gesture and across gestures as the situated discourse proceeds.
- 3 *Syntactic Ambiguity*: Signal doesn't determine a unique representation of form.
- 4 *Underspecification*: The compositional semantics of gesture is highly underspecified, even when form is disambiguated:
  - 1 Form of iconic gesture depicts many alternatives.
  - 2 Underspecified bits must be resolved to yield a specific, coherent interpretation.

## What We Re-use from Linguistics

- 1 *Rhetorical Relations*: knit gesture and synchronous speech into a single thought.
- 2 *Dynamic Semantics*: models constraints on co-reference between speech and gesture and across gestures as the situated discourse proceeds.
- 3 *Syntactic Ambiguity*: Signal doesn't determine a unique representation of form.
- 4 *Underspecification*: The compositional semantics of gesture is highly underspecified, even when form is disambiguated:
  - 1 Form of iconic gesture depicts many alternatives.
  - 2 Underspecified bits must be resolved to yield a specific, coherent interpretation.

# An Incoherent Example

adapted from Numack corpus Kopp et al. (2004)

You walk out the doors.

1: *flat hand, vertical palm,  
fingers pointing right.*

2: *two fists, move in downwards arc  
1 or 2, but not 1 and 2!!*



- (8)
- a. Walk out the door.
  - b. Turn right.
  - c. ??Push the door handles down.

*Narration(a, b) ∧ Elaboration(a, c) violates right frontier.*

## Example Logical Form

There are these very low-level phonological errors that tend not to get reported.



$\pi_1$  :  $\exists y(\text{low-level}(y) \wedge \text{phonological}(y) \wedge \text{errors}(y) \wedge$   
 $\text{go-unreported}(e, y))$

$\pi_2$  :  $[\mathcal{G}]\exists x(\text{continuous}(x) \wedge \text{below-awareness}(x) \wedge \text{process}(x) \wedge$   
 $\text{sustain}(e', x, y))$

$\pi_0$  :  $\text{Explanation}(\pi_1, \pi_2)$

*How do we get content of  $\pi_2$  from gestural form, and how do we get gestural form from the hand movement?*

# Iconic Gesture: Form

No hierarchical structure McNeill (1992), Kopp et al. (2004):

- Form is a multi-dimensional matrix.
- Each component represents an aspect of the gesture's form which potentially reveals things about its meaning.
  - Hand shape, finger direction, palm direction, position (relative to torso), path of movement. . .

## An Example

Gesture for (1) and (2) is:

(9)  $\left[ \begin{array}{l} \textbf{qualitative-characterising-gesture} \\ \text{hand-shape} : \textit{asl-a} \\ \text{finger-direction} : \textit{down} \\ \text{palm-direction} : \textit{left} \\ \text{trajectory} : \textit{sagittal-circle} \\ \text{movement-direction} : \{ \textit{iterative}, \textit{clockwise} \} \\ \text{location} : \textit{central-right} \end{array} \right]$

- Each attribute value element may convey a specific analogous piece of descriptive content.
  - *asl-a* hand-shape can resolve to a 1-place predicate (for (2)), 3-place predicate (for (1)), ...

How do we express this?

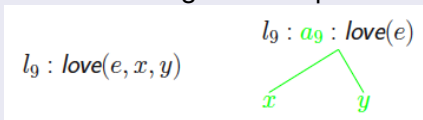


## Shallow Semantics with RMRS: POS Tagger

Every\_AT1 black\_JJ cat\_NN1 loved\_VVD some\_DD dog\_NN1.

Missing lexical subcategorisation  $\rightsquigarrow$  unknown arity of predicate

- Predicate's arguments specified separately via *anchors*:



- $l_9 : a_9 : \text{love}(e)$ ,  $\text{ARG1}(a_9, x)$ ,  $\text{ARG2}(a_9, y)$

No syntactic dependencies  $\rightsquigarrow$  No semantic dependencies

- Label and variable equalities can be specified separately too.

- $l_4 : \text{black}(x)$ ,  $l_4 : \text{cat}(x)$       $l_{41} : a_{41} : \text{black}(x_{41})$ ,  $l_{42} : a_{42} : \text{cat}(x_{42})$   
 $l_{41} = l_{42}$ ,  $x_{41} = x_{42}$

# Shallow Semantics with RMRS: POS Tagger

Every\_AT1 black\_JJ cat\_NN1 loved\_VVD some\_DD dog\_NN1.

RMRS from POS Tagger RMRS from Deep Parser

$l_1 : a_1 : \text{every}(x_1), \text{RESTR}(a_1, h_1), \text{BODY}(a_1, h_2)$

$l_2 : a_2 : \text{black}(x_2)$

$l_3 : a_3 : \text{cat}(x_3)$

$l_4 : a_4 : \text{loved}(e), \text{ARG1}(a_4, x_4), \text{ARG2}(a_4, x_5)$

$l_5 : a_5 : \text{some}(x_6), \text{RESTR}(a_5, h_3), \text{BODY}(a_5, h_4)$

$l_6 : a_6 : \text{dog}(x_7)$

$x_1 = x_2, x_2 = x_3, x_3 = x_4, x_5 = x_6, x_6 = x_7$

$l_2 = l_3, h_1 =_q l_3, h_3 =_q l_6$

# Shallow Semantics with RMRS: POS Tagger

Every\_AT1 black\_JJ cat\_NN1 loved\_VVD some\_DD dog\_NN1.

RMRS from POS Tagger RMRS from Deep Parser

$l_1 : a_1 : \text{every}(x_1), \text{RESTR}(a_1, h_1), \text{BODY}(a_1, h_2)$

$l_2 : a_2 : \text{black}(x_2)$

$l_3 : a_3 : \text{cat}(x_3)$

$l_4 : a_4 : \text{loved}(e), \text{ARG1}(a_4, x_4), \text{ARG2}(a_4, x_5)$

$l_5 : a_5 : \text{some}(x_6), \text{RESTR}(a_5, h_3), \text{BODY}(a_5, h_4)$

$l_6 : a_6 : \text{dog}(x_7)$

$x_1 = x_2, x_2 = x_3, x_3 = x_4, x_5 = x_6, x_6 = x_7$

$l_2 = l_3, h_1 =_q l_3, h_3 =_q l_6$

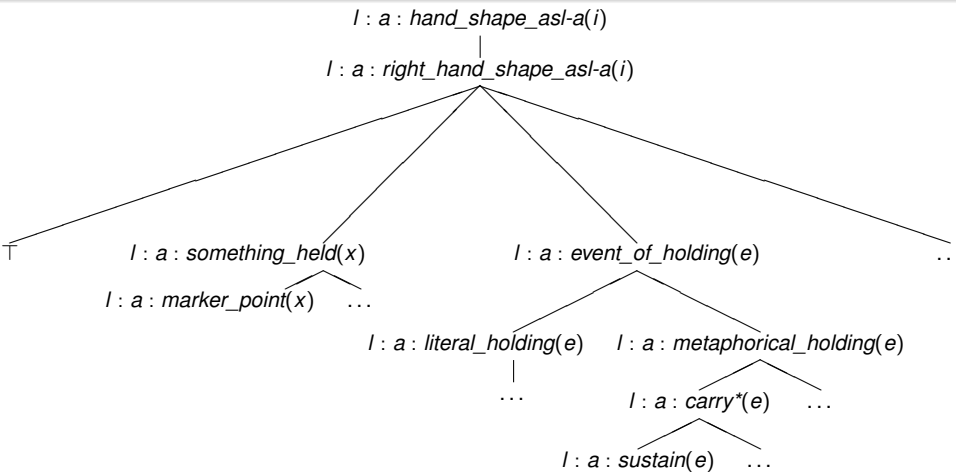
# Underspecified Semantics of Iconic Gesture

- Each attribute-value element may convey a specific analogous piece of descriptive content.
- Convention reads the underspecified predicate directly off the feature structure:

$$I_1 : a_1 : \textit{hand\_shape\_asl-a}(I_1)$$

with each attribute value having a unique label and semantic index (cf semantics from a POS tagger).

# Hierarchy for Resolving Underspecified Predicates



*marker\_point*: 1-place predicate

*sustains*: 3-place predicate

# Form of Deictic Gesture

## identifying-gesture

right-hand-shape : *loose-asl-5-thumb-open*

right-finger-direction : *forward*

right-palm-direction : *up-left*

right-location :  $\vec{c}$

## Ambiguity in Form

- iconic vs. deictic vs. combination of both.
- Portion of movement that's the stroke
  - Includes hand trajectory?

## Form of Deictic Gesture

### **identifying-gesture**

right-hand-shape : *loose-asl-5-thumb-open*

right-finger-direction : *forward*

right-palm-direction : *up-left*

right-location :  $\vec{c}$

### Underspecified Meaning

- Designation of space: hand vs. distant
- Mapping from physical space to space depicted in meaning
- Relation between gestured individual and depicted space (e.g., literal location vs. metaphorical)
- Relation between gestured individual and individual in speech

## The Form of Gesture and Speech Combined

- Synchronicity carries meaning, and should be defined in terms of *prosody* and *constituency*.

Example

(Loehr corpus)

They made everything in the room GREASY

- In general there are attachment ambiguities for combining gesture with speech.
- Pragmatics (sometimes) resolves this syntactic ambiguity.

# Interpreting Gesture

Search for resolutions of underspecified content that maximise coherence

*The mouse ran on the wheel for several minutes*

- This interpretation:

*gesture depicts a point on the wheel moving*

supports an inference to *Elaboration*:

- Both constituents to be related describe physical movement of the same object.

## Same Gesture, Different Context, Different Meaning

*There are these very low-level phonological errors that tend not to get reported.*

- Inferring *Elaboration* via a gesture interpretation denoting physical movement is impossible.

- But the gesture can visualise why (1) is true:

|                        |    |                                 |
|------------------------|----|---------------------------------|
| fist                   | ~> | x <i>sustains</i> speech errors |
| circular path          | ~> | x is continuous;                |
| iterative movement     | ~> | x is iterative                  |
| (low) central position | ~> | x is subconscious               |

yielding *Explanation*.

# Challenges for Automating Signal to Form (cf. ASR)

- 1 Identifying which movements are communicative and which aren't.
- 2 Identifying the temporal extent of the *stroke* of the gesture
- 3 Mapping speck values during stroke into a feature structure (perhaps more than one; cf. *n*-best list in ASR)
  - Qualitative values; quantitative values
- 4 For combining speech and gesture:
  - 1 Computing the relative timing of the stroke to speech string.
  - 2 Identifying pitch accents in speech (challenge for ASR).
  - 3 Computing syntactic structure of the speech (challenge for linguistic parsing)

## Challenges for Automating Signal to Form (cf. ASR)

- 1 Identifying which movements are communicative and which aren't.
- 2 Identifying the temporal extent of the *stroke* of the gesture
- 3 Mapping speck values during stroke into a feature structure (perhaps more than one; cf. *n*-best list in ASR)
  - Qualitative values; quantitative values
- 4 For combining speech and gesture:
  - 1 Computing the relative timing of the stroke to speech string.
  - 2 Identifying pitch accents in speech (challenge for ASR)
  - 3 Computing syntactic structure of the speech (challenge for linguistic parsing)

## Challenges for Automating Signal to Form (cf. ASR)

- 1 Identifying which movements are communicative and which aren't.
- 2 Identifying the temporal extent of the *stroke* of the gesture
- 3 Mapping speck values during stroke into a feature structure (perhaps more than one; cf. *n*-best list in ASR)
  - Qualitative values; quantitative values
- 4 For combining speech and gesture:
  - 1 Computing the relative timing of the stroke to speech string.
  - 2 Identifying pitch accents in speech (challenge for ASR)
  - 3 Computing syntactic structure of the speech (challenge for linguistic parsing)

## Challenges for Automating Signal to Form (cf. ASR)

- 1 Identifying which movements are communicative and which aren't.
- 2 Identifying the temporal extent of the *stroke* of the gesture
- 3 Mapping speck values during stroke into a feature structure (perhaps more than one; cf. *n*-best list in ASR)
  - Qualitative values; quantitative values
- 4 For combining speech and gesture:
  - 1 Computing the relative timing of the stroke to speech string.
  - 2 Identifying pitch accents in speech (challenge for ASR)
  - 3 Computing syntactic structure of the speech (challenge for linguistic parsing)

## Challenges for Automating Signal to Form (cf. ASR)

- 1 Identifying which movements are communicative and which aren't.
- 2 Identifying the temporal extent of the *stroke* of the gesture
- 3 Mapping speck values during stroke into a feature structure (perhaps more than one; cf. *n*-best list in ASR)
  - Qualitative values; quantitative values
- 4 For combining speech and gesture:
  - 1 Computing the relative timing of the stroke to speech string.
  - 2 Identifying pitch accents in speech (challenge for ASR).
  - 3 Computing syntactic structure of the speech (challenge for linguistic parsing)

## Challenges for Automating Signal to Form (cf. ASR)

- 1 Identifying which movements are communicative and which aren't.
- 2 Identifying the temporal extent of the *stroke* of the gesture
- 3 Mapping speck values during stroke into a feature structure (perhaps more than one; cf. *n*-best list in ASR)
  - Qualitative values; quantitative values
- 4 For combining speech and gesture:
  - 1 Computing the relative timing of the stroke to speech string.
  - 2 Identifying pitch accents in speech (challenge for ASR).
  - 3 Computing syntactic structure of the speech (challenge for linguistic parsing)

## Challenges for Automating Signal to Form (cf. ASR)

- 1 Identifying which movements are communicative and which aren't.
- 2 Identifying the temporal extent of the *stroke* of the gesture
- 3 Mapping speck values during stroke into a feature structure (perhaps more than one; cf. *n*-best list in ASR)
  - Qualitative values; quantitative values
- 4 For combining speech and gesture:
  - 1 Computing the relative timing of the stroke to speech string.
  - 2 Identifying pitch accents in speech (challenge for ASR).
  - 3 Computing syntactic structure of the speech (challenge for linguistic parsing)

## Challenges for Automating Signal to Form (cf. ASR)

- 1 Identifying which movements are communicative and which aren't.
- 2 Identifying the temporal extent of the *stroke* of the gesture
- 3 Mapping speck values during stroke into a feature structure (perhaps more than one; cf. *n*-best list in ASR)
  - Qualitative values; quantitative values
- 4 For combining speech and gesture:
  - 1 Computing the relative timing of the stroke to speech string.
  - 2 Identifying pitch accents in speech (challenge for ASR).
  - 3 Computing syntactic structure of the speech (challenge for linguistic parsing)

# Conclusions

- Gesture meaning modelled with devices that are needed anyway for modelling language.
- Form of iconic gesture requires *qualitative* values.
- Form of deictic gesture requires *quantitative* values.
- Ambiguity in form and synchrony.
- Specks can provide a rich source of informative features for learning mappings from multimodal signals to multimodal syntactic form.

- J. Carletta. Unleashing the killer corpus: Experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation Journal*, 41(2):181–190, 2007.
- K. B. Emmorey, B. Tversky, and H. Taylor. Using space to describe space: Perspective in speech, sign and gesture. *Spatial Cognition and Computation*, 2(3):157–180, 2000.
- A. Kendon. *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004.
- S. Kopp, P. Tepper, and J. Cassell. Towards integrated microplanning of language and iconic gesture for multimodal output. In *Proceedings of ICMI, 2004*.
- D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, 1992.