Formal Semantics for Iconic Gesture

Alex Lascarides[1]
[1]Human Communication Research Centre
University of Edinburgh
Edinburgh EH8 9LW, UK
alex@inf.ed.ac.uk

Matthew Stone[1,2]
[2]Department of Computer Science
Rutgers University
Piscataway, NJ 08845-8019
Matthew.Stone@rutgers.edu

Abstract

We present a formal analysis of iconic coverbal gesture. Our model describes the incomplete meaning of gesture that's derivable from its form, and the pragmatic reasoning that yields a more specific interpretation. Our formalism builds on established models of discourse interpretation to capture key insights from the descriptive literature on gesture: synchronous speech and gesture express a single thought, but while the form of iconic gesture is an important clue to its interpretation, the content of gesture can be resolved only by linking it to its context.

1 Introduction

Speakers use their whole bodies to present their ideas. Utterance (1), drawn from a lecture about speech errors,[1] shows how speakers can combine speech and gesture to flesh out their arguments in visible form.

(1) There are these very low level phonological errors that tend not to get reported.
The right hand is held in a fist and positioned below the mouth, where the previous gesture was performed; the hand iteratively moves in the sagittal plane (i.e., vertically outwards) in clockwise circles (as viewed from left).

In context, the gesture seems to visualise the continuous processes, operating below the level of awareness, that give rise to unreported errors.

Descriptive work on such gestures makes three key observations, which any theoretical account must respect. First, speech and gesture combine to express a single thought. Their contents fit together, forming the speaker's overall message (McNeill, 1992; Kendon, 2004). For example, in (1) the gesture visualises the subconscious nature of processes that cause low-level phonological errors, thereby explaining why they don't get reported.

Second, these gestures take a form that directly or metaphorically depicts what is described (McNeill, 1992; Kopp et al., 2004). For example, the iterative movement in (1) is a metaphorical depiction of a continuous process. However, not all aspects of a gesture have to be meaningful; e.g., the clockwise direction of motion in (1) doesn't contribute to interpretation.

Third, apart from conventionalised gestures (e.g., "thumbs up"), the form of a gesture on its own is insufficient for a coherent interpretation. For example, the gesture in (1) would be uninterpretable without simultaneous speech. A specific and coherent interpretation of gesture arises by linking it to simultaneous speech, and so it changes meaning in different speech contexts:

(2) The mouse ran on the wheel for a few minutes.
Gesture as in (1)

In (2), the gesture is still iconic: the physical movement of the hand depicts the path of the wheel's motion. But its interpretation is different from that in (1), and in particular the direction of motion now carries important information whereas it didn't in (1). A further kind of context-dependence arises through spatial distinctions maintained across multiple gestures (Emmorey et al., 2000). In (1), we recognise that the processes depicted are low-level in part by linking

[1]http://www.talkbank.org/media/Class/Lecture-unlinked/feb02/feb02-8.mov

the gesture here to earlier gestures which have depicted the production of *noteworthy* errors through a trajectory leading from the mouth *upward*.

This paper describes a formal analysis of gesture that respects these three principles. In formalising these principles, we go beyond previous work—whether descriptive (McNeill, 1992; Kendon, 2004), psychological (Lozano and Tversky, 2004), or applied to embodied agents (Cassell, 2001; Kopp et al., 2004)—by drawing on formal models of semantics and pragmatics in discourse interpretation. Specifically, we argue in Section 2 that *rhetorical relations* provide a theoretical construct to explicate how speech and gesture cohere into a single thought. We explain in Section 3 how *underspecified representations of meaning* let us specify both how the form of gesture constrains its content and how the resulting representation needs to be augmented by contextual information to obtain a coherent logical form (LF). In Section 4 we represent LFs with *dynamic semantics* to capture the evolving structure of objects and spatial relationships that inform gesture interpretation. And in Section 5, this formal apparatus allows us to model how gesture is interpreted by drawing on its mappings from form to (underspecified) meaning, a context of salient objects and relationships, and rhetorical connections to synchronous speech.

While the resulting architecture captures descriptive insights into gesture, it in fact instantiates a general end-to-end model of pragmatic interpretation (Asher and Lascarides, 2003). We believe that these same principles apply to the interpretation of all communication—in whatever medium it takes place.

2 Relating gesture to speech

For Asher and Lascarides (2003), rhetorical relations are kinds of speech acts. That is, they offer an inventory of things that a speaker might be doing by providing content in discourse: he might be elaborating it, explaining it, continuing a narration, drawing a contrast, and so forth. When hearers infer rhetorical relations, they recognise the speaker's communicative intention and so discover why the discourse is coherent.

We propose that gesture is rhetorically related to simultaneous speech. For example, the gesture in (1) can be understood as providing an *explanation* in support of what is being said. The gesture in (2) can be understood as an *elaboration* that complements what is being said. On our view, the rhetorical connection is a tool which lets us formalise the intuition that the gesture is a communicative action which plays a part in the speaker's overall intention: rhetorical connections knit gesture and speech into a single thought.

Rhetorical relations are a vehicle for predicting implicatures, because their semantic consequences go beyond the compositional semantics of the utterances (and gestures) they connect, and inferring rhetorical relations during discourse interpretation involves commonsense reasoning with compositional semantics and contextual information such as world knowledge. Rhetorical relations also create a hierarchical structure to the discourse, where some communicative actions are completed and others remain open. This structure thus constrains the alternative ways coherent discourse can progress. The theory of rhetorical relations therefore serves to operationalise Grice's (1975) theory of communication as rational behaviour, articulating a precise interface between compositional semantics and pragmatics.

In essence, inferring rhetorical connections and inferring a gesture's specific meaning are logically co-dependent tasks. For example, interpreting the gesture in (1) as a continuous subconscious process causing speech errors supports an inference that the gesture and speech are related with *explanation*. This inference is justified partly by the semantics of *explanation* and partly by world knowledge: errors won't get reported if they aren't perceived; and the effects of continuous subconscious processes are normally hard to perceive.

Note that this specific content is compatible with the gesture's underspecified meaning as revealed by its form: as we shall see in Section 3, the fist can be interpreted as depicting the phonological errors being caused by something; the iterative, continuous motion of the hand can be interpreted as conveying that this cause is iterative and continuous; and the relatively low position of the hand can be interpreted as conveying that it is 'low down' or subconscious. However, the *explanation* relation predicts that the clockwise motion does not depict anything in this context.

There may be alternative specific interpretations of the gesture in (1), which in turn support inferences to alternative rhetorical connections, but as Asher and Lascarides (2003) argue, discourse in-

terpretation is governed by a general principle of maximising coherence: one interprets discourse so that the highest possible quality of rhetorical connections is achieved (see Section 5 for further details). Of course, calculating a preferred interpretation using this principle does require formalising all the commonsense background involved.

Rhetorical relations thus help to model how context yields a more specific interpretation of the gesture from its underspecified meaning as revealed by its form. The remainder of this paper puts the case in formal terms. Now, generalising from Asher and Lascarides (2003), we would also expect that rhetorical relations can help to characterise the interpretation of speech and gesture in other ways—such as predicting when the interpretation of a gesture is coherent and when it is not in a way that other pragmatic knowledge sources, such as world knowledge, cannot do on their own; or modelling how a gesture can resolve ambiguities in synchronous speech. We leave these suggestions to future work.

3 Underspecifying iconic meaning

Underspecification is a common representational approach to interface an abstract linguistic meaning to its specific, contextualised interpretation e.g., (Alshawi and Crouch, 1992; Reyle, 1993). The contextualised interpretation is represented as a logical formula in a standard formal language; this plays the role of an LF in the model. (We will combine rhetorical relations with dynamic semantics to represent LFs; see Section 4.) The grammar, however, does not explicitly construct the LF. Instead, it builds a *partial description* of it, leaving open multiple alternatives. In this sense, the exact interpretation is left underspecified by compositional semantics. Accordingly, the underspecified elements in the description must be *resolved* pragmatically in interpretation.

We adopt *Robust minimal recursion semantics* (RMRS) as a formalism for underspecified semantic representation (Copestake, 2003). Like many formalisms, RMRS can underspecify semantic scope. In addition, it can represent partial information about *which predicates* appear in LF, *what arity* they have, and *what sorts of arguments* they take, a flexibility that isn't fully supported by other formalisms (e.g., Asher and Lascarides (2003) do not underspecify arity). We show that the form of iconic gesture constrains, but does not determine, all these aspects of interpretation.

Following earlier work (McNeill, 1992; Kopp et al., 2004), we characterise the link between the form and iconic meaning of gesture by representing gesture form in a multidimensional matrix. The rows in this matrix describe aspects of a gesture's form which potentially reveal things about its meaning—the hand shape, the orientations of the palm and finger, the position of the hands relative to the speaker's torso, the paths of the hands and the direction in which the hands move along those paths. For example, we represent the gesture form of (1) as the feature structure in (3).

(3)
$$\begin{bmatrix} \text{hand-shape} : \textit{asl-s} \\ \text{finger-direction} : \textit{down} \\ \text{palm-direction} : \textit{left} \\ \text{trajectory} : \textit{sagittal-circle} \\ \text{movement-direction} : \{\textit{iterative}, \textit{clockwise}\} \\ \text{location} : \textit{central-right} \end{bmatrix}$$

Here each of the six attributes takes a particular value which characterises the physical realisation of the gesture. The matrix formalism highlights that the gesture morphology does not yield a hierarchical structure; rather, elements of the description combine via unification or 'conjunction'.

The gesture's iconicity consists in the fact that each of these attribute-value elements may convey a specific, analogous piece of content. With RMRS, we can formalise this in two straightforward steps. First, to each attribute-value element, we associate an *underspecified abstract predication* that must be resolved to a particular formula in the logical form of gesture. We introduce a convention that reads this underspecified predication directly off the feature structure, as in (4):

(4) h_1:*hand_shape_asl-s*(i_1)

Here h_1 is a uniquely indexed label that underspecifies the scope of the predication; i_1 is a uniquely indexed *metavariable* that underspecifies the main argument of the predication (an object, eventuality, etc); and *hand_shape_asl-s* underspecifies the property of i_1 that's depicted through the gesture's fist-shape. The compositional meaning of a gesture is just the conjunction of the underspecified predications associated with each of its form features. These predications must be resolved to give the gesture a specific interpretation.

Second, we constrain the possible resolutions of the underspecified predicates to a restricted inventory that states what alternative qualities we can depict with aspects of the gesture's form.

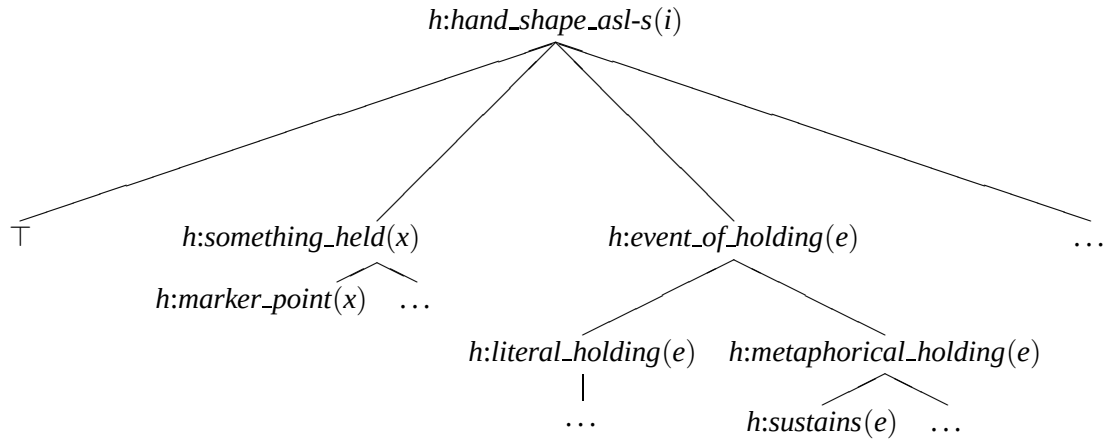

Figure 1: Part of the hierarchy of underspecified and fully-specified predications for *hand_shape_asl-s*.

We expect that each underspecified predicate admits a hierarchy of increasingly specified resolutions, as in Figure 1. While some of the leaves in this hierarchy correspond to fully specific interpretations, the creative use of metaphor makes interpretation open-ended. Therefore, some of the hierarchy's leaves correspond to more vague interpretations, and we envisage that either the speaker and hearer sometimes settle on a coherent but vague interpretation, or additional logical axioms will resolve a vague interpretation to a more specific one in the particular discourse context. To capture the (metaphorical) contribution of the fist in (1), we resolve *hand_shape_asl-s* as depicting a holding event, metaphorically interpreted as the event *e* of a process *x sustaining* speech errors *y* ("bearing them with it", as it were). At the same time, we can capture the contribution of the fist to the depiction of (2) by resolving *hand_shape_asl-s* as depicting something held, in particular a *marker-point x* indicating a designated location on the mouse's spinning wheel. Finally, all underspecified predications are resolvable to $\top$—the valid formula—since they might not contribute meaning in context. Underspecified predicates may also share specific resolutions: e.g., *marker-point* is also one way of resolving the underspecified predicate corresponding to a flat hand, and thus the gesture in (2) could have been performed with a flat hand instead of a fist.

Crucially, Figure 1 reflects the fact that, like all dimensions of iconic gesture, the fist shape doesn't determine how many entities are involved in the specific semantic relation it resolves to. The specific predications in Figure 1 vary in the number of arguments they take, and the factorised notation of RMRS lets us express this. In RMRS, additional arguments to predicates, over and above the 'primary' one, are expressed as separate binary relations: e.g., *sustains* is a 3-place predicate and h:*sustains*(e,x,y) is a notational variant of h:*sustains*(e), *ARG1*(h,x), *ARG2*(h,y), while *marker-point* is a 1-place predicate, and therefore h:*marker-point*(x), *ARG1*(h,y) is unsatisfiable. One can also underspecify the position of a variable in a predication: the binary relation *ARGn*(h,x) means that x is an argument to the predicate labelled by h, but its argument position is unknown (so *ARG1* $\sqsubset$ *ARGn*).

The divergent resolutions of the same gesture in different contexts highlight how we capture insights from previous work: we represent gesture meaning compositionally and iconically, yet in an underspecified form that requires context to resolve. You can compare predications like *hand_shape_asl-s* to Kopp et al.'s (2004) *image description features*, an abstract representation, distinct from form and content, that captures gesture meaning. By using RMRS, we can reinterpret these representations as analogous, both formally and substantively, to existing underspecified semantic representations for linguistic items. In particular, we show in Section 5 that we can therefore build reasoning mechanisms that combine information from speech and gesture to derive a single, overall coherent resolution of the logical form of discourse.

4 Representing meaning in context

In portraying objects and relationships, gesture exploits not just the iconic meaning of physical actions, but also the evolving discourse context.

For example, gesture, like speech, has access to the salient objects that have been evoked by noun phrases in the previous discourse. However, one striking difference between gesture and speech is that gesture is profoundly limited in its ability to introduce new entities into the context. We adapt the formalism of *segmented discourse representation structures* (SDRS) (Asher and Lascarides, 2003) to precisely model these similarities and differences between gesture and speech. An SDRS specifies a collection of update expressions which partially describe the evolution of context during the discourse. The SDRS also links these updates together using rhetorical relations to further constrain the interpretation and structure of discourse. We focus here on the updates themselves.

Individuals which are introduced in gesture seem to be subject to similar constraints on acceptability as *definite descriptions* in language: in both cases, the entities so-introduced must be related to an available antecedent through one of a constrained set of semantic relationships—including equality, in which case the entity is coreferent with its antecedent. We call these *bridging* relations, after Clark (1977). For instance, we infer in (2) that the *marker-point* represented by the fist indicates a *part of* the wheel. Thus there is a bridging relationship *part-of* between the gestural depiction and the noun phrase *the wheel* in the utterance.

The form of the gesture doesn't determine the bridging relation nor the antecedent, just as the form of definite descriptions doesn't. And so the form of gesture (and of definite descriptions) must impose the constraint that there is such a bridging relation, but underspecify its value. We follow Asher and Lascarides' (2003) realisation of this. For the sake of simplicity, we simply mention the notation and gloss its interpretation in words: $R(x,y) \wedge R =? \wedge x =?$ means that y is related to an (available) individual x with a relation R, but the *values* of x and R are underspecified. Following Chierchia's (1995) compositional semantics of definite descriptions, we include bridging constraints in the LF of gesture. These can be added to the RMRS produced by the grammar. We assume this addition occurs *outside* the grammar because bridging relations don't affect semantic composition from syntax. Rather, they impose constraints on the process of constructing the LF of discourse, stipulating that a particular relation to a particular available antecedent must be found for each individual variable.

At the same time, use of a gesture changes the referents available to subsequent discourse. This is the bread-and-butter of dynamic semantics—see e.g. Groenendijk and Stokhof (1991)—and we handle it in the usual way. We interpret formulae as transitions that update an input context to yield an output context. Among other things, these changing contexts make explicit what referents are available. However, an object introduced in gesture, like the point on the wheel in (2), can appear in subsequent gestural figurations, but cannot be picked up by a pronoun in subsequent speech. So we follow Asher and McCready (2006) in structuring our contexts to distinguish kinds of reference: we have one set of referents f available to speech and another set g (a superset in fact) available to gesture—see the dynamic semantic definition of indefinite quantification in (5a). Correspondingly, we annotate LFs for speech and gesture to indicate which kind of reference they participate in. That is, we introduce a 'gesture' modality $[\mathcal{G}]$, and the dynamic semantics of $[\mathcal{G}]\phi$ ensures that ϕ updates only the set g of referents available to *gesture*; see (5b):

(5) a. $\langle f,g\rangle [\![\exists x]\!]^M \langle f',g'\rangle$ iff
$dom(f') = dom(f) \cup \{x\}$ and
$\forall y \in dom(f),\ f'(y) = f(y)$
(i.e., $f \subseteq_x f'$), $g \subseteq_x g'$, and $f'(x) = g'(x)$.

b. $\langle f,g\rangle [\![[\mathcal{G}]\phi]\!]^M \langle f',g'\rangle$ iff $f = f'$ and $\langle g,g\rangle [\![\phi]\!]^M \langle g',g'\rangle$

One of the most interesting kinds of context dependence is the way successive gestures can establish a common frame of reference for spatial depiction (Emmorey et al., 2000; Kopp et al., 2004). We believe that dynamic semantics will provide an attractive formal setting in which to capture such connections precisely, since dynamic semantics has already proved an effective tool for modelling the evolving perspective in discourse—in time, space and information (Bittner, 2006). However, a model of spatial context in gesture will need substantial formal development, requiring a suitable formal ontology of space, a corresponding characterisation of spatial context, and rules for interpreting gesture meaning in terms of this spatial context. We leave this for the future, and here limit ourselves to the formalism sketched so far, which we can more immediately carry over from Asher and Lascarides (2003) and which in fact suffices to account for examples (1) and (2).

5 Interpreting gesture

We now address the problem of how the underspecified semantics revealed by form gets resolved to fully specific meanings in context. In Asher and Lascarides' (2003) SDRT model, this occurs as a byproduct of *discourse update*: the process of constructing the logical form of discourse.

Discourse update in SDRT starts from the compositional semantics derived from the grammar. To handle situated language, we work with the semantics for gesture derived from its form by iconicity. The compositional semantics of the gesture in (1) and (2) is shown in (6).

(6) h_g:$[\mathcal{G}](h)$,
$h \geq h_j$, for $1 \leq j \leq 6$,
h_1:*hand_shape_asl-s*(i_1),
h_2:*finger_dir_down*(i_2),
h_3:*palm_dir_left*(i_3),
h_4:*traj_sagittal_circle*(i_4),
h_5:*move_dir_iterative*(i_5),
h_5:*move_dir_clockwise*(i_5),
h_6:*loc_central-right*(i_6)

In outline, this formula says that the final meaning will contain an expression h_g giving information specified through gesture, and that this information will resolve how the hand shape, finger direction, path, trajectory, direction of motion and location of the gesture (as labelled by $h_1 \ldots h_6$) work to describe salient generalised individuals (as labelled by $i_1 \ldots i_6$) from the context. Observe that the modality $[\mathcal{G}]$ outscopes the predications labelled h_1 to h_6, as required by the dynamic semantics in (5) of any of its resolved forms.

We assume, following Kopp et al. (2004), that gesture combines with its synchronous speech *within the grammar*, producing a single derivation tree. This assumption is necessary both to predict the fine-grained temporal synchrony between speech and gesture, and to capture the distinctive constraints on coreference and other semantic relations that apply to units of speech and gesture in coordination (e.g., a gesture and its synchronous speech cannot be combined with disjunction). Here the grammar yields the predication h:*iconic_rel*(h_s, h_g), where h_s labels the content of the speech. This predication underspecifies the rhetorical connection between the gesture and speech and must resolve to a value that's licensed by iconic gesture: e.g., *Explanation* or *Elaboration*, but not *Contrast* or *Disjunction*.

Discourse update derives an LF through commonsense reasoning, drawing on non-linguistic information, such as world knowledge, as well as compositional semantics. This reasoning is formalised using nonmonotonic inference rules that predict possible rhetorical connections from (shallow) representations of linguistic meaning and non-linguistic information. We refer collectively to this system as the *glue logic*. Its rules have the following form, where $A > B$ can be read as *If A then normally B*, and the symbols α and β are metavariables ranging over the labels of discourse segments in the SDRS representation:

$(\lambda{:}?(\alpha,\beta) \wedge \varphi) > \lambda{:}R(\alpha,\beta)$
(`Glue Logic Schema`)

In words: if the segment labelled β is to be connected to the segment labelled α with a rhetorical relation, and the result is to appear as part of the logical scope labelled λ, but we don't know the value of this relation yet, and moreover φ holds of the content labelled by λ, α and β, then normally the rhetorical relation is R. The conjunct φ is cashed out in terms of the (underspecified) LFs that α and β label, and the rules are justified either on the basis of underlying linguistic knowledge, world knowledge, or knowledge of the cognitive states of the conversational participants. Thus glue logic axioms encapsulate *prima facie* default inferences about which type of speech act was performed, on the basis of the content and context of the utterances.

In SDRT the inferences can flow in one of several directions. For example, if the premises of a glue logic axiom is satisfied by the information already available (e.g., by the underspecified semantics derived from the grammar), then one can infer a particular rhetorical relation and from its semantics infer how the underspecified conditions of the utterance or gesture are resolved. Alternatively, there are cases where the premises for inferring rhetorical relations are not satisfied by the underspecified compositional semantics. In this case, one can resolve the underspecified content so as to support an inference to a rhetorical relation. If one adopts this strategy, and moreover there is a choice of which way to resolve the underspecified content so as to infer a rhetorical relation from it, then one chooses an interpretation which maximises the *quality* of the rhetorical relations one can infer from it (see Asher and Lascarides (2003) for details).

Here, we indicate how discourse update can resolve the underspecified meaning of gesture with speech. Let's start with the analysis of the situated utterance (2). We introduce a glue logic axiom which captures the following intuition: if two propositions are rhetorically related somehow, and they both describe a movement event with the same participant and which can occur simultaneously, then there is evidence in the discourse that these events are in a subtype relation (following Asher and Lascarides (2003), we assume a notation where e_α and e_β are respectively the semantic indices of α and β):

(7) $(\lambda{:}?(\alpha,\beta) \wedge h{:}\mathit{movement}(e_\alpha) \wedge \mathit{ARGn}(h,x) \wedge h'{:}\mathit{movement}(e_\beta) \wedge \mathit{ARGn}(h',x) \wedge \mathit{temporally\text{-}compatible}(e_\alpha,e_\beta)) \rightarrow \mathit{Subtype}_D(\beta,\alpha)$

The predication $\mathit{Subtype}_D(\beta,\alpha)$ does not entail that β and α are *actually* in a subtype relation; only that there is evidence in the discourse that they are. Note that the rule is monotonic, because either the evidence is present in the discourse, or it's not. This predicate is used to infer *Elaboration*:

(8) $(\lambda{:}?(\alpha,\beta) \wedge \mathit{Subtype}_D(\beta,\alpha)) > \lambda{:}\mathit{Elaboration}(\alpha,\beta)$

If $\mathit{Elaboration}(\alpha,\beta)$ is inferred, then an *actual* subtype relation among their events follows.

Now returning to the situated utterance (2), the grammar imposes a constraint that the contents of speech and gesture are rhetorically connected by one of the relations that's licensed for gesture (as encapsulated in *iconic_rel*). So for (2) to be coherent, one must infer a particular rhetorical relation between them and also infer specific interpretations that support this relation.

In (2), the underspecified content on its own is insufficient for inferring a rhetorical relation, for although the gesture depicts movement, some of its possible specific interpretations do not entail physical movement (e.g., the movement could be metaphorical, or indeed the movement could resolve to $\top$ as explained in Section 3). Nor does the gesture's form specify the movement's participants. However, one of the possible resolved meanings of the gesture is one which satisfies the axiom (7). This is because one can resolve e_β (i.e., the semantic index of the gesture) to be the movement of the wheel in a circular, iterative clockwise direction, where the wheel is also the location of the running described in the sentence. This possible interpretation of the gesture is supported by world knowledge, which stipulates that when a marker point on a rigid object moves then so does that object. Moreover, world knowledge suggests that the moving object that's depicted cannot be the mouse, since the mouse runs on the spot. Thus with this specific interpretation of the gesture, the antecedent to (7) is satisfied by the content of the utterance and the gesture, with *x* in this axiom instantiated by the wheel. If the gesture is interpreted this way, then the axioms (7) and (8) lead to a (nonmonotonic) inference that the utterance and gesture are related with *Elaboration*. Suppose that this is the *only* possible resolved interpretation of the gesture that leads to an inference about which rhetorical relation connects the utterance and the gesture. Then discourse update in SDRT forces this specific interpretation (see (Asher and Lascarides, 2003) for formal details). Thus discourse update resolves the hand shape to $\mathit{marker\text{-}point}(y)$ and the accompanying bridging relation $\mathit{part\text{-}of}(y,x) \wedge \mathit{wheel}(x)$, where *x* is co-referent with the wheel denoted in (2); it resolves the underspecified predicate $\mathit{traj\_sagittal\_circle}(i)$ to $\mathit{move}(e_\beta,x) \wedge \mathit{path}(e_\beta,z) \wedge \mathit{sagittal\_circle}(z)$, and it resolves the underspecified predications $\mathit{move\text{-}dir\_iterative}(j)$ and $\mathit{move\text{-}dir\_clockwise}(j)$ to $\mathit{direction}(e_\beta,w) \wedge \mathit{iterative}(w) \wedge \mathit{clockwise}(w)$. Thus the gesture provides more information about the movement described in the utterance: the wheel is in a vertical plane (and fixed at a central point), and moves in a clockwise direction several times.

The analysis of (1) is similar to that of (2).

(1) There are these very low level phonological errors that tend not to get reported.

However, the specific interpretation of the gesture in (1) cannot satisfy the axiom (7) this time, because the sentence is not about physical movement. So another specific interpretation is needed to support a particular rhetorical connection between the speech and gesture. As we explained in Sections 2 and 3, the underspecified content of the gesture can resolve to denote a continuous, subconscious process which causes the phonological errors mentioned in (1). This particular interpretation satisfies the antecedent of an axiom whose consequent is $\mathit{Cause}_D(\beta,\alpha)$—i.e., there is evidence in the discourse of a causal relation. This in turn supports a default inference to *Explanation*:

(9) $(\lambda{:}?(\alpha,\beta) \wedge \mathit{Cause}_D(\beta,\alpha)) >$
$\lambda{:}\mathit{Explanation}(\alpha,\beta)$

If this is the specific interpretation which maximises the quality of the connection between the constituents, then discourse update dictates that the logical form of the discourse resolves the interpretations this way.

6 Conclusion and future work

We have provided a formal semantic analysis of iconic gesture which captures several compelling features that are described in the literature. First, it predicts that iconic gesture on its own doesn't receive a coherent interpretation: this is achieved by assigning a very underspecified content to iconic gesture as revealed by its form. Second, it predicts that speech and gesture together form a 'single thought'. This is achieved by integrating the content of gesture and synchronous speech in the grammar, and ensuring that their denotations are semantically related. The model then demands that one must compute the value of this rhetorical relation, using compositional semantics and contextual information as clues. Reasoning about this rhetorical connection leads to the gesture's underspecified content being resolved to a specific interpretation. Finally, we exploited discourse structure and the dynamics in dynamic semantics to account for dependencies on co-reference across speech and gesture and among different gestures in the discourse.

One virtue in our analysis is to demonstrate that existing mechanisms for representing the content of language can be exploited to model gesture as well. However, much future work needs to be done. For example, we need to specify in more detail the construction rules in the grammar which combine speech and gesture, and the meaning postulates which convey the range of possible meanings that the various dimensions of iconic gesture can depict. Concretely, that requires us to specify a hierarchy as in Figure 1 more fully, and to link the hierarchy to a family of interpretive instances of the Glue Logic Schema so as to predict a wide range of natural interpretations. In the dynamic semantic component, we need to integrate the interpretation of gesture with a commonsense view of space. We would also like to explore in more detail how a gesture's interpretation is constrained by prior gestures, as well as speech, and extend the analysis to other types of gesture, such as deixis and beats.

Acknowledgments

Thanks to Paul Tepper and anonymous reviewers for comments. Supported in part by the Leverhulme Trust, Rutgers University and NSF HLC0308121.

References

H. Alshawi and R. Crouch. 1992. Monotonic semantic interpretation. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 32–39, Delaware.

N. Asher and A. Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

N. Asher and E. McCready. 2006. Were, would, might and a compositional account of counterfactuals. technical report; material will be presented at ESSLLI 2006.

M. Bittner. 2006. Online update: temporal, modal and *de se* anaphora in polysynthetic discourse. In C. Barker and P. Jacobson, editors, *Direct Compositionality*. Oxford.

J. Cassell. 2001. Embodied conversational agents: Representation and intelligence in user interface. *AI Magazine*, 22(3):67–83.

G. Chierchia. 1995. *Dynamics of Meaning: Anaphora, Presupposition and the Theory of Grammar*. University of Chicago Press, Chicago.

H. Clark. 1977. Bridging. In P. N. Johnson-Laird and P. C. Wason, editors, *Thinking: Readings in Cognitive Science*, pages 411–420. Cambridge University Press.

A. Copestake. 2003. Report on the design of rmrs. Technical Report EU Deliverable for Project number IST-2001-37836, WP1a, Computer Laboratory, University of Cambridge.

K. B. Emmorey, B. Tversky, and H. Taylor. 2000. Using space to describe space: Perspective in speech, sign and gesture. *Spatial Cognition and Computation*, 2(3):157–180.

H. P. Grice. 1975. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Synax and Semantics Volume 3: Speech Acts*, pages 41–58. Academic Press.

J. Groenendijk and M. Stokhof. 1991. Dynamic predicate logic. *Linguistics and Philosophy*, 14:39–100.

A. Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge.

S. Kopp, P. Tepper, and J. Cassell. 2004. Towards integrated microplanning of language and iconic gesture for multimodal output. In *Proceedings of ICMI*.

S. C. Lozano and B. Tversky. 2004. Communicative gestures benefit communicators. In *Proceedings of Cognitive Science*.

D. McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago.

U. Reyle. 1993. Dealing with ambiguities by underspecification: Construction, interpretation and deduction. *Journal of Semantics*, 10:123–179.