A Formal Semantic Analysis of Gesture

Alex Lascarides
School of Informatics,
University of Edinburgh

Matthew Stone
Computer Science,
Rutgers University

Abstract

The gestures that speakers use in tandem with speech include not only conventionalised actions with identifiable meanings (so called *narrow gloss gestures* or *emblems*) but also productive iconic and deictic gestures whose form and meanings seem largely improvised in context. In this paper, we bridge the descriptive tradition with formal models of reference and discourse structure so as to articulate an approach to the interpretation of these productive gestures. Our model captures gestures' partial and incomplete meanings as derived from form, and accounts for the more specific interpretations they derive in context. Our work emphasises the commonality of the pragmatic mechanisms for interpreting both language and gesture, and the place of formal methods in discovering the principles and knowledge that those mechanisms rely on.

1 Introduction

Face-to-face dialogue is the primary setting for language use, and there is increasing evidence that theories of semantics and pragmatics are best formulated directly for dialogue. For example, many accounts of semantic content see extended patterns of interaction rather than individual sentences as primary (Kamp, 1981, Asher and Lascarides, 2003, Ginzburg and Cooper, 2004, Cumming, 2007). Likewise, many pragmatic theories derive their principles from cognitive models of interlocutors who must coordinate their interactions while advancing their own interests (Lewis, 1969, Grice, 1975, Clark, 1996, Asher and Lascarides, 2003, Benz, Jäger and van Rooij, 2005). But face-to-face dialogue is not just words. Speakers can use facial expressions, eye gaze, hand and arm movements and body posture intentionally to convey meaning; see e.g. (McNeill, 1992). This raises the challenge of fitting a much broader range of behaviours into formal semantic and pragmatic models. We take up this challenge in this paper.

We focus on a broad class of *improvised, coverbal, communicative actions*, which seem both particularly important and particularly challenging for models of meaning in face-to-face dialogue. We distinguish *communicative actions* from other behaviours that people do in conversation, such as practical actions and incidental 'nervous' movements, following a long descriptive tradition (Goffman, 1963, Ekman and Friesen, 1969, Kendon, 2004). This allows us to focus on a core set of behaviours—which we call *gestures* following Kendon 2004—that untrained viewers are sensitive to (Kendon, 1978), that linguists can reliably annotate (Carletta, 2007), and that any interpretive theory must account for. Gestures have what Kendon calls "features of manifest deliberate expressiveness" (Kendon, 2004, p. 15), including the kinematic profile of the movement, as an excursion from and back to a rest position; its dy-

namics, including pronounced onset and release; and the attention and treatment interlocutors afford it.

Coverbal gestures are those that are performed in synchrony with simultaneous speech. Gestures can also be performed without speech (see (Kendon, 2004, Ch. 14) for examples), in the pauses between spoken phrases (see (Engle, 2000, Ch. 3) for examples), or over extended spans that include both speech and silence (see Oviatt, DeAngeli and Kuhn (1997) for examples). Coverbal gestures show a fine-grained alignment with the prosodic structure of speech (Kendon (1972); (Kendon, 2004, Ch. 7)). The gesture typically begins with a preparatory phase where the agent moves the hands into position for the gesture. It continues with a stroke (which can involve motion or not), which is that part of the gesture that is designed to convey meaning—we focus in this paper on interpreting strokes. Finally, it can conclude with a post-stroke phase where the hands retract to rest. Speakers coordinate gestures with speech so that the phases of gesture performance align with intonational phrases in speech and so that strokes in particular are performed in time with nuclear accents in speech. This coordination may involve brief pauses in one or the other modality, orchestrated to maintain synchrony between temporally extended behaviours (Kendon, 2004, Ch. 7). The active alignment between speech and gesture is indicative of the close semantic and pragmatic relationship between them.

Finally, we contrast *improvised* gestures both with other gestures whose content is emblematic and completely conventionalised, such as the 'thumbs up' gesture, and with *beat* gestures, which merely emphasise important moments in the delivery of an utterance. Improvised gestures may involve deixis, where an agent designates a real, virtual or abstract object, and iconicity, where the gesture's form or manner of execution mimics its content. Deixis and iconicity sometimes involve the creative introduction of correspondences between the body and depicted space. Nevertheless, as Kendon's (2004) fieldwork shows, even in deixis and iconic representation, speakers recruit specific features of form consistently to convey specific kinds of content. The partial conventionalisation involved in these correspondences is revealed not only in consistent patterns of use by individual speakers but also in cross-cultural differences in gesture form and meaning.

Researchers have long argued that speakers use language and gesture as an integrated ensemble to negotiate a single contribution to conversation—to "express a single thought" (McNeill, 1992, Engle, 2000, Kendon, 2004). We begin with a collection of attested examples which lets us develop this idea precisely (Section 2). We show that characterising the interpretation of such examples demands a fine-grained semantic and pragmatic representation, which must encompass content from language and gesture and formalise scope relationships, speech acts, and contextually-inferred referential connections. Our approach adopts representations that use *dynamic semantics* to capture the evolving structure of salient objects and spatial relationships in the discourse and a *segmented structure* organised by *rhetorical connections* to characterise the semantic and pragmatic connections between gesture and its communicative context. We motivate and describe these logical forms in Section 3.

We then follow up our earlier programmatic suggestion (Lascarides and Stone, in press), that such logical forms should be derivable from underspecified semantic representations that capture constraints on meaning imposed by linguistic and gestural form, via constrained inference which reconstructs how language and gesture are rhetorically connected. Our underspecified semantic representations, described in Section 4, capture the incompleteness of meaning that's revealed by gestural form while also capturing, very abstractly, what a gesture must convey given its particular pattern of shape and movement. We describe the resolution

from underspecified meaning to specific interpretation in Section 5: a *glue logic* composes the logical form of discourse from underspecified semantic representations via default rules for inferring rhetorical connections; and as a byproduct of this reasoning underspecified aspects of meaning are disambiguated to pragmatically preferred, specific values. These formal resources parallel those required for recognising how sentences in spoken discourse are coherent, as developed by Asher and Lascarides (2003) *inter alia*.

The distinctive contribution of our work, then, is to meet the challenge, implicit in descriptive work on non-verbal communication, of handling gesture within a framework that's continuous with and complementary to purely linguistic theories. This is both a theoretical and methodological contribution. In formalising the principles of coherence that guide the interpretation of gesture, we go beyond previous work—whether descriptive (McNeill, 1992, Kendon, 2004), psychological (So, Kita and Goldin-Meadow, in press, Goldin-Meadow, 2003), or applied to embodied agents (Cassell, 2001, Kopp, Tepper and Cassell, 2004). Such a logically precise model is crucial to substantiating the theoretical claim that speech and gesture convey an integrated message.

Such a model is also crucial to inform future empirical research. The new data afforded by gesture calls for refinements to theories of purely linguistic discourse, potentially resulting in more general and deeper models of semantics and pragmatics. But a formal model is often indispensable for formulating precise hypotheses to guide empirical work. For instance, our framework raises for the first time a set of logically precise constraints characterising reference and content across speech and gesture and sequences of gesture in embodied discourse. Testing these constraints empirically will have a direct influence not only on the development of formal theory but on our understanding of the fundamental pragmatic principles underlying multimodal communication. A hybrid research methodology, combining empirical research and logically precise models to mutual benefit, has proved highly successful in analysing language. We hope the same will be true for analysing gesture.

2 Dimensions of Gesture Meaning in Interaction with Speech

We begin with an overview of the possible interpretations of improvised coverbal gestures. We emphasise that the precise reference and content of these gestures is typically indeterminate, so that multiple consistent interpretations are often available. It is the details and commonalities of these alternative interpretations that we aim to explain. We argue that they reveal three key generalisations about gesture and its relationship to speech.

1. Gestures can depict the referents of expressions in the synchronous speech, inferentially related individuals, or salient individuals from the prior context.

2. Gestures can show what speech describes, or they can complement speech with distinct but semantically related information.

3. Gesture and speech combine into integrated overarching speech acts with a uniform force in the dialogue and consistent assignments of scope relationships.

These principles—which we defend in more detail in Lascarides and Stone (2006, in press)—underpin the formalism we present in the rest of the paper.

Our discussion follows McNeill (2005, p. 41) in characterising *deixis* and *iconicity* as *two dimensions* of gesture meaning, rather than two kinds of gesture. Deixis is that dimension of

gesture meaning that locates objects and events with respect to a consistent spatial reference frame; our first examples highlight the semantic interaction of this spatial reference with the words that accompany coverbal gestures. Iconicity, meanwhile, is the dimension of gesture meaning which depicts aspects of form and motion in a described situation through a natural correspondence with the form and motion of the gesture itself; we consider iconicity in relatively 'pure' examples later in this section. So characterised, of course, deixis and iconicity are not mutually exclusive, so our formalism must allow us to regiment and combine the deictic and iconic contributions to gesture interpretation.

We start with utterance (1), which is taken from Kopp et al.'s (2004) corpus of face-to-face direction-giving dialogues and visualised in Figure 1 (in this paper, we use square brackets to indicate the temporal alignment of speech and gesture, and where relevant smallcaps to mark pitch accents):[1]

(1) And [Norris]$_1$ [is exactly across from the library.]$_2$
First: The left arm is extended into the left periphery; the left palm faces right so that it and fingers are aligned with the forearm in a flat, open shape. Meanwhile, the right hand is also held flat, in line with the forearm; the arm is held forward, with elbow high and bent, so that the fingers are directly in front of the shoulder.
Second: The left hand remains in its position while the right hand is extended to the extreme upper right periphery.

"And Norris..." *"...is exactly across from the library."*

Figure 1: Hand gestures place landmarks on a virtual map.

The utterance concludes a direction-giving episode in which the speaker has already used gestures to create a virtual map of the Northwestern university campus (more of this episode appears later as examples (10) and (11)). Throughout the utterance, the speaker's left hand is positioned to mark the most salient landmark of this episode—the large lagoon at the centre of the campus, which she has introduced in previous speech and gesture. The gesture on the left of Figure 1 positions the right hand at a location that was initially established for Norris Hall, while the next gesture moves the right hand to resume an earlier demonstration of the library. One could interpret these gestures as demonstrating the buildings, their locations, or even the eventualities mentioned in the clause. But any of these interpretations result in a multimodal communicative act where the right hand indexes entities referenced in the accompanying speech and indicates the spatial relationship *have* that the sentence describes. To capture the deictic dimension of gesture form and meaning here, we characterise the form

[1] Video is available at homepages.inf.ed.ac.uk/alex/Gesture/norris-eg.mov.

of the gesture as targeting a specific region of space, and then use a referent to that region of space to characterise the content that the gesture conveys.

Example (2), from Engle (2000, p. 37), illustrates a less constrained relationship between the contents conveyed by speech and gesture (pitch accents are shown with smallcaps).

(2) They [have SPRINGS.]
Speaker places right pinched hand (that seems to be holding a small vertical object) just above left pinched hand (that seems to be holding another small vertical thing).

The speaker here describes how the cotter pins on a lock are held in position. The utterance refers to the set of pins with *they* and the whole set of corresponding springs with *springs*. The gesture, however, depicts a *single* spring and pin in combination, highlighting the vertical relationship through which a spring pushes its corresponding pin into the key cylinder to help hold the cylinder in place. As is common, the gesture remains ambiguous; it is not clear which hand represents the spring and which the pin.[2] But even allowing for this ambiguity, we know that the gesture *elaborates* on the speech by showing the vertical spatial relationship maintained in a prototypical case of the relationship described in speech. Furthermore, as Engle notes, the gesture serves to disambiguate the plural predication in the accompanying sentence to a *distributive interpretation*.

Gestures maintain semantic links to questions and commands, as well as assertions. Such examples underscore the need to integrate the reference and content of gesture precisely with semantic and pragmatic representations of linguistic units. Consider the following example, taken from the AMI corpus (dialogue ES2002b, Carletta (2007)), in which a group of four people are tasked with designing a remote control:

(3) C: [Do you want to switch places?]
While C speaks, her right hand has its index finger extended; it starts at her waist and moves horizontally to the right towards D and then back again to C's waist, and this movement is repeated, as if to depict the motion of C moving to D's location and D moving to C's location.

Intuitively, the gesture in (3) adds content to C's question: it is not about whether D wants to switch places with someone unspecified, but rather switch places with C (the speaker). So overall, the multimodal action means "Do you want to switch places with me?". A different gesture, involving C's hand moving between agents D and A, would have resulted in a different overall question: Do you want to switch places with A? To capture this interaction in logical form, the interrogative operator associated with the question must have a referential or scopal dependence which allows its contribution to covary with the content of the gesture.

The following example is taken from the same dialogue as (1):

(4) [You walk out the doors]
The gesture is one with a flat hand shape and vertical palm, with the fingers pointing right, and palm facing outward.

The linguistic component expresses an instruction. And intuitively, the interpretation of the gesture is also an instruction: "and then immediately turn right". The inferential connection here must integrate both the semantic and pragmatic relationships between gesture and speech. Semantically, the two modalities respect a general presentational constraint that the

[2] In fact, springs push pins down into the cylinder, as diagrammed in the explanation this subject relied on, and as required by the need to make locks robust against gravity.

time and place where the first event ends (in this case, where the addressee would be once he walks out the door) overlaps with where the second event starts (turning right). Pragmatically, they are interpreted as presenting an integrated instruction. (Note that if we were to replace the clause in (4) with an indicative such as "John walked out the doors", then although the content of the utterance and the gesture would exhibit the same semantic relationship, the gesture would now have the illocutionary effects of an assertion.) Both aspects can be characterised by representing the content of the two as connected by a *rhetorical relation*: this reflects both their integrated content and the integrated speech act that the speaker accomplishes with the two components of the utterance (Asher and Lascarides, 2003). In this case, that relation is Narration, the prototypical way to present an integrated description of successive events in discourse.

Our examples thus far show how the hands can establish a consistent deictic frame to indicate objects and actions in a real or virtual space. Many gestures use space more abstractly, to depict aspects of form and motion in a described situation. A representative example is the series of gestures in (5)—Example 6 and Figure 8.4 in (Kendon, 2004, p. 136) that is an extract of the story of Little Red Riding Hood:

(5)
a. and [took his] [HATCHET and
First: Speaker's right hand grasps left hand, with wrists bent.
Second: Speaker lifts poised hands above right shoulder.

b. with] [a mighty SWEEP]
First: Hands, held above right shoulder, move back then forward slightly.

c. [(pause 0.4 sec)] [SLICED the wolf's stomach open]
First: Speaker turns head
Second: Arms swing in time with sliced; then are held horizontally at the left

In this example, the speaker assumes what McNeill (1992, p. 118) calls *character viewpoint*: she mirrors the actions of the woodsman as he kills the wolf. In (5a) and (5b), the speaker's hands, in coming together, depict the action of grabbing the handle of the hatchet and then the action of drawing the hatchet overhead, ready to strike; in (5c), the speaker's broad swinging motion depicts the woodsman's effort in delivering his blow to the wolf with the hatchet. The whole discourse thus exhibits a consistent dramatisation of the successive events, with the speaker understood to act out the part of the woodsman, and her hands in particular understood as holding and wielding the woodsman's hatchet. However, there seems to be no implication that these actions are demonstrated in the same spatial frame as previous or subsequent depictions of events in the story. Crosscultural studies, as in the work of Haviland (2000) among others, suggest that narrative traditions differ in how freely gestures can depart from a presupposed anchoring to a consistent real or virtual space, with examples like (5) in English representing the most liberal case. Following the descriptive literature on gesture, we represent the form of such gestures with *qualitative* features that mirror elements of the English descriptions we give for these gestures. In (5a), for example, we indicate that the speaker's hands are held above the right shoulder, that the right is grabbing the left, that both hands are in a fist shape as though grabbing something. Iconicity is captured by the relationship between these gesture elements and a naturally-related predication that each element contributes to the interpretation of the gesture as a whole.

Gestures with iconic meanings, like those with deictic meanings, must be interpreted in tight semantic and pragmatic integration with the accompanying utterances. Consider utterance (6) from the AMI corpus (dialogue ES2005b):

(6) 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.

The content of *D*'s gesture is presented in semantic interaction with the scope-bearing elements introduced in the sentence. Intuitively, the gesture depicts a keyboard, not anchored to any specific virtual space. There is nothing within the form of the gesture itself that depicts negation. Nevertheless, *D*'s overall multimodal act entails *not with a keyboard*. This requires the negation that's introduced by the word *not* to semantically outscope the content depicted by the gesture. This scope relation can be achieved only via a unified semantic framework for representing verbal and gestural content. In fact, we will argue in Section 4.3 that example (6) calls for an integrated, compositional description of utterance form and meaning that captures both linguistic and gestural components. That way, established and well-understood methods for composing semantic representations from descriptions of the part–whole structure of communicative actions can determine the semantic scope between gestured content and linguistically-introduced negation.

Iconicity gives rise to the same interpretive underspecification that we saw with deixis. For example consider the depiction of 'trashing' in the following example from a psychology lecture:[3]

(7) I can give you other books that would totally trash experimentalism.
When the speaker says *trash*, both hands are in an open flat handshape (ASL 5), with the palms facing each other and index fingers pointing forward. The palms are at a 45 degree angle to the horizontal. The hands start at the central torso, and move in parallel upwards and to the left.

While there is ambiguity in what the speaker's hands denote—they could be hands metaphorically holding experimentalism itself, or a representation of a statement of experimentalism such as a book, or the content of such a book—the gesture is clearly coherent and depicts experimentalism being thrown away.

The following example (8) illustrates the possibility for deictic and iconic imagery to combine in complex gestures. It is extracted from Kendon's Christmas cake narrative (2004, Fig. 15.6, pp. 321–322), where a speaker describes how his father, a small-town grocer, would sell pieces of a giant cake at Christmas time:

(8) a. and it was [pause 1.02 sec] this sort of [pause 0.4 sec] size
during the pauses, the speaker frames a large horizontal square using both hands; his index fingers are extended, but other fingers are drawn in, palms down.

b. and [he'd cut it off in bits]
the speaker lowers his right hand, held open, palm facing to his left, in one corner of the virtual square established in the previous gesture

The gesture in (8b) involves both iconic and deictic meaning. The iconicity comes in the configuration and orientation of the speaker's right hand, which mirrors a flat vertical surface involved in cutting: perhaps the knife used to cut the cake; or the path it follows through the cake; or the boundary thereby created. (We are by now familiar with such underspecification.) The deixis comes in the position of the speaker's hand, which is interpreted by reference to the virtual space occupied by the cake, as established by the previous gesture.

[3]See www.talkbank.org/media/ClassTalk/Lecture-unlinked/feb07/feb07-1.mov

We finish with an example, taken from a lecture on speech,[4] that—like all our examples—underscores how gesture interpretation is dependent on both its form and its coherent links to accompanying linguistic context:

(9) So there are these very low level phonological errors that tend to not get reported. The hand is in a fist with the thumb to the side (ASL A) and moves iteratively in the sagittal plane in clockwise circles (as viewed from left), below the mouth.

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

Figure 2: Hand gestures depicting speech errors.

One salient interpretation is that the gesture depicts the iterative processes that cause low-level phonological errors, slipping beneath everyone's awareness. In previous utterances, the speaker used both words and gestures to show that anecdotal methods for studying speech errors are biased towards *noticeable* errors like Spoonerisms. Those noticeable errors were depicted with the hand emerging *upward from the mouth* into a *prominent* space between the speaker and his audience. If we take the different position of the gesture in (9), below the mouth, nearer the speaker, as intended to signal a contrast with this earlier case, then we derive an interpretation of this gesture as depicting low-level phonological errors as less noticeable. At the same time, as in (8b), we might understand the hand shape iconically, with the fist shape suggesting the action of processes of production in bringing forth phonological material. This ensemble represents a coherent interpretation because it allows us to understand the gesture as providing information that directly supports what's said in the accompanying speech—the fact that these errors are less noticeable *explains why* anecdotal methods would not naturally detect them.

Of course, this interpretation is just one of several coherent alternatives for this gesture. Another plausible interpretation of the gesture in (9) is that it depicts the low level of the phonological errors, rather than the fact that these errors are less noticeable. This alternative interpretation is also coherently related to the linguistic content: like (1) it *depicts* objects that are denoted in the sentence. In fact, this interpretation would be supported by a distinct view of the gesture's form, where instead of conceiving it as a single stroke (as our prior interpretation requires since the repeated movement was taken to depict an iterative process), it is several strokes—a sequence of identical gestures, each consisting of a fist moving in exactly one circle, and each circle demonstrating a distinct low-level phonological error. This

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

alternative interpretation demonstrates how ambiguity can persist in a coherent discourse at all levels, from form to interpretation. But crucially, all plausible interpretations must satisfy the dual constraints that (a) the interpretation offer a plausible iconic rendition of the gesture's form; and (b) the interpretation be coherently related to the content conveyed by its synchronous speech. Accordingly, while computing the interpretation of gesture via *unification* with the content of synchronous speech may suffice for examples where gesture coherence is achieved through conveying the *same content* as speech (Kopp, Tepper and Cassell, 2004), on its own it cannot account for the gestures in examples such as (4) and (6) that evoke distinct, but related, objects and properties to those in the speech. Rather, computing an interpretation of the gesture that is coherently related to the content conveyed in speech will involve commonsense reasoning.

Whether the full inventory of rhetorical relations that are attested in linguistic discourse are also attested for relating a gesture to its synchronous speech is an empirical matter. But we rather suspect that certain relations are excluded—for instance interpreting a gesture so that it connects to its synchronous speech with *Disjunction* seems implausible (although Disjunction could relate one multimodal discourse unit that includes a gestural element to another). But this doesn't undermine the role of coherence relations in interpreting gesture any more than it does for interpreting other purely linguistic constructions that signal the presence of one of a strict subset of rhetorical connections. For example, sense ambiguous discourse connectives such as *and* are like this: *and* signals the presence of a rhetorical relation between its complements; it underspecifies its value, but it cannot be Disjunction (Carston, 2002). Similarly, Kortmann (1991) argues that the interpretation of free adjuncts (e.g., "opening the drawer, John found a revolver") involves inferring coherence relationships between the subordinate and main clauses, but certain relationships such as Disjunction are ruled out. It isn't surprising that synchronous speech and gesture likewise signals the presence of a coherence relation whose value is not fully determined by form, although certain relations are ruled out.

3 The Logical Form of Multimodal Communication

The overall architecture of our formalism responds to the claim, substantiated in Section 2, that gesture and speech present complementary, inferentially-related information as part of an integrated, overarching speech act with a uniform force and consistent assignments of scope relationships. We formalise this integration of gesture and speech by developing an integrated logical form (LF) for multimodal discourse, which, like the LF of purely linguistic discourse, makes explicit the illocutionary content that the speaker is committed to in the conversation. As in theories of linguistic discourse, we give a central place in LF to *rhetorical relations* between discourse units. Here rhetorical relations must not only link linguistic material together, but also gestures to synchronous speech and to other material in the ongoing discourse.

A rhetorical relation represents a type of (relational) speech act (Asher and Lascarides, 2003). Examples include Narration (describing one eventuality and then another that is in contingent succession); Background (a strategy like Narration's save that the eventualities temporally overlap); and Contrast (presenting related information about two entities, using parallelism of syntax and semantics to call attention to their differences). The inventory also includes *metatalk* relations, that relate units at the level of the speech acts rather than

content. For instance, you might follow "Chris is impulsive" with "I have to admit it"—an explanation of why you said Chris is impulsive, not an explanation of why Chris is impulsive. These are symbolised with a subscript star—Explanation$_*$ for this example.

To extend the account to gesture, we highlight an additional set of connections which specify the distinctive ways embodied communicative actions connect together. The examples from Section 2 provide evidence for three such relations. First, *Depiction* is the strategy of using a gesture to visualise exactly the content conveyed in speech. Example (1) is an illustrative case. The speaker says that Norris is across from the library at the same time as she depicts their relative locations across from one another. Technically, Depiction might be formalised as a special case of Elaboration, where the gesture does not present *additional* information to that in the speech. We distinguish Depiction, however, because Depiction does not carry the implicatures normally associated with redundancy in purely spoken discourse (Walker, 1993)—it is helpful, not marked.

Second, *Overlay* relates one gesture to another when the latter continues to develop the same virtual space. Example (1), which is preceded in the discourse by (10), illustrates this:

(10) a. Norris is like up here—
The right arm is extended directly forward from the shoulder with forearm slightly raised; the right palm is flat and faces up and to the left.

b. And then the library is over here.
After returning the right hand to rest, the right hand is re-extended now to the extreme upper right periphery, with palm held left.

The speaker evokes the same virtual space in (1) as in the preceding (10) by designating the same physical positions when naming the buildings. The rhetorical connection Overlay captures the intuition that commonalities in the use of space marks the coherent use of gesture. Here, a logical form that features Overlay between the successive gestures in (10ab) and (1) captures the correct spatial aspects of the discourse's content.

Our third new relation is *Replication*, which relates successive gestures that use the body in the same way to depict the same entities. The gestures of example (5) illustrate Replication. The initial gesture adopts a figuration in which the speaker represents the woodsman, with her hands modelling his grip on the handle of the hatchet. While the subsequent speech no longer explicitly mentions the hatchet or even the woodsman, subsequent gestures continue with the imagery adopted in the earlier gesture. Connecting subsequent gestures to earlier ones by Replication captures the coherence of this consistent imagery.

Our plan for this section is to formalise this programmatic outline, by developing representations of logical form that combine these rhetorical relations with appropriate models of spatial content (Section 3.1), dynamic semantics (Section 3.2), and discourse structure (Section 3.3). We then show that these logical forms allow for the interpretive links in reference, content and scope that we observed in Section 2. The section culminates in the presentation of a language $\mathcal{L}_{sdrs}$ (Section 3.4) for describing the content of multimodal discourse that is based on that of SDRT (Asher and Lascarides, 2003).

3.1 Spatial Content

We begin by formalising the spatial reference that underpins deixis as a dimension of gesture meaning. Our formalisation adds symbols for places and paths through space, variables that map physical space to virtual or abstract spaces, and predicates that record the propositional

information that gestures offer in locating entities in real, virtual and abstract spaces. This section presents each of these innovations in turn.

We begin by adding to the model a spatiotemporal locality $L \subset \mathcal{R}^4$ within which individuals and events can be located. We also add to the language a set of constants $\vec{p}_1, \vec{p}_2, \ldots$, which are mapped a subset of L by the model's interpretation function I—i.e., $[\![\vec{p}]\!]^M = I^M(\vec{p}) \subseteq L^M$. Whenever we need to formalise the place or path in physical space designated by the gesture, we use a suitable constant $\vec{p}$. We'll return shortly to how physical locations map to locations in the situation the speaker describes.

In Section 2, we used the gestures in (1) as representative examples of spatial reference, since the position of the right hand in space signals the (relative) locations of Norris Hall and the library in interpretation. To formalise this, we can use a spatial reference $\vec{p}_n$ denoting a place in front of the speaker's shoulder, and $\vec{p}_l$ denoting a place up and to her right. The contrast is now evident between the use of space in (1) and the mimicry in examples such as (5) and (7). Whereas (5) and (7) portray 'non-spatial' content through qualitative aspects of movement, (1) expresses intrinsically spatial information through explicit spatial reference.

For now, we remain relatively eclectic about how a speaker uses movement to indicate a spatiotemporal region. In (1) the speaker designates the *position* of the hand. But in (11)—a description of the library from the discourse preceding (1)—the speaker designates the *trajectory* followed by the hand:

(11) It's the weird-looking building over here.
The left hand shape is ASL 5 open, the palm facing right and fingers facing forward; the hand sweeps around to the right as though tracing the surface of a cylinder.

This trajectory is meant to represent the cylindrical exterior of the library—a fact that must be captured in LF via a suitable constant $\vec{p}_s$. As we saw in Section 2, such alternative methods of spatial reference give rise to ambiguities in the form and interpretation of gestures—ambiguities that may never be fully resolved. Accordingly, it may not be possible or desirable to draw inferences from *lf*s involving spatial constants such as $\vec{p}_s$ that depend on the constants' exact values.

A further crosscutting distinction is whether the speaker indicates the location of the hand itself, as in (1) and (11), or uses the hand to designate a distant region. The typical pointing gesture, with the index finger extended (the ASL 1-index hand shape) is often used in this way. Take the case of deferred reference illustrated in (12), after (Engle, 2000, Table 8, p38):

(12) [These things] push up the pins.
The speaker points closely at the frontmost wedge of the line of jagged wedges that runs along the top of a key as it enters the cylinder of a lock.

It seems clear that the speaker aims to call attention to the spatial location $\vec{p}_w$ of the first wedge, not the spatial location of the finger.[5]

Utterance (12) also contrasts with (1) and (11) in whether they link up with the real places or establish a virtual space that models the real world. In (12) the speaker locates the actual wedge on the key. In (1) and (11), however, the speaker is not actually pointing at the buildings—she marks their imagined locations in the space in front of her. The information

[5]This demonstrative noun phrase and accompanying demonstration is attested in Engle's data. Unfortunately Engle does not report the entire sentence context in which the gesture is used; the continuation is based on other examples she reports. Further examples of the variety of spatial reference in gesture are provided by Johnston et al. (1997), Johnston (1998), Lücking, Rieser and Staudacher (2006).

that (12) gives about the real world can therefore be characterised directly in terms of the real-world region $\vec{p}_w$ that the speaker's gesture designates. By contrast, the content of (1) and (11) can only be described in terms of context-dependent mappings v_c and v_s from the space in front of the speaker to (in this case) the Northwestern University campus. The relationship between the real positions of the speaker's hands during her demonstrations $\vec{p}_n$ and $\vec{p}_l$ in (1) thus serves to signal the corresponding relationship between the actual location of Norris Hall $v_c(\vec{p}_n)$ and the actual location of the library $v_c(\vec{p}_l)$. A related (perhaps identical) mapping v_s is at play in (11) when the speaker characterises the actual shape of the library facade $v_s(\vec{p}_s)$ in terms of the curved path $\vec{p}_s$ of her hand.

These spatiotemporal mappings are the second formal innovation of the language to represent gesture meaning. Formally, variables such as v_c and v_s are part of an infinite family of variables that denote transformations over L. They simplify the relationship between the form of a gesture and its semantics considerably. We do not have to assume an ambiguity between reference to physical space vs. virtual space. Rather, gesture always refers to physical space and always invokes a mapping between this physical space and the described situation—e.g., the gesture in (12) makes the relevant mapping the identity function v_I.

The values of these variables $v_1, v_2, \ldots$ are determined by context. Some continuations of discourse are coherent only when the speaker continues to use the space in his current gesture in the same way as his previous gestures. Other continuations are coherent even though the current gesture uses space in a different way. The values of $v_1, v_2, \ldots$ are therefore provided by assignment functions, which in our dynamic semantics mediate the formal treatment of context dependence and context change since they are a part of the context of evaluation (see Section 3.4). To respect the iconicity, the possible values for a mapping v is tightly constrained: they can rotate and re-scale space but not effect a mirroring transformation. At the same time (given their origin in human cognition and bodily action), we would not expect mappings to realise spatial relationships exactly. Here we simply assume that there is a suitably constrained set of mappings T in any model, and so where f is an assignment function, $f(v) \in T$.

As is standard in dynamic semantics (Williamson, 1994, Kyburg and Morreau, 2000, Barker, 2002), we understand the context-dependence of mapping variables to offer a solution to the problem of vagueness in spatial mappings. These variables take on precise values, given a precise context. However, interlocutors don't typically nail down a precise context, and if the denotation of a variable v is not determined uniquely, the *lf* will be correspondingly vague about spatial reference. A range of spatial reference will be possible and the interpretation of the gesture will fit a range of real-world layouts. For instance, utterances (1) and (11) can either exemplify a statement at a particular time, or extend their spatiotemporal reference throughout a wider temporal interval (Talmy, 1996). We also discussed in Section 2 that the gestures in (1) can be interpreted as demonstrating the buildings or the locations of the buildings. These alternatives correspond to distinct values for the mapping v that might both be supported by the discourse context. Interlocutors can converse to resolve these vagaries (Kyburg and Morreau, 2000, Barker, 2002). But eventually, even if some vagueness in interpretation persists, interlocutors understand each other well enough for the purposes of the conversation (Clark, 1996).

We complete this formalisation of spatial reference in gesture by introducing two new predicates: *loc* and *classify* respectively describe the literal and metaphorical use of space to locate an entity. *loc* is a 3-place predicate, and $loc(e, x, \vec{p})$ is true just in case at each moment spanned by the temporal interval e, x is spatially contained in the region specified by $\vec{p}$. For

example, (13a) represents the interpretation of the gestures in (1):

(13) a. $loc(e_1, n, v_c(\vec{p}_n)) \wedge loc(e_2, l, v_c(\vec{p}_l))$

b. $loc(e_3, f, v_s(\vec{p}_s)) \wedge facade(l, f)$

c. $loc(e_4, w, v_I(\vec{p}_w))$

In words, e_1 is the state of n, the discourse referent for Norris Hall that's introduced in the clause, being contained in the spatiotemporal image $v_c(\vec{p}_n)$ on the speaker's virtual map of a point $\vec{p}_n$ in front of her left shoulder; e_2 is the state of the library l being contained in the spatiotemporal image $v_c(\vec{p}_l)$ of the designated point $\vec{p}_l$ further up and to the right. (13b) states that the facade f of the library lies in the real-world cylindrical shell $v_s(\vec{p}_s)$ determined by the speaker's hand movement $\vec{p}_s$ in (11). The predication $facade(l, f)$ is not contributed by an explicit element of gesture morphology but is, as we describe in more detail in Section 3.2, the result of pragmatic inference that connects individuals introduced in the gesture to antecedents from the verbal content (i.e., the library). The logical form (13c) of the deictic gesture in (12) locates the frontmost wedge w at the (distant) location $\vec{p}_w$ where the speaker is pointing. The deferred reference from that one wedge w to the entire set of wedges at the top of the key as denoted by the linguistic phrase *these things* is obtained via pragmatics: context resolves to a specific value an underspecified relation between the deictic referent and the NP's referent, this underspecified relation being a part of the compositional semantics of the multimodal act (see Section 4.2). Indeed, identifying the gesture's referent and spatial mapping as w and v_I, resolving the referent of *these things*, and resolving this underspecified relation between them (to *exemplifies*) are logically co-dependent (see Section 5).

Speakers can also use positions in space as a proxy for abstract predications. Such metaphorical uses are formalised through the predicate *classify*. A representative illustration can be found in the conduit metaphor for communication (Reddy, 1993), where information contributed by a particular dialogue agent is metaphorically located with that agent in space, as shown in the naturally-occurring utterance (14):

(14) We have this one ball, as you said, Susan.
The speaker sits leaning forward, with the right hand elbow resting on his knee and the right hand held straight ahead, in a loose ASL-L gesture (thumb and index finger extended, other fingers curled) pointing at his addressee.

(14) is part of an extended explanation of the solution to a problem in physics. The speaker's explanation describes the results of a thought experiment that his addressee Susan had already introduced into the dialogue, and he works to acknowledge this in both speech and gesture. More precisely, both the gesture in (14) and the adjunct clause "as you said" function as meta-comments characterising the original source of "We have this one ball". The gesture, by being performed close to Susan and away from the speaker's body, shows that his contribution here is metaphorically located with Susan; that is, it recapitulates content she contributed.

Formally, we handle such metaphorical reference to space by assuming a background of meaning postulates that link spatial coordinates with corresponding properties. The predicate *classify* is used to express instantiations of this link. For example, corresponding to the conduit metaphor is a virtual space v_m that associates people with their contributions to conversation. If $\vec{p}_i$ is the location of any interlocutor i, then $classify(e, u, v_m(\vec{p}_i))$ is true exactly when utterance u presents content that is originally due to a contribution by i. The generative mapping v_m between interlocutors' locations and contributed content shows why

the metaphor depends on spatial reference. The LF of (14) will therefore use the formula $\mathit{classify}(e_6, u', v_m(\vec{p}_S))$, where u' denotes an utterance whose content entails "We have this one ball", and $\vec{p}_S$ denotes Susan's location.

This treatment of metaphor is continuous with models of linguistic metaphor as context-dependent indirect reference (Stern, 2000, Glucksberg and McGlone, 2001). This indirect reference depends on the conventional referent (here a point in space) and a *generating principle* taken from context that maps the conventional referent into another domain (here the domain of contributing to conversation). As revealed within cognitive linguistics (Lakoff and Johnson, 1981, Gibbs, 1994, Fauconnier, 1997), such mappings are typically richly-structured but flexible and open-ended. Thus, interlocutors will no more fix a unique way to understand the metaphor than they will fix other aspects of context-dependent interpretation. So metaphorical interpretations on our account—though represented precisely in terms of context-dependent reference—will remain vague.

3.2 Dynamic Semantics

We proceed by formalising a shared set of constraints on co-reference in discourse that describes both deictic and iconic gesture meaning. Our formalisation distinguishes between entities introduced in speech and those introduced in gesture. As a provisional account of our empirical data and linguistic intuitions, we articulate a model in which entities introduced in gesture must be bridging-related to entities introduced explicitly in accompanying speech (Clark, 1977). These inferred entities can figure in the interpretation of subsequent gestures but do not license pronominal anaphora in subsequent speech. This provisional model offers a lens with which to focus future empirical and theoretical research.

In Section 2, we presented a range of examples in which gestures seem most naturally understood as depicting entities that are not directly referenced in the accompanying speech, but which stand in close relations to them given commonsense knowledge. For instance in (2), the prototypical spring and the prototypical pin depicted in gesture are inferable from the set of springs and the set of pins explicitly referenced in words. Conversely in (12), the referent of "these things" is inferable from but not identical to the specific individual wedge that's denoted by the speaker's gesture. The grip of the woodsman's hands on the handle of the hatchet, we suggested, is inferable but not explicitly referenced in the words "took his hatchet" of (5a). The knife-edge depicted by the hand in (8b) is also inferable but not explicitly referenced in the accompanying statement about the grocer's work with the cake, "he'd cut it off into bits".

While we do not believe gestures refer only to entities evoked in speech, we do think some inferential connection to prior discourse is necessary. Not only does this characterise all the examples we have investigated, but to go beyond inference would place an uncharacteristically heavy burden on gesture meaning, which is typically ambiguous and open-ended except when interpreted in close connection to its discourse context. Thus, we will formalise initial references to new entities in gesture by analogy to definite references to inferable entities in discourse—what is known as *bridging* in the discourse literature (Clark, 1977). Our techniques are standard; see e.g. Chierchia (1995). But nevertheless they offer the chance to engage future empirical work on reference in gesture with a related formal and empirical approach to linguistic discourse.

Entities depicted initially in gesture remain available for the interpretation of subsequent gestures. For example, the lagoon, located as a landmark on the speaker's left in the initial

segment of the direction-giving episode excerpted in (1) serves as the referent demonstrated by the speaker's left hand in both gestures of (1)—despite the fact that the speaker does not continue to reference the lagoon in speech. Similarly, once introduced in (5a), the grip of the woodsman's hands on the handle of the hatchet continues to guide the interpretation of the gestures of (5b), even though the speaker does not mention the hatchet again.

By contrast, inferable entities evoked only in gesture seem not to be brought to prominence in such a way as to license the use of a pronoun in subsequent speech. Such examples would be very surprising in light of the tradition in discourse semantics—see e.g., Heim (1982)—that sees pronominal reference as a reflex of a formal link to a linguistic antecedent. In fact, we have found no such examples. And our intuitions find hypothetical attempts at such reference quite unnatural. Try, for example, following (8b) "he'd cut it off into bits" with ¿'and it would get frosting all over it", with *it* understood to pick out the cutting edge demonstrated in gesture in (8b). We show in this section how to formalise such a constraint.

Our formalism builds an existing *dynamic semantic* model of anaphora in discourse, since with dynamic semantics—unlike alternatives such as Centering Theory (Grosz, Joshi and Weinstein, 1995) and Optimality Theory (Buchwald et al., 2002)—we can build on previous work that integrates anaphoric interpretation with rhetorical relations and discourse structure (Asher and Lascarides, 2003). We use a dynamic semantics where context is represented as a partial variable assignment function (van Eijck and Kamp, 1997). As a discourse is interpreted the model remains fixed but the input assignment function changes into a different output one. The basic operations are to *test* that the input context satisfies certain conditions (with respect to the model), to *extend* the input variable assignment function one by defining a value for a new variable, and to *sequence* two actions together, thereby composing their individual effects on their input contexts. These primitives are already used to model the anaphoric dependencies across sentence boundaries; here we will use them to model anaphoric dependencies between spoken utterances and gesture and between sequences of gestures.[6]

To distinguish between a set of prominent entities introduced explicitly in speech and a set of background entities that's depicted only in gesture, we take our cue from Bittner's 2001 formalisation of centering morphology as distinguishing foreground and background entities. This involves splitting the context into two assignment functions $\langle f, g\rangle$ to distinguish referents of different status. For us, the first function f records the entities that can be used to interpret pronouns and other anaphors in speech; the second one g records the entities that are the basis for referential depiction in gesture. To ensure that linguistic indefinites can license depiction in gesture (e.g., the indefinite *springs* in (2) licenses the depiction of a prototypical spring in gesture), existential quantifiers in the logic will trigger an update to both f and g. Meanwhile, to delimit the scope of gesture, we introduce an *operator* $[\mathcal{G}]$ over formulae, which semantically restricts the context-updating operations that happen within its scope to only the second function g (see Section 3.4 for details). $[\mathcal{G}]$ is extensional, not modal. It allows us to capture the different status of referents introduced in different ways

[6]Our framework aims for the simplest possible dynamic semantic formalisation. This involves taking up generalisations from linguistic discourse as a provisional guide to the behaviour of gesture. For example, we assume that a gesture that's outscoped by a negation, like the *keyboard* gesture of (6), does not license subsequent anaphora, just as an indefinite in speech that's outscoped by a negation does not license a subsequent pronoun (Groenendijk and Stokhof, 1991). Other analyses are possible, with more complex dynamic semantics, modeled for example after treatments of so-called specific indefinites in language; see Farkas (2002). More generally, our developments are compatible with more complex architectures for dynamic semantics. But the formalism we provide is already sufficient for interpreting the key examples we consider here.

while allowing us to treat gesture and speech within a common logical representation—as we must, given the inferential and scopal dependencies we observed in Section 2. The need for this formal device is independent of the use of rhetorical relations to integrate content from different modalities together. In particular, the need for a suitable dynamic semantics does not undermine our claim that gesture and speech are rhetorically connected. Modal subordination and grammatically-marked focus systems in language also block individuals in one clause from being an antecedent to anaphora in another, even when there's a clear rationale for a rhetorical connection. So the anaphoric constraints across speech and gesture are no more a counterargument to rhetorical connections than modal subordination and focus give counterarguments to using rhetorical relations to model linguistic discourse.

3.3 Rhetorical Relations and Discourse Structure

There are several existing frameworks that use rhetorical relations; e.g., Mann and Thompson (1987), Hobbs et al. (1993). We will use Segmented Discourse Representation Theory (SDRT, Asher and Lascarides (2003)) as our starting point, for three main reasons. First, SDRT fully supports *semantic underspecification*. This is useful because the meaning of a gesture as revealed by its form is highly underspecified—we can re-use SDRT's existing techniques for resolving underspecified content to pragmatically preferred values in our model of gesture interpretation (see Section 5). Secondly, SDRT acknowledges that ambiguity can persist in a coherent discourse. Its motivation for doing so stems originally from observing that there may be more than one maximally coherent interpretation of linguistic discourse. We have seen that the same is true of gesture, and so a framework where coherence constrains interpretation but doesn't necessarily resolve it uniquely is essential. Finally, SDRT offers a dynamic semantic interpretation of logical forms. We have seen already in Section 3.2 that dynamic semantics offers an elegant way of modelling both the vagueness in gesture interpretation and constraints on co-reference.

In SDRT, logical forms consist of *labels* $\pi_1, \pi_2, \ldots$ that each represent a unit of discourse, and a function that associates each label with a formula that represents the unit's interpretation—these formulae can be rhetorical relations between labels. We will treat individual clauses and gestures as units of discourse and so they each receive a label.

Rhetorical connections among units of discourse create discourse segments: π_1 *immediately outscopes* π_2 if π_1's formula includes $R(\pi, \pi_2)$ or $R(\pi_2, \pi)$ for some R and π. While a segment may consist of (or outscope) a *continuous* set of discourse units, this isn't necessary; see Asher and Lascarides (2003) for many examples. Gestures likewise structure discourse in flexible but constrained ways. As we have seen, gestures like those in (10ab) followed by (1) will bear a rhetorical relation both to simultaneous speech and to previous gestures. In all cases, however, the outscopes relation over labels in an LF cannot contain cycles and must have a single root—i.e., the unique segment that is the entire discourse.

Each rhetorical relation symbol receives a semantic interpretation that's defined in terms of the semantics of its arguments. For instance, a discourse unit that is formed by connecting together smaller units π_1 and π_2 with a *veridical* rhetorical relation R entails the content of the smaller units, as interpreted in dynamic succession, and goes on to add a set of conditions $\varphi_{R(\pi_1,\pi_2)}$ that encode the particular illocutionary effects of R. For example, $\mathit{Explanation}(\pi_1, \pi_2)$ transforms an input context C into an output one C' only if $K_{\pi_1} \wedge K_{\pi_2} \wedge \varphi_{\mathit{Explanation}(\pi_1,\pi_2)}$ also does this, where $\wedge$ is dynamic conjunction, and K_{π_1} and K_{π_2} are the contents of labels π_1 and π_2 respectively. The formula $\varphi_{\mathit{Explanation}(\pi_1,\pi_2)}$ is a test

on its input context. Meaning postulates constrain its interpretation: e.g., K_{π_2} must be an answer to the question *Why* K_{π_1}*?* The formalisation of the three new rhetorical relations is straightforward in this setting; see Section 3.4. The content of the entire discourse is then interpreted in a compositional manner, by recursively unpacking the truth conditions of the formula that's associated with the unique root label. This natural extension of the formal tools for describing discourse coherence fits what we see as the fundamental commonality in mechanisms for representing and establishing coherence across all modalities.

The structure induced by the labels and their rhetorical connections impose constraints and preferences over interpretations: in other words, discourse structure guides the resolution of ambiguity and semantic underspecification that's induced by form. For example, SDRT gives a characterisation of *attachment* that limits the rhetorical links interpreters should consider for a new discourse unit, based on the pattern of links in preceding discourse. Rhetorical connections also restrict the *availability* of referents as antecedents to anaphora. Both of these ingredients of SDRT carry over to embodied discourse (Lascarides and Stone, in press).

Anticipating the arguments of Section 4, we assume a distinction between *identifying* gestures, which simply demonstrate objects, from more general *visualising* gestures, which depict some aspect of the world. Interpreting a gesture may involve resolving an ambiguity in its form, as to whether it is identifying or visualising. Identifying gestures are interpreted in construction with a suitable linguistic constituent. (We remain agnostic about the temporal and structural constraints between speech and gesture that may apply here.) The joint interpretation introduces an underspecified predicate symbol—call it *id_rel*—that relates the referent of the identifying gesture to the semantic index of the corresponding linguistic unit. Constructing the discourse's LF then involves resolving the underspecified relation *id_rel* to a specific value. The complex demonstrative in (12) represents such a case. The speaker's identifying gesture refers to the frontmost wedge on the key. That referent exemplifies the set denoted by the demonstrative NP *these things*; so in this case *id_rel* resolves to *exemplifies*. Discourse structure and commonsense reasoning guide this process of resolution.

Similarly, visualising gestures are also interpreted in construction with a suitable linguistic constituent (often a clause). The joint interpretation introduces an underspecified *rhetorical connection vis_rel*(π_s, π_g) between the spoken part π_s and the gesture part π_g. Thus for a visualising gesture, constructing the LF of the discourse involves achieving (at least) four, logically co-dependent tasks that are all designed to make the underspecified logical form that's derived from its form more specific. First, one resolves the underspecified content of the gesture to specific values. Second, the underspecified rhetorical connection *vis_rel* is resolved to one of a set of constrained values (e.g., Narration is OK, Disjunction is not). Third, one identifies a label for this rhetorical connection: if it's a new label which in turn attaches to some label in the context then π_s and π_g start a new discourse segment; otherwise it's an existing label and π_s and π_g continue that existing segment. And finally, one computes whether π_g and/or π_s are also rhetorically connected to other labels.

Discourse structure imposes constraints on all of these decisions. In SDRT, the available labels in the context for connections to new ones are either (i) the last label π_l that was added, or (ii) any label that dominates π_l via a sequence of the outscopes relation and/or subordinating relations (e.g., Elaboration, Explanation, Background are subordinating, while Narration is not). This corresponds to the *right frontier* when the discourse structure is diagrammed as a graph with outscoping and subordinating relations depicted with downward

arcs.[7] However, extending SDRT to handle gesture introduces a complication, because the last label is not unique: while a linguistic discourse imposes a linear order on its minimal discourse units (one must say or write one clause at a time), this linear order breaks down when one gestures and speaks at the same time. As the most conservative possible working hypothesis, we assume there that new attachments remain limited to the right frontier, the only difference being that instead of one last label there are two: the label π_l^s for the last minimal spoken unit, and the label π_l^g for its synchronous gesture (if there was one). Since there are two last labels there are now two right frontiers. The 'spoken' right-frontier Π^s is the set of all labels that dominate π_l^s via outscopes and subordinating relations, and the 'gesture' frontier Π^g is the set of all labels that dominate π_l^g. Thus the available labels are $\Pi^s \cup \Pi^g$ (note that $\Pi^s \cap \Pi^g$ is non-empty and contains at least the root π_0). In other words, when an utterance attaches to its context, the dependencies of its speech and gesture are satisfied either through the connection to the discourse as a whole, or to one another, or to the continued organisation of communication in their respective modalities. Given this definition of attachment, antecedents to anaphora can be controlled as in Asher and Lascarides 2003—roughly, the antecedent is in the same discourse unit or in one that's rhetorically connected to it.

This definition combines with the dynamic semantics from Section 3.2 to make precise predictions about how logical structure and discourse structure constrain co-reference. For instance, the co-reference across gestures that we observed in the extended discourse (10ab) followed by (1) satisfies these constraints—each spoken clause is connected to its prior one with the subordinating relation Background, making all the spoken labels on the 'spoken' right frontier; and each gesture is connected to the prior one with the subordinating relation Overlay (this entails that space is used in the same way in the gestures), placing them on the 'gesture' right frontier. So all labels remain available. Each gesture also connects to its synchronous clause with Depiction—a coordinating relation because the content of one argument is not more fine-grained or 'backgrounded' relative to the other. Nevertheless, all labels are on the right frontier because of the Background and Overlay connections.

Conversely, availability constrains the interpretation of the gestures in (15) given in Figure 3—which is taken from the same dialogue as (10) and (1)—in ways that match intuitions. To see this, we discuss the interpretation of each multimodal action in turn. Intuitively, the gesture in (15a) is related to its synchronous clause with Depiction, since it demonstrates the direction to turn into. (15b) attaches to the spoken and gesture labels of (15a) with the subordinating relation Acknowledgement—thus A's utterance (15c) can connect to (15a). And indeed it does: the speech unit in (15c) is related to that in (15a) with the coordinating relation Narration—so it is interpreted as a command to keep walking *after* turning right. Furthermore, the gesture connects to (15a)'s gesture with Overlay so that it conveys keep walking rightward from the position where you turned.

But (15d) marks a change in how the the physical location of the hands map to the locations of landmarks on the university campus. The discourse cue phrase *and then* in (15d) implicates that the linguistic unit attaches to the prior one with Narration, and its gesture attaches to the content of this clause with Depiction to capture the intuition that it depicts the road—the object introduced in the clause. But crucially, the physical location of this depiction bears no relation to the spatial setup given by the prior two gestures: technically, it does not attach to (15c) with Overlay, reflecting the fact that the mapping v' from physical

[7] Parallel, Contrast and discourse subordination relax this right-frontier constraint, but we ignore this here.

(15)

a. A: So then, once you get to that parking lot you turn [right].
When A says the word *right*, her right hand is held in flat open shape (ASL-5) with the palm facing forward and fingers pointing right, and the hand is held far out to the right of her torso.

b. B: Right

c. A: And you [keep walking].
When A says *keep walking*, she repeats the gesture from (15a).

d. A: And then there's like a [road]
When A says *road*, both her hands are brought back towards the centre of her torso, with flat open hand shapes (ASL-5) and palms held vertically and facing each other, with fingers facing forwards.

e. B: U-huh

f. A: [It will just kinda like consolidate, you know, like come into a road].
A's hands are in ASL-5, and they start with the palms facing each other at an angle (as if the hands form two sides of a equilateral triangle), very close to her central torso. They then sweep out towards B, and the hands go from being at an angle to being parallel to each other.

g. A: [Just stay on the road and then walk for a little bit].
A's right hand starts at the centre of her torso and sweeps out to the right.

h. A: [There are buildings over here]
A's right hand goes to her right, to the same place where her hand was in the (15a)'s gesture . Her hand is in a loose claw shape with the palm facing down.

Figure 3: An extract from a direction giving dialogue that features Narration.

space to virtual space that is a part of its interpretation is different from the mapping v that was used earlier. To put this another way, the road that is demonstrated in (15d) is not to the left of the walking path that's demonstrated in (15c), even though the hands in (15d) are to the left of where they were in (15c). These rhetorical connections mean that the gestures in (15ac) are no longer on the right frontier. And thus, according to our model, the mapping v that's evoked by (15ac) is no longer available for interpreting subsequent gestures (while the mapping v' that's used in (15d) is available). Interestingly, this prediction matches our intuitions about how the gestures in (15fg) are interpreted. In particular, even though the gesture in (15g) places the right hand in the same place as it was in (15a), it does *not* demonstrate that the buildings it denotes are co-located with the place where the agent is to turn right (i.e., at the parking lot). The right frontier constraint likewise predicts that the clause in (15g) cannot connect to the clause in (15a); it cannot be interpreted in the same way as the discourse "So then, once you get to the parking lot you turn right; there are buildings over here".

3.4 Summary: Formalism and Examples

We now complete our presentation of the logical form for embodied discourse by giving formally precise definitions. We start with the syntax of the language $\mathcal{L}_{sdrs}$ for expressing logical forms, which is based on that of SDRT. It is extended to include spatial expressions (see Section 3.1), and the two last labels (see Section 3.3). The dynamic semantics of $\mathcal{L}_{sdrs}$ is similar

to that in Asher and Lascarides (2003), except that a context of evaluation is refined to include two partial variable assignment functions rather than one; these track salient entities for interpreting language and gesture respectively (see Section 3.2). We close with worked examples to illustrate the formalism.

Definition 1 Vocabulary and Terms

The following vocabulary provides the syntactic atoms of the language $\mathcal{L}_{sdrs}$:

- A set $\mathcal{P}$ of predicate symbols $(P_1, P_2, \ldots)$ each with a specified *sort* giving its arity and the type of term in each argument position;
- A set $\mathcal{R}$ of (2-place) rhetorical relation symbols over labels (e.g., *Contrast*, *Explanation*, *Narration*, *Overlay*, ...);
- Individual variables $(x_1, x_2, y, z \ldots)$; eventuality variables $(e_1, e_2 \ldots)$; and constants for spatiotemporal regions $(\vec{p}_1, \vec{p}_2 \ldots)$;
- Variables over mappings between spatiotemporal regions $(v_1, v_2 \ldots)$;
- The boolean operators $(\neg, \wedge)$; the operator $[\mathcal{G}]$; and quantifiers $\forall$ and $\exists$;
- Labels $\pi_1, \pi_2 \ldots$.

We also define *terms* from this vocabulary, each with a corresponding sort. Individual variables are individual terms; eventuality variables are eventuality terms; and if $\vec{p}$ is a constant for a spatiotemporal region and v is a variable over mappings, then $\vec{p}$ and $v(\vec{p})$ are place terms.

A logical form (LF) for discourse is a Segmented Discourse Representation Structure (SDRS). This is constructed from SDRS-formulae in $\mathcal{L}_{sdrs}$ as defined in Definition 2:

Definition 2 SDRS-Formulae

The definition of the SDRS-formulae $\mathcal{L}_{sdrs}$ starts with a definition of a subset $\mathcal{L}_{base} \subset \mathcal{L}_{sdrs}$ of SDRS-formulae that feature no rhetorical relations:

1. If $P \in \mathcal{P}$ is an n-place predicate and $i_1, i_2, \ldots, i_n$ are terms of the appropriate sort for P, then $P(i_1, \ldots, i_n) \in \mathcal{L}_{base}$.
2. If $\phi, \psi \in \mathcal{L}_{base}$ and u is a variable, then $\exists u\phi, \forall u\phi \in \mathcal{L}_{base}$
3. If R is a rhetorical relation and π_1 and π_2 are labels, then $R(\pi_1, \pi_2) \in \mathcal{L}_{sdrs}$
4. If $\phi, \psi \in \mathcal{L}_{sdrs}$, then $\phi \wedge \psi, \neg\phi, [\mathcal{G}]\phi \in \mathcal{L}_{sdrs}$.

An SDRS is a set of labels (two of which are designated to be last), and a set of SDRS-formulae associated with each label:

Definition 3 SDRS

An SDRS is a triple: $\langle A, F, last \rangle$, where:

- A is a set of labels;
- F is a mapping from A to $\mathcal{L}_{sdrs}$; and
- *last* is set containing at most two labels $\{\pi_s, \pi_g\} \subseteq A$, where π_s labels the content of a token linguistic unit, and π_g the content of a token gesture (intuitively, this is the last multimodal act and *last* will contain no gesture label if the act had no gesture).

We say that π *immediately outscopes* π' iff $F(\pi)$ contains π' as a literal. Its transitive closure $\succ$ must be a well-founded partial order with a unique root (that is, there is a unique $\pi_0 \in A$ such that $\forall \pi \in A$, $\pi_0 \succeq \pi$).

The unique root makes an SDRS the logical form of a single discourse: the segment that the root label corresponds to is the entire discourse. The outscopes relation need not form a tree, reflecting the fact that a single communicative act can play multiple illocutionary roles in its context (see Section 3.3). When there is no confusion we may omit *last* from the specification of an SDRS, writing it $\langle A, F \rangle$. We may also write $F(\pi) = \phi$ as $\pi : \phi$. And we will continue occasionally to use K_π as notation for the the content $F(\pi)$. An example SDRS is shown in (9′); this represents one of the plausible interpretations of (9) (see Section 2)—the gesture depicts the subconscious nature of the processes that sustain low-level phonological errors:

(9) So there are these very low level phonological errors that tend to not get reported. The hand is in a fist with the thumb to the side (ASL A) and moves iteratively in the sagittal plane in clockwise circles (as viewed from left), below the mouth.

(9′) $\pi_1 : \exists y(\textit{low-level}(y) \wedge \textit{phonological}(y) \wedge \textit{errors}(y) \wedge \textit{go-unreported}(e, y))$
$\pi_2 : [\mathcal{G}]\exists x(\textit{continuous}(x) \wedge \textit{below-awareness}(x) \wedge \textit{process}(x) \wedge \textit{sustain}(e', x, y))$
$\pi_0 : \textit{Explanation}(\pi_1, \pi_2)$

We will shortly discuss the much more incomplete representation of meaning that is revealed by (9)'s form, and how commonsense reasoning uses that together with contextual information to construct the SDRS (9′). But first, we give details of the model theory of SDRSs, ensuring in particular that the dynamic semantics of (9′) is as intended.

Definition 4 Model

A model is a tuple $\langle D, L, T, I \rangle$ where:

- D consists of eventualities (D_E) and individuals (D_I).
- $L \subset \mathcal{R}^4$ is a spatiotemporal locality.
- T is a set of constrained mappings from L to L (i.e., they can expand, contract and rotate space, but not invert it).
- I is an interpretation function that maps non-logical constants from $\mathcal{L}_{base}$ to denotations of appropriate type (e.g., $I(\vec{p}) \subseteq L$).

Note that I does not assign denotations to rhetorical relations; we'll return to them shortly. But the semantics of all SDRS-formulae ϕ relative to a model M will specify a *context-change potential* that characterises exactly when ϕ relates an input context to an output one. A context is a pair of partial variable assignment functions $\langle f, g \rangle$ (see Section 3.2 for motivation); these define values for individual variables ($f(x) \in D_I$), eventuality variables ($f(e) \in D_E$) and spatial mappings ($f(v) \in T$).

As is usual in dynamic semantics, all atomic formulae and $\neg\phi$ are tests on the input context. The existential quantifier $\exists x$ extends the input functions $\langle f, g \rangle$ to be defined for x, and dynamic conjunction is composition. Hence $\exists x \phi$ is equivalent to $\exists x \wedge \phi$. The operator $[\mathcal{G}]$ for gesture ensures that all formulae in its scope act as tests or updates only on the function g in the input context $\langle f, g \rangle$, but leave f unchanged. This means that the denotations for each occurrence of x in $([\mathcal{G}]\exists x P(x)) \wedge ([\mathcal{G}]Q(x))$ are identical, but they do not co-refer in $([\mathcal{G}]\exists x P(x)) \wedge Q(x)$. This matches intuitions about co-reference in discourse (across gestures vs. from gesture to subsequent speech respectively) that we discussed in Section 3.2.

Definition 5 Semantics of SDRS-formulae without rhetorical relations

1. Where i is a constant term, $\langle f, g\rangle[\![i]\!]^M = I(i)$.
2. Where i is a variable, $\langle f, g\rangle[\![i]\!]^M = f(i)$.
3. Where $v(\vec{p})$ is a spatial term, $\langle f, g\rangle[\![v(\vec{p})]\!]^M = f(v)(I(\vec{p}))$.
4. For a formula $P^n(i_1, \ldots, i_n)$, $\langle f, g\rangle[\![P^n(i_1, \ldots, i_n)]\!]^M\langle f', g'\rangle$ iff $\langle f, g\rangle = \langle f', g'\rangle$ and $\langle\langle f, g\rangle[\![i_1]\!]^M, \ldots, \langle f, g\rangle[\![i_n]\!]^M\rangle \in I(P^n)$
5. $\langle f, g\rangle[\![\exists x]\!]^M\langle f', g'\rangle$ iff:
   (a) $dom(f') = dom(f) \cup \{x\}$, and $\forall y \in dom(f)$, $f'(y) = f(y)$ (i.e., $f \subseteq_x f'$);
   (b) $dom(g') = dom(g) \cup \{x\}$ and $\forall y \in dom(g)$, $g'(y) = g(y)$ (i.e., $g \subseteq_x g'$);
   (c) $f'(x) = g'(x)$.
6. $\langle f, g\rangle[\![\phi \wedge \psi]\!]^M\langle f', g'\rangle$ iff $\langle f, g\rangle[\![\phi]\!]^M \circ [\![\psi]\!]^M\langle f', g'\rangle$.
7. $\langle f, g\rangle[\![\neg\phi]\!]^M\langle f', g'\rangle$ iff $\langle f, g\rangle = \langle f', g'\rangle$ and for no $\langle f'', g''\rangle$, $\langle f, g\rangle[\![\phi]\!]^M\langle f'', g''\rangle$
8. $\langle f, g\rangle[\![[\mathcal{G}](\phi)]\!]^M\langle f', g'\rangle$ iff $f = f'$ and $\exists g''$ such that $\langle g, g\rangle[\![\phi]\!]^M\langle g'', g'\rangle$

Finally, we address the semantics of rhetorical relations. Unlike the predicate symbols in $\mathcal{P}$, these do *not* impose tests on the input context. As speech acts, they change the context just like actions generally do. We emphasise veridical relations:

Definition 6 Semantic Schema for Rhetorical Relations

Let R be a veridical rhetorical relation (i.e., *Narration*, *Background*, *Elaboration*, *Explanation*, *Contrast*, *Parallel*, *Depiction*, *Overlay*, *Replication*). Then:

$$\langle f, g\rangle[\![R(\pi_1, \pi_2)]\!]^M\langle f', g'\rangle \text{ iff } \langle f, g\rangle[\![K_{\pi_1} \wedge K_{\pi_2} \wedge \varphi_{R(\pi_1,\pi_2)}]\!]^M\langle f', g'\rangle$$

In words, $R(\pi_1, \pi_2)$ transforms an input context $\langle f, g\rangle$ into an output one $\langle f', g'\rangle$ if and only if the contents K_{π_1} followed by K_{π_2} followed by some particular illocutionary effects $\varphi_{R(\pi_1,\pi_2)}$ also do this. Meaning postulates then impose constraints on the illocutionary effects $\varphi_{R(\pi_1,\pi_2)}$ for various relations R. For instance, the meaning postulate for $\varphi_{Narration(\pi_1,\pi_2)}$ stipulates that individuals are in the same spatio-temporal location at the end of the first described event e_{π_1} as they are at the start of the second described event e_{π_2}, and so e_{π_1} temporally precedes e_{π_2} (we assume *prestate* and *poststate* are functions that map an eventuality to the spatiotemporal regions in L of its prestate and poststate respectively):[8]

- *Meaning Postulate for Narration*
  $\varphi_{Narration(\pi_1,\pi_2)} \rightarrow overlap(poststate(e_{\pi_1}), prestate(e_{\pi_2}))$

So, for instance, representing (16) with $\pi_0 : Narration(\pi_1, \pi_2)$, where π_1 and π_2 label the contents of the clauses, ensures its dynamic interpretation matches intuitions: John went out the door, and then from the other side of the door he turned right.

(16) π_1. John went out the door.
π_2. He turned right.

[8] In fact, this axiom is stated here in simplified form; for details, see Asher and Lascarides (2003).

It's important to stress, however, that these interpretations of rhetorical relations are defined only with respect to *complete interpretations*: K_{π_1} and K_{π_2} must be SDRS-formulae, with all underspecified aspects of content that are revealed by form fully resolved. Accordingly, the type of speech act that is performed is a property of a contextually resolved interpretation of an utterance (or a gesture) rather than a property of its form. This belies the fact that in linguistic discourse, it is possible to align certain linguistic forms with certain types of speech acts: e.g., indicatives tend to be assertions while interrogatives tend to be questions. Such alignments are not possible with gesture, and our theory reflects this: the form of a gesture on its own is insufficient for inferring anything about its illocutionary effects; it is only when it is combined with context that clues about the speech act are revealed. In short, the syntax and model theory of $\mathcal{L}_{sdrs}$ are designed only for representing the pragmatically preferred interpretations. And when there is more than one pragmatically plausible interpretation, there is more than one logical form expressed in $\mathcal{L}_{sdrs}$. We will examine shortly how these pragmatic interpretations are inferred from form and context.

Rhetorical relations that are already a part of SDRT—like Explanation—can now relate contents of gesture. We also argued earlier for three relations whose arguments are restricted to gesture: Depiction, Overlay and Replication. These are all veridical relations, and the meaning postulates that define their illocutionary effects match the informal definitions given earlier. For instance, $\mathit{Depiction}(\pi_1, \pi_2)$ holds only if π_1 labels the content of a *spoken* unit, π_2 labels the content of a *gesture*, and K_{π_1} and K_{π_2} are nonmonotonically equivalent; we omit formal details because it requires a modal model theory for $\mathcal{L}_{sdrs}$ as described in Asher and Lascarides (2003). We assume that a discourse unit labels speech only if all the minimal units outscoped by it label speech; similarly for gesture. $\mathit{Overlay}(\pi_1, \pi_2)$ holds only if π_1 and π_2 are gestures, and K_{π_2} continues to develop the same virtual space as K_{π_1}: in other words, K_{π_1} and K_{π_2} entail contingent formulae containing the same mapping v. Finally, $\varphi_{\mathit{Replication}(\pi_1,\pi_2)}$ holds only if π_1 and π_2 are gestures, and they depict common entities in the same way. More formally, there is a partial isomorphic mapping μ from the constructors in K_{π_1} to those in K_{π_2} such that for all constructors c from K_{π_1}, c and $\mu(c)$ are semantically similar. We forego a formal definition of semantic similarity here.

Definition 3 now formalises the interpretation of an SDRS:

Definition 7 The Dynamic Interpretation of an SDRS

Let $S = \langle A, F, \mathit{last}\rangle$ be an SDRS, and let $\pi_0 \in A$ be its unique root. Then:

$$\langle f, g\rangle [\![S]\!]^M \langle f', g'\rangle \text{ iff } \langle f, g\rangle [\![F(\pi_0)]\!]^M \langle f', g'\rangle$$

If the SDRS features only veridical rhetorical relations, then it transforms an input context into an output one only if the contents of each clause and gesture also do this.

As discussed in Section 3.3, we minimise the changes to SDRT's original constraints on the parts of a discourse context to which new material can rhetorically connected—the so-called notion of availability. In other words, the available labels in embodied discourse are those on the right frontier of at least one last label:

Definition 8 Availability for Multimodal Discourse

Let $S = \langle A, F, \mathit{last}\rangle$ be an SDRS for multimodal discourse (and so by Definition 3, *last* is a non-empty set of at most two labels). Where $\pi, \pi' \in A$, we say that $\pi > \pi'$ iff either π immediately outscopes π' or there is a label $\pi'' \in A$ such that $F(\pi'')$ contains the literal $R(\pi, \pi')$ for some subordinating rhetorical relation R (e.g.,

Elaboration or *Explanation* but not *Narration*). Let $>^*$ be the transitive closure of $>$. Then $\pi \in A$ is available in S iff $\pi \geq^* l$, where $l \in last$.

In keeping with our strategy for minimising changes that are engendered by gesture, the constraints on anaphora match exactly SDRT's constraints for purely linguistic discourse:

Definition 9 Antecedents to Anaphora

Suppose that K_β contains an anaphoric condition φ. Then the *available antecedents* to φ are terms that are:

1. in K_β and DRS-accessible to φ (as defined in Kamp and Reyle (1993)); or
2. in K_α, DRS-accessible to any sub-formula in K_α, and there is a formula $R(\alpha, \gamma)$ in the SDRS such that $\gamma \succeq \beta$.

In other words, an antecedent must be in the same discourse unit as the anaphor, or accessible in a distinct unit that is rhetorically connected to a unit that contains the anaphor.

To illustrate this formalism, we give the precise semantics for some key discourses that guided its development. We start with the discourse (5); its SDRS is shown in (5′)—with some simplification, since we have ignored tense and presuppositions:

(5)
a. π_1: and took his hatchet
π_2: Speaker's right hand grasps left hand, with wrists bent.
π_3: Speaker lifts poised hands above right shoulder.

b. π_4: and with a mighty sweep
π_5: Hands, held above right shoulder, move back then forward slightly.

c. π_6: SLICED the wolf's stomach open
π_7 : Arms swing in time with sliced; then are held horizontally at the left

(5′) $\langle\{\pi_0, \pi, \pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6, \pi_7\}, F, \{\pi_6, \pi_7\}\rangle$, where F is as follows:

π_1: $\exists hw[took(e_1, w, h) \wedge hatchet(h)]$

π_2: $[\mathcal{G}](\exists lra[left\_hand(l, w) \wedge right\_hand(r, w) \wedge handle(h, a) \wedge grab(e_2, w, a) \wedge instrument(e_2, l) \wedge instrument(e_2, r)])$

π_3: $[\mathcal{G}](\exists d[right\_shoulder(d, w) \wedge lift(e_3, w, h) \wedge goal\_location(e_3, d) \wedge instrument(e_3, l) \wedge instrument(e_3, r)])$

π_4: $\exists s[sweep(s) \wedge mighty(s) \wedge with(e_4, s)]$

π_5: $[\mathcal{G}](coiled\_backswing(e_5, w) \wedge instrument(e_5, l) \wedge instrument(e_5, r))$

π_6: $\exists ft[slice\text{-}open(e_4, w, t) wedgestomach(t, f) \wedge wolf(f)]$

π_7: $[\mathcal{G}][slice\text{-}open(e_4, w, t)])$

π_0: $Elaboration(\pi_1, \pi) \wedge Narration(\pi_1, \pi_4) \wedge Explanation(\pi_4, \pi_5) \wedge Replication(\pi, \pi_5) \wedge Narration(\pi, \pi_5) \wedge Background(\pi_4, \pi_6) \wedge Depiction(\pi_6, \pi_7) \wedge Replication(\pi_5, \pi_7) \wedge Narration(\pi_5, \pi_7)$

π: $Replication(\pi_2, \pi_3)$

This logical form is built from the underspecified content that's revealed by linguistic and gestural form via complex pragmatic inference. In fact, resolving the underspecified content and identifying the rhetorical connections are logically co-dependent. We will discuss such inference in Section 5. Here, we simply explain (5′)'s dynamic semantic interpretation. First, observe the referential connection between the initial grasping gesture and the clause in (5a): the woodsman w and the hatchet h are bound by quantifiers in the content $F(\pi_1)$ of the

linguistic component of (5a). But the dynamic semantics of *Explanation*(π_1, π) (and the value of $F(\pi)$, which outscopes the content $F(\pi_2)$ of the gesture) then ensures that the functions f and g in the input context $\langle f, g\rangle$ for interpreting the gesture $F(\pi_2)$ assign the same values to w and h as is used to satisfy the body of the formula $F(\pi_1)$—the speech and gesture are about the same woodsman and hatchet. Furthermore, by Definition 9, w and h are available antecedents for the bridging references to the woodsman's hands l and r and the handle a of the hatchet that form part of the content $F(\pi_2)$ of the gesture. Similarly, the continued references to these individuals throughout the rest of the gestures is licensed by the sequence of Replication relations connecting π_2 to π_3 and then to π_5 and finally to π_7 (and these rhetorical connections are licensed by Definition 8). These connections also entail that all the gestures depict the same mimicry—here, the woodsman's embodied actions in attacking the wolf with the hatchet. The fragments of speech also rhetorically connect together: the first clause π_1 describes a first event; the next adjunct π_4 continues the narrative, indicating that the sweep immediately follows the taking described in π_1; the Background relation between π_4 and π_6 entails that the sweep is part of the action that accomplishes the slicing. Finally, we have an additional layer of rhetorical connections that describe the interaction of gesture and speech. We assume that the two gestures in π_2 and π_3 show how the woodsman takes his hatchet: by grabbing the handle with his hands and hoisting it over his right shoulder. Then, we assume that the coiled backswing demonstrated in gesture π_5 shows how the woodsman is able to deliver such a mighty swing—so this gesture serves as an Explanation of the synchronous speech. The final "slicing" gesture of π_7 is a direct Depiction of the event described in the utterance segment π_6 that accompanies it, and the Narration connection to the gesture π_5 entails that the slicing happens after the coiled backswing. Again, Definition 8 makes this rhetorical structure possible.

Now consider an example with identifying spatial reference:

(12) [These things] push up the pins.
The speaker points closely at the frontmost wedge of the line of jagged wedges that runs along the top of a key as it enters the cylinder of a lock.

(12′) $\pi_0 : \exists sp(\mathit{things}(s) \wedge \mathit{pins}(p) \wedge \mathit{push\_up}(e, s, p)) \wedge$
$[\mathcal{G}]\exists w(\mathit{exemplifies}(w, s) \wedge \mathit{loc}(e, w, v_I(\vec{p}_w)))$

The construction rules for multimodal utterances introduce an underspecified anaphoric condition for *these things*, and an underspecified condition *id_rel* for relating the denotation of *these things* to the referent of the synchronous identifying gesture. Resolving *id_rel* to *exemplifies*, identifying the denotation of *these things* to the set of wedges s on the top surface of the key, identifying the demonstrated object w as the frontmost wedge, and identifying the gesture as directly demonstrating a copresent object in real space (so that the spatial mapping is v_I) are all logically co-dependent tasks. The individual w that's referenced in the identifying gesture but not in synchronous speech is outscoped by $[\mathcal{G}]$; this predicts the anomaly of continuing (12) with, e.g., *??It has the right height* via Definition 5.

The logical form (14′) of example (14) formalises the metaphorical use of spatial reference. We divide the speech into two segments: π_1 labels "we have one ball"; and π_2 elaborates the speaker's act in providing this information—it's something Susan has already said.

(14) We have this one ball, as you said, Susan.
The speaker sits leaning forward, with the right hand elbow resting on his knee and the right hand held straight ahead, in a loose ASL-L gesture (thumb and index finger extended, other fingers curled) pointing at his addressee.

(14′) $\pi_1 : \exists wb(we(w) \wedge have(e, w, b) \wedge one(b) \wedge ball(b))$
$\pi_2 : \exists us(susan(s) \wedge said(e', u, s))$
$\pi_3 : [\mathcal{G}]classify(e'', u, v_m(\vec{p}_i))$
$\pi : Depiction(\pi_2, \pi_3)$
$\pi_0 : Elaboration_*(\pi_1, \pi)$

The gesture π_3 offers a metaphorical depiction of "as you said Susan": it classifies the speaker's utterance u as associated with the virtual space of Susan's contributions. In fact, given the illocutionary effects of $Elaboration_*(\pi_1, \pi)$ (formal details of which we omit here), it is satisfied only if the content u of what Susan said entails K_{π_1}.

4 Underspecified Meaning for Gesture

The logical forms presented in Section 3 capture specific interpretations in context. These result from inference that reconciles the abstract meanings that are revealed by linguistic and gestural forms with overarching constraints on coherent communication and commonsense background knowledge. In this section, we formalise the abstract level of gesture meaning that's revealed by its form.

The formalisation follows (Kendon, 2004, Kopp, Tepper and Cassell, 2004, McNeill, 2005) in locating iconicity and deixis within individual aspects of the form of a gesture. For example, Kendon (2004) finds interpretive generalisations across related gestures with similar handshapes or hand orientations. Kopp et al. (2004) describe *image description features* that offer an abstract representation of the iconic significance of a wider range of form features. Section 4.1 reviews how the descriptive literature analyses gestures as complexes of form features, and gives a formal realisation.

Section 4.2, meanwhile, formalises the significance of these form features as constraints on interpretation. We emphasise that principles of gesture meaning such as iconicity cannot be formalised transparently at the level of truth-conditional content. Consider, for example, the interpretive effect of holding the right hand in a fist while performing a gesture. The fist itself might depict a roughly spherical object located in a virtual space. For example, McNeill (1992, Ex 8.3 p 224) offers a case where a speaker narrating the events of a cartoon uses a fist to depict a bowling ball. Alternatively, the fist might mirror the described pose of a character's hand as a fist. Threatening a punch is such a case, as when Jackie Gleason, playing the role of Ralph on *The Honeymooners*, holds up a fist to his wife and announces "You're going to the moon!". Finally, the fist can depict a grip on a (perhaps abstract) object, as in the woodsman's grip on the hatchet in (5) or our metaphorical understanding of low-level processes as carrying speech errors with them in (9). Logically, the different cases involve qualitatively different relationships, with different numbers and sorts of participants; so the iconicity shared by all these examples must reside in an abstract *description* of iconic meaning, rather than any specific iconic content shared by all the cases.

Finally, in Section 4.3, we formalise the additional constraints on interpretation that emerge when gesture and speech are used together in synchrony. Our formalism represents these constraints through abstract, underspecified relationships that connect content across modalities.

The semantic constraints contributed by iconicity, deixis, and synchrony describe logical form and provide input to the processes of establishing discourse coherence described in Section 5. The semantic constraints identify a range of alternative possible specific interpre-

tations. Recognising why the communicative action is coherent and identifying which of the possible specific interpretations are pragmatically preferred are then logically co-dependent tasks.

4.1 The Form of Gesture

By the *form* of gesture, we mean the cognitive organisation that underlies interlocutors' generative abilities to produce and recognise gestures in an unbounded array. This definition shows a clear parallel to natural language grammar, and we build on that parallel throughout. But the differences between gesture form and natural language syntax are also important. In particular, gesture seems to lack the arbitrary correspondence between the order of actions and their interpretation as an ensemble, as mediated by a hierarchical formal structure, which is characteristic of natural language syntax (McNeill, 1992).

Instead, gesture form is at heart multidimensional. A gesture involves various features of performance—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 of movement. These form features are interpreted jointly; not through arbitrary 'syntactic' conventions, but through creative reasoning about the principles of iconicity, deixis, and coherence. Following Kopp, Tepper and Cassell (2004), we represent this multidimensionality by describing the form of each gesture stroke with a feature structure. The feature structure contains a list of attribute–value pairs characterising the physical makeup of the performance of the gesture. For example, we represent the form of the right-hand gesture identifying Norris Hall in (1) with the feature structure (17).

(17)
$$\begin{bmatrix} \textbf{identifying-gesture} \\ \text{right-hand-shape} : \textit{loose-asl-5-thumb-open} \\ \text{right-finger-direction} : \textit{forward} \\ \text{right-palm-direction} : \textit{up-left} \\ \text{right-location} : \vec{c} \end{bmatrix}$$

Here $\vec{c}$ is the spatio-temporal coordinate in $\mathcal{R}^4$ of the tip of the right index finger (i.e., up and to the right of the speaker's shoulder) that, together with the values of the other attributes, serves to to identify the region $\vec{p}_n$ in space that is designated by the gesture. Unlike Kopp, Tepper and Cassell (2004), our representations are typed: e.g., (17) is typed **identifying-gesture**. We particularly use this to distinguish between form features that are interpreted in terms of spatial reference, like the feature *right-location* in (17), and those that are interpreted via iconicity, perhaps like the feature *right-hand-shape* in (17). Kendon (2004, Ch. 11) observes that hand shape in pointing gestures often serves as an indication of the speaker's communicative goal in demonstrating an object—distinguishing, presenting, orienting, directing attention—through a broadly metaphorical kind of meaning-making.

The organisation of gesture is recognised or constructed by our perceptual system. So parsing a non-verbal signal into a contextually appropriate description of its underlying form is a highly complex task where many ambiguities must be resolved—or left open—just as in parsing language. We can see this by recapitulating the ambiguity in form associated with the iterative circling gesture of (9).[9] One account of its form is (18):

[9]Note that some of the values are expressed as *sets* (e.g., the movement direction). This allows us to capture

(18)
$$\left[\begin{array}{l}\textbf{qualitative-characterising-gesture}\\ \text{right-hand-shape} : \textit{asl-a}\\ \text{right-finger-direction} : \textit{down}\\ \text{right-palm-direction} : \textit{left}\\ \text{right-trajectory} : \textit{sagittal-circle}\\ \text{right-movement-direction} : \{\textit{iterative}, \textit{clockwise}\}\\ \text{right-location} : \textit{central-right}\end{array}\right]$$

This treats the hand movement in (9) as one stroke and it abstracts away from the *number* and *exact spatial trajectory* followed in each circling of the hand. We might also analyse this gesture as a composition of several identical strokes. Furthermore, the repetition of the movement is captured in (18) via the value *iterative*; another licensed representation lacks this value but instead makes *trajectory* be exactly two sagittal circles. Finally, this gesture has a licensed representation whose values are spatiotemporal coordinates as in (17) rather than qualitative as in (18) (and accordingly it will have a distinct root type); this form, in contrast to (18), yields spatial constants in semantics. Our theory tolerates and indeed welcomes such ambiguities.

Similarly, we regard synchrony as an underlying perceptual judgement about the relationship between gesture and speech. Because we regard form as an aspect of perceptual organisation, we do not need to assume that perceived synchrony between speech and gesture necessarily involves strict temporal constraints. In fact, there is no clear answer about the conditions required for a gesture to be perceived as synchronous with a linguistic phrase (Oviatt, DeAngeli and Kuhn, 1997, Sowa and Wachsmuth, 2000, Quek et al., 2002). Interlocutors' judgements are influenced by the relative time of performance of the gesture to speech, the type of syntactic constituent of the linguistic phrase (and possibly the type of gesture), prosody, and perhaps other factors. We remain neutral about these details.

4.2 The Meaning of Gesture

We formalise gesture meaning using the technique of *underspecification* from computational semantics (Egg, Koller and Niehren, 2001). With underspecification, knowledge of meaning determines a *partial description* of the LF of an utterance. This partial description is expressed in a distinct language $\mathcal{L}_{ulf}$ from the language $\mathcal{L}_{sdrs}$ of LFs. Each model M for $\mathcal{L}_{ulf}$ corresponds to a unique formula in $\mathcal{L}_{sdrs}$, and M satisfies $\phi \in \mathcal{L}_{ulf}$ if and only if ϕ (partially) describes the unique formula corresponding to M. Semantic underspecification languages are typically able to express partial information about semantic scope, anaphora, ellipsis and lexical sense; e.g., Koller, Mehlhorn and Niehren (2000). For example, pronoun meaning will stipulate that the LF must include an equality between the discourse referent that interprets the pronoun and some other discourse referent that's present in the LF of the context, but it will not stipulate *which* contextual discourse referent this is. Thus the partial description is compatible with several LFs: one for each available discourse referent in the LF of the context. Such partial descriptions are useful for ensuring that the grammar neither under-determines nor over-determines content that's revealed by form.

generalisations over clockwise movements on the one hand (iterative or not), and iterative movements on the other (where *iterative* can represent a finite repetition of movement, as in this case). More generally, if features *change* during a stroke, we can specify feature values as *sequences* as well.

We illustrate the technical resources of underspecification with a sentence (19a) whose syntax underdetermines semantic scope. In a typical underspecified representation, so-called Minimal Recursion Semantics (MRS, Copestake et al. (1999)), the description given in (19b) underspecifies semantic scope: (i) each predication is labelled (l_1, etc); (ii) scopal arguments are holes (h_1, h_2 etc); (iii) there are scope constraints ($h \geq l$ means that h outscopes l's predication); and (iv) the constraints admit two ways of equating holes with labels.

(19) a. Every black cat loved some dog.

b. $l_1 : \_every\_q(x, h_1, h_2)$
$l_2 : \_black\_a\_1(e_1, x)$
$l_2 : \_cat\_n\_1(x)$
$l_3 : \_loved\_v\_1(e_2, x, y)$
$l_4 : \_some\_q(y, h_3, h_4)$
$l_5 : \_dog\_n\_1(y)$
$h_2 \geq l_2$, $h_3 \geq l_5$

c. $l_1 : a_1 \_every\_q(x)$, $RESTR(a_1, h_1)$, $BODY(a_1, h_2)$
$l_{21} : a_{21} : \_black\_a\_1(e_1)$, $ARG1(a_{21}, x_1)$
$l_{22} : a_{22} : \_cat\_n\_1(x_2)$
$l_3 : a_3 : \_loved\_v\_1(e_2)$, $ARG1(a_3, x_3)$, $ARG2(a_3, y_1)$
$l_4 : a_4 : \_some\_q(y)$, $RESTR(a_4, h_3)$, $BODY(a_4, h_4)$
$l_5 : a_5 : \_dog\_n\_1(y_2)$
$h_2 \geq l_2$, $h_3 \geq l_5$
$x = x_1$, $x = x_2$, $x = x_3$, $x_1 = x_2$, $x_2 = x_3$, $y = y_1$, $y = y_2$, $y_1 = y_2$, $l_{21} = l_{22}$

Intersective modification is achieved by sharing labels across predications (e.g., l_2 in (19b)), meaning that in any fully specific LF $\_black\_a\_1(e_1, x)$ and $\_cat\_n\_1(x)$ are connected with logical conjunction. Observe also the naming convention for the predicate symbols, based on word lemmas, part-of-speech tags and sense numbers. Our approach to gesture also leverages this ability to regiment the links between form and meaning.

An extension of these ideas is explored in Robust Minimal Recursion Semantics (RMRS, Copestake (2003))—RMRS can also underspecify the *arity* of the predicate symbols and *what sorts of arguments* they take. Since the iconic meaning of gesture constrains, but doesn't fully determine, all these aspects of interpretation, we adopt RMRS as the underlying semantic formalism $\mathcal{L}_{ulf}$. RMRS is fully compatible with the language $\mathcal{L}_{sdrs}$ of SDRT (Asher and Lascarides, 2003, p. 122)—indeed, SDRT's existing glue logic supports any description language, and so can construct from the RMRSs of discourse units an SDRS (or a set of them if ambiguities persist) that captures the pragmatic interpretation (we don't use the specific description language from Asher and Lascarides (2003) here because it doesn't underspecify arity).

The RMRS corresponding to (19b) is (19c): RMRSs offer a more factorised representation where the base predicates are unary and the other arguments are represented by separate binary relations on the unique *anchor* of the relevant predicate symbols ($a_1, a_2, \ldots$) together with variable and label equalities (e.g., $x = x_1$, $l_{21} = l_{22}$). This factored representation allows one to build semantic components to shallow parsers, where lexical or syntactic information that contributes to meaning is absent. An extreme example would be a part-of-speech (POS) tagger: one can build its semantic component simply by deriving lexical predicate symbols from the word lemmas and their POS tags, as given in (20):

(20) a. Every_AT1 black_JJ cat_NN1 loved_VVD some_DD dog_NN1

b. $l_1 : a_1 : \_every\_q(x)$,
$l_{21} : a_{21} : \_black\_a(e_1)$,
$l_{22} : a_{22} : \_cat\_n(x_2)$
$l_3 : a_3 : \_loved\_v(e_2)$,
$l_4 : a_4 : \_some\_q(y)$,
$l_5 : a_5 : \_dog\_n(y_2)$

Semantic relations, sense numbers and the arity of the predicates are missing from (20b) because the POS tagger doesn't reveal information about syntactic constituency, word sense or lexical subcategorisation. But the RMRSs (19c) and (20b) are entirely compatible, the former being more specific than the latter. In particular, the model theory of *rmrs* restricts the possible denotations of *_lemma_tag_sense* (which are all constructors in the fully specific language $\mathcal{L}_{sdrs}$) to being a subset of those of *_lemma_tag*.

To regiment the interpretation of gesture formally in *rmrs*, we assume that interpretive constraints apply at the level of form features. Each attribute–value element yields an *rmrs* predication, which must be resolved to a formula in the LF of gesture in context.[10] If the element is interpreted by *iconcity*, we constrain the resolution to respect possibilities for depiction. If the element is interpreted by *spatial reference*, we interpret it as locating an underspecified individual via an underspecified correspondence to the physical point in space-time that is designated by that feature of the gesture. We treat attributes with set values (e.g., the movement-direction attribute in (18)) like intersective modification in language (e.g., *black cat* in (19b)). This captures the intuition that the different aspects of the direction of movement in (18) must all depict properties of the same thing in interpretation.

So, more formally, each iconic attribute value pair introduces an underspecified predication that corresponds directly to it; for instance the hand shape in (18) introduces the predication (21) to the RMRS for this gesture:

(21) $l_1 : a_1 : right\_hand\_shape\_asl\text{-}a(i_1)$

Here, l_1 is a unique label that underspecifies the scope of the predication; a_1 is the unique anchor that provides the locus for specifying the predicate's arguments; i_1 is a unique *metavariable* that underspecifies the sort of the main argument of the predication (it could be an individual object or an eventuality); and *hand_shape_asl-a* underspecifies reference to a property that the entity i_1 has and that can be depicted through the gesture's fist shape.

We then represent the possible resolutions of the underspecified predicates via a hierarchy of increasingly specific properties, as in Figure 4. The hierarchy of Figure 4 captures the metaphorical contribution of the fist to the depiction of the process in (9), by allowing *right_hand_shape_asl-a* to depict a holding event, metaphorically interpreted as the event e of a process x *sustaining* errors y in speech production ("bearing them with it", as it were). Following Copestake and Briscoe (1995), this treats metaphorical interpretations as a specialisation of the underspecified predicate symbol that's produced by form, as opposed to coercion on a specific literal interpretation of a word that in turn contradicts information in the context; e.g., Hays and Bayer (2001). We have argued elsewhere for treating metaphor in

[10] This suggests that the form of an iconic gesture is like a bag of words. Kopp, Tepper and Cassell (2004) liken it to a bag of *morphemes*, on the grounds that the resolved interpretations of the features can't be finitely enumerated. But word senses can't be enumerated either (Pustejovsky, 1995); hence our analogy with words is also legitimate.

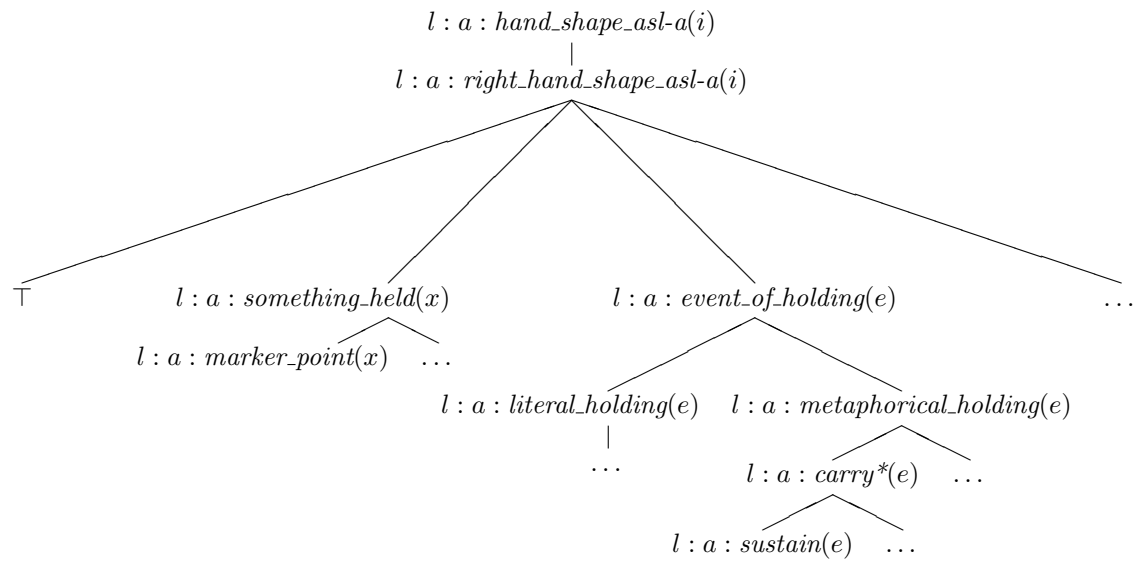

Figure 4: Possible resolutions for *hand_shape_asl-a*.

linguistic discourse in this way (Asher and Lascarides, 1995), and choose to treat metaphor in gesture in the same way so as to maintain as uniform a pragmatics for speech and gesture as possible.

At the same time, we can capture the contribution of a fist as depicting something held by resolving *right_hand_shape_asl-a* accordingly; e.g., if the gesture in (9) were to accompany the utterance "the mouse ran round the wheel", then the underspecified predicate symbol would resolve to a *marker-point* x indicating a designated location on the mouse's spinning wheel. Finally, all underspecified predications are resolvable to validity ($\top$), since any form feature in a gesture may contribute no meaning in context (e.g., the clockwise motion in (9)). We assume that resolving *all* predications to $\top$ is pragmatically dispreferred compared with logically contingent resolutions (see Section 5). Underspecified predicates may also share certain specific resolutions: e.g., *marker-point* is also one way of resolving the underspecified predicate corresponding to a flat hand—*hand_shape_asl-5*.

Figure 4 reflects the fact that, like all dimensions of iconic gesture, the fist shape underspecifies how many entities it relates in its specific semantic interpretation. The predicates in Figure 4 vary in the number of arguments they take and the factorised notation of RMRS lets us express this. For example, *sustain* is a 3-place relation and so $l : a : sustain(e)$ entails $l : a : sustain(e)$, $ARG1(a, x)$, $ARG2(h, y)$ for some x and y, while *marker-point* is a 1-place property, and therefore $l : a : marker\text{-}point(x)$, $ARG1(a, y)$ is unsatisfiable.

Figure 4 represents a special kind of commonsense background knowledge; namely, general possibilities for iconic representation. Technically, the interpretation of *hand_shape_asl-a* is not defined at all with respect to the dynamic semantics given in Definition 5 because it is not a part of the language $\mathcal{L}_{sdrs}$ of fully specific logical forms; rather, the distinct and static model theory for RMRS ensures that *hand_shape_asl-a* denotes a constructor from $\mathcal{L}_{sdrs}$ or a combination thereof that is compatible with the hierarchy in Figure 4. More

informally, you can compare predications like *hand_shape_asl-a* to Kopp et al.'s (2004) *image description features*—an abstract representation that captures gesture meaning. While some of the leaves in this hierarchy correspond to fully specific interpretations, others represent vague ones. (Following Kopp, Tepper and Cassell (2004), we do not believe that the specific interpretations that are licensed by a (unique) underspecified semantic representation can be finitely enumerated.) 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.

Let's now consider the compositional semantics of a spatially-locating component of gesture meaning. In words, (22) states that x denotes an individual which is spatially located at the coordinates $v(\vec{p})$, where $\vec{p}$ is the physical location actually designated by the gesture, and v is a mapping from this physical point to the space depicted in meaning (with the value of v being resolved through discourse interpretation).

(22) $l_2 : a_2 : \mathit{sp\_ref}(e),\ \mathit{ARG1}(a_2, x),\ \mathit{ARG2}(a_2, v(\vec{p}))$

The predicate *sp_ref* in (22) can resolve to the constructor *loc* or to the constructor *classify* in $\mathcal{L}_{sdrs}$, with its dynamic semantics defined in Section 3.1. The constant $\vec{p}$ in (22) is determined as a complex function of the grounding of gesture form in physical space. For instance, a pointing gesture (with hand shape 1-index) will make $\vec{p}$ denote a cone whose tip is the *hand-location* coordinate $\vec{c}$, with the cone expanding out from $\vec{c}$ in the same direction as the value of *finger-direction* (Kranstedt et al., 2006).

Finally, in Section 3.2 we motivated the introduction of an operator $[\mathcal{G}]$ for each stroke, which must outscope the content conveyed by the gesture's form features. This was necessary for constraining co-reference. We translate each gesture overall using an instance of this operator, constrained to outscope each predication that's contributed by its form features. Thus the (underspecified) content arising from the form of the visualising gesture in (18) is the RMRS (23):

(23) $l_0 : a_0 : [\mathcal{G}](h)$
$l_1 : a_1 : \mathit{right\_hand\_shape\_asl\text{-}a}(i_1)$,
$l_2 : a_2 : \mathit{right\_finger\_dir\_down}(i_2)$,
$l_3 : a_3 : \mathit{right\_palm\_dir\_left}(i_3)$,
$l_4 : a_4 : \mathit{right\_traj\_sagittal\_circle}(i_4)$,
$l_5 : a_{51} : \mathit{right\_move\_dir\_iterative}(i_5)$,
$l_5 : a_{52} : \mathit{right\_move\_dir\_clockwise}(i_5)$,
$l_6 : a_6 : \mathit{right\_loc\_central\text{-}right}(i_6)$,
$h \geq l_j$, for $1 \leq j \leq 6$

To handle identifying gestures, we add an overall layer of quantificational structure as in (24) so that we model the gesture as identifying an appropriate entity x.

(24) $l_0 : [\mathcal{G}](h_1)$,
$l_1 : a_1 : \mathit{deictic\_q}(x),\ \mathit{RESTR}(a_1, h_2),\ \mathit{BODY}(a_1, h_3)$
$l_2 : a_2 : \mathit{sp\_ref}(e),\ \mathit{ARG1}(a_2, x),\ \mathit{ARG2}(a_2, v(\vec{p}))$
$h_1 \geq l_2,\ h_2 \geq l_2$

More generally, anywhere the context-resolved interpretation of a gesture introduces an individual that is not co-referent with any individual in the synchronous speech (by the bridging

inference described in Section 3.2), then we must have a quantifier and inference relation to introduce this individual, outscoped by $[\mathcal{G}]$ so that availability (Definition 8) and semantic interpretation (Definition 5) constrain anaphoric dependencies correctly—i.e., the gestured individual cannot be an antecedent to a pronoun in subsequent speech. This scopal constraint can be expressed as part of discourse update—the logic that builds a specific LF of discourse from the underspecified logical forms of its units—although we forego details here.

Mapping the syntactic representation of gestures such as (18) to their unique RMRS (23) is very simple. Computing the predications from each attribute value pair in (18) involves exactly the same techniques as used by Copestake (2003) to build the semantic component of a POS tagger. Adding the scopal predication $[\mathcal{G}]$ and its $\geq$-conditions is triggered by the gestural-type of the feature structure: e.g., **qualitative-characterising-gesture** introduces $[\mathcal{G}]$ via scopal modification as defined in the semantic algebra for RMRS (Copestake, Lascarides and Flickinger, 2001).

Our representations of gesture meaning are analogous, both formally and substantively, to the underspecified meanings that computational semanticists already use to represent language. In particular, as we show in Section 5, we can therefore build reasoning mechanisms that combine information from language and gesture to derive integrated logical forms for multimodal discourse. Differences remain across modalities, however, in the kinds of semantic underspecification that are present in the RMRS representation of a phrase versus that of gesture. A complete and disambiguated syntactic representation of a linguistic phrase fully specifies predicate argument structure. But gestures lack hierarchical syntactic structure and individual form features, unlike words, don't fix their subcategorisation frames. Consequently, a complete and disambiguated syntactic analysis of gesture underspecifies all aspects of content, including predicate argument structure.

4.3 Combining Speech and Gesture in the Grammar

Like Kopp, Tepper and Cassell (2004), we believe that we need a formal account of the integration of speech and gesture that directly describes the organisation of multimodal communicative actions into complex units that contribute to discourse. Kopp et al. argue for this on the grounds of *generation*; our motivation stems from issues in *interpretation*. First, our observations about the relative semantic scope of negation and the content depicted by the gesture in (6) suggests that scope bearing elements introduced by language can outscope gestured content. It is very straightforward to derive such a semantics from a single derivation of the structure of a multimodal utterance: use a construction rule to combine the gesture with *computery* and a further construction rule to combine the result with *not*. Standard methods for composing semantic representations from syntax—e.g., Copestake, Lascarides and Flickinger (2001)—would then make the (scopal) argument of the negation outscope both the atomic formula $computery(x)$ and the gesture modality $[\mathcal{G}]$, as required. Of course, other analyses may be licensed by the grammar: for instance, the gesture might combine with the phrase *not computery*. (We don't address the resolution of such ambiguities of form here.)

Secondly, we assume, as is standard in dynamic semantic theories, that discourse update has access to semantic representations but no direct access to form. But synchrony is an aspect of form that conveys meaning, and consequently we need to give a formal description of this form–meaning relation as part of utterance structure. For instance, we suggested in Section 2 that the content of a characterising gesture must be related to its synchronous linguistic phrase with one of a subset of the full inventory of rhetorical connections (for instance, Disjunction is

excluded). This means that the synchrony that connects speech and gesture together conveys semantic information that's similar to that conveyed by a highly sense-ambiguous discourse connective or a free adjunct in language: they both introduce rhetorical connections between their syntactic complements, but do not fully specify the value of the relation. We must represent this semantic contribution that's revealed by form.

For identifying gestures, meanwhile, synchrony identifies which verbally-introduced individual y is placed in correspondence with the individual x designated in gesture . Moreover, synchrony serves to constrain the relationship between x and y. Sometimes the relationship is equality but not always—as in *these things* said while pointing to an exemplar (see (12)). We treat the semantic relationship as underspecified but not arbitrary (Nunberg, 1978).

While we don't give details here, we assume a unification or constraint–based representation of utterance structure, following Johnston (1998). Construction rules in this specification describe the form and meaning of complex communicative actions including both speech and gesture. We assume that the construction rule for combining a characterising gesture and a linguistic phrase contributes its own semantics. In other words, as well as the daughters' RMRSs being a part of the RMRS of the mother node (as is always the case), the construction rule introduces the new predication (25) to the mother's RMRS, where h_s is the top-most label of the content of the 'speech' daughter, and h_g the top-most label of the gesture:

(25) $\quad l : a : \mathit{vis\_rel}(h_s),\ \mathit{ARG1}(a, h_g)$

The underspecified predicate *vis_rel* must then be resolved via discourse update to a specific rhetorical relation, where we assume at least that Disjunction is not an option.

Similarly, we assume that the construction rule that combines an identifying gesture with an NP contributes its own semantics: the labelled predication (26), where l_2 is the label of the spatial condition *sp_ref* introduced by the RMRS of the gesture (see (24)) and y is the semantic index of the NP.

(26) $\quad l_2 : a_{21} : \mathit{id\_rel}(x),\ \mathit{ARG1}(a_{21}, y)$

So, for instance, the RMRS for a multimodal constituent consisting of the NP *these things* combined with the pointing gesture in (12) is the following:

(27)
$$
\begin{array}{l}
l_0 : [\mathcal{G}](h_1), \\
l_1 : a_1 : \mathit{deictic\_q}(x),\ \mathit{RESTR}(a_1, h_2),\ \mathit{BODY}(a_1, h_3) \\
l_2 : a_2 : \mathit{sp\_ref}(e),\ \mathit{ARG1}(a_2, x),\ \mathit{ARG2}(a_2, v(\vec{p})) \\
l_2 : a_{21} : \mathit{id\_rel}(x),\ \mathit{ARG1}(a_{21}, y) \\
l_3 : a_3 : \mathit{\_these\_q}(y),\ \mathit{RESTR}(a_3, h_4),\ \mathit{BODY}(a_3, h_5), \\
l_4 : a_4 : \mathit{\_things\_n\_1}(y) \\
h_1 \geq l_2,\ h_2 \geq l_2,\ h_4 \geq l_4
\end{array}
$$

Identifying gestures that combine with other kinds of linguistic syntactic categories, such as PPs and VPs are also possible in principle, although we leave the details to future work.

5 Establishing Coherence through Default Inference

SDRT describes which possible ways of resolving an underspecified semantics are pragmatically preferred. This occurs as a byproduct of *discourse update*: the process by which one constructs the logical form of discourse from the (underspecified) compositional semantics

of its units. So far, SDRT's discourse update has been used to model linguistic phenomena. Here, we indicate how it can resolve the underspecified meaning of both language and gesture.

Discourse update in SDRT involves nonmonotonic reasoning and is defined in a constraint-based way: Where $\phi \in \mathcal{L}_{ulf}$ represents the (underspecified) old content of the context, which is to be updated with the (underspecified) new content $\psi \in \mathcal{L}_{ulf}$ of the current utterance, the result of update will be a formula $\chi \in \mathcal{L}_{ulf}$ that is entailed by both ϕ, ψ and the consequences of a nonmonotonic logic—known as the *glue logic*—when ϕ and ψ are premises in it. This constraint-based approach allows an updated interpretation of discourse to exhibit ambiguities. This is well established in the literature for being necessary for linguistic discourse; here, we observed via examples like (1), (7) and (9) that it is necessary for gesture as well.

The glue logic axioms specify which speech act $\lambda : R(\alpha, \beta)$ was performed, given the content and context of utterances. And being a nonmonotonic consequence of the glue logic, $\lambda : R(\alpha, \beta)$ becomes part of the updated LF. Formally, the glue logic axioms typically have the shape schematised below, where $A > B$ means *If A then normally B* (note that without loss of generality, we omit anchors when the arguments are specified; so for instance $\lambda : R(\alpha, \beta)$ is a notational variant of $\lambda : a : R(\alpha)$, $ARG1(a, \beta)$):

- `Glue Logic Schema:` $(\lambda :?(\alpha, \beta) \wedge$ some stuff$) > \lambda : R(\alpha, \beta)$

In words, this axiom says: if β is to be connected to α with a rhetorical relation, and the result is to appear as part of the logical scope labelled λ, but we don't know what the value of that relation is yet, and moreover "some stuff" holds of the content labelled by α and β, then normally the rhetorical relation is R. The "some stuff" is derived from the (underspecified) logical forms (expressed in $\mathcal{L}_{ulf}$) that α and β label (in our case, this language is that of RMRS), and the rules are justified either on the basis of underlying linguistic knowledge, world knowledge, or knowledge of the cognitive states of the dialogue agents.

For example, the glue logic axiom `Narration` stipulates that one can normally infer Narration if the constituents that are to be rhetorically connected describe eventualities which are in an occasion relation. That is, there is a 'natural event sequence' such that events of the sort described by α lead to events of the sort described by β:

- `Narration:` $\lambda :?(\alpha, \beta) \wedge occasion(\alpha, \beta)) > \lambda : Narration(\alpha, \beta)$.

Schank's (1977) scripts attempted to capture information about which eventualities occasion which others; in SDRT such scripts are default axioms. For example, we assume that the underspecified logical forms of the clauses in (16) verify the antecedents of a default axiom whose consequence is $occasion(\pi_1, \pi_2)$, yielding $\pi_0 : Narration(\pi_1, \pi_2)$ via `Narration`:

(16) π_1. John went out the door.

π_2. He turned right.

Indeed, the glue logic axioms that do this should be neutral about about sentence mood, so that they also predict that the *imperatives* in (28) are connected by Narration:

(28) Go out the door. Turn right.

In the model theory of SDRT, such a logical form for (28) entails that (a) both imperatives are commanded (because Narration is veridical), and (b) the command overall is to turn right

immediately after going out the door.[11] This is exactly the discourse interpretation we desire for the multimodal act (4) from the NUMACK corpus:

(4) You walk out the doors
The gesture is one with a flat hand shape and vertical palm, with the fingers pointing right, and palm facing outward.

So our aim now is to ensure that discourse update in SDRT supports the following two co-dependent inferences in constructing the LF of (4): (a) the contents of the clause and gesture are related by Narration; and (b) the (underspecified) content of the gesture as revealed by its form resolves to *turn right* in this context. We'll see how shortly.

Explanation is inferred on the basis of evidence in the discourse for a causal relation:

- `Explanation:` $(\lambda :?(\alpha, \beta) \wedge \mathit{cause}_D(\beta, \alpha)) > \lambda : \mathit{Explanation}(\alpha, \beta)$

Note that $\mathit{cause}_D(\beta, \alpha)$ does not entail that β actually *did* cause α; the latter causal relation would be inferred if Explanation is inferred. The formula $\mathit{cause}_D(\beta, \alpha)$ is inferred on the basis of *monotonic* axioms (monotonic because the evidence for causation is present in the discourse, or it's not), where the antecedent to these axioms are expressed in terms of the (underspecified) content of α and β. We assume that there will be such a monotonic axiom for inferring $\mathit{cause}_D(\pi_2, \pi_1)$ for the discourse (29) (we omit details), which bears analogies to the embodied utterance (9).

(29) π_1. There are low-level phonological errors which tend not to get reported.
π_2. They are created via subconscious processes.

In SDRT the inferences can flow in one of several directions. If the premises of a glue logic axiom is satisfied by the underspecified semantics derived from the grammar, then a particular rhetorical relation follows and its semantics yields inferences about how underspecified parts of the utterance and gesture contents are resolved. Alternatively, there are cases where the underspecified compositional semantics is insufficient for inferring any rhetorical relation. In this case, discourse update allows inference to flow in the opposite direction: one can resolve the underspecified content to a more specific interpretation that supports an inference to a rhetorical relation. If 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* and *quantity* of the rhetorical relations; see Asher and Lascarides (2003) for details. There may be more than one such interpretation, in which case ambiguity persists in the updated discourse representation.

Of course, this inferential flow from possible resolved interpretations of speech and gesture to rhetorical connections represents a *competence* model of discourse interpretation only. Any implementation of SDRT's discourse update would have to restrict drastically the massive search space of fully resolved interpretations that are licensed by an underspecified logical form for discourse. We have just begun to explore such issues in related work (Schlangen and Lascarides, 2002).

Let's illustrate the inference from underspecified content to complete LF with the example (9). As described in Section 4.2, the grammar yields (23) for the content of the gesture, an RMRS for the compositional semantics of the clause (which is omitted here for

[11] We did not give the model theory for imperatives in order to stay within the confines of the extensional version of SDRT in this paper. See Asher and Lascarides (2003) for details.

reasons of space), and the construction rule that combines them contributes the predication $l : \mathit{vis\_rel}(h_s, l_0)$, where h_s outscopes all labels in the RMRS of the clause, and l_0 labels the scopal modifier $[\mathcal{G}]$ in (23). Producing a fully specific LF from this therefore involves, among other things, resolving the underspecified predications in (23), and the underspecified predicate *vis_rel* must resolve to a rhetorical relation that's licensed by it—so not Disjunction.

Even though the RMRS (23) fails to satisfy the antecedent of any axiom for inferring a particular rhetorical relation, one can consider alternative ways of resolving it so as to support a rhetorical connection. One alternative is to resolve it to denote a continuous, subconscious process which sustains the phonological errors as shown in (30) where y is the low-level phonological errors introduced in the clause:

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

This particular interpretation is licensed by the predications in the RMRS (23), via hierarchies such as the one shown in Figure 4 (using y in (30) is also licensed by Definitions 8 and 9). And similarly to discourse (29), this and the compositional semantics of the clause satisfy the antecedent of an axiom whose consequent is $\mathit{cause}_D(\beta, \alpha)$. And so an Explanation relation is inferred via `Explanation`, resulting in the logical form (9′) shown earlier. As stated earlier, an alternative specific interpretation of the gesture (in fact, one that can stem from an alternative analysis of its form) entails that the gesture depicts the low level of the phonological errors. This specific interpretation would validate an inference in the glue logic that the gesture and speech are connected with Depiction (this would be on the general grounds in the glue logic that the necessary semantic consequences of a rhetorical connection are normally sufficient for inferring it). If both of these interpretations are equally coherent, then discourse update predicts that the multimodal utterance, while coherent, is ambiguous. If, on the other hand, the interpretation given in (9′) yields a more coherent interpretation (and we believe that it does because it supports additional Contrast relations with prior gestures that are in the context), then discourse update predicts this fully specific interpretation.

Finally, let's examine deictic gesture: discourse update must support inferences which resolve the underspecified relation *id_rel* between the denotations of an NP and its accompanying deictic gesture to a specific value. This is easily achieved via default axioms such as the following (we have omitted labels and anchors for simplicity):

- `Co-reference:` $(\mathit{id\_rel}(x, y) \wedge \mathit{loc}(e, y, \vec{p}) \wedge P(x) \wedge P(y)) > x = y$

In words, `Co-reference` stipulates that if x and y are related by *id_rel*, and moreover, the individual y that is physically located at $\vec{p}$ shares a property with x, then normally x and y are co-referent. Other default axioms can be articulated for inferring $\mathit{exemplifies}(x, y)$ rather than $x = y$ in the interpretation of (12).

6 Conclusion

We have provided a formal semantic analysis of co-verbal iconic and deictic gesture which captures several observations from the descriptive literature. For instance, three features in our analysis encapsulate the observation that speech and gesture together form a 'single thought'. First, the content of language and gesture are represented jointly in the same logical language. Secondly, rhetorical relations connect the content of iconic gesture to that of its synchronous speech. And finally, language and gesture are interpreted jointly within an

integrated architecture for linking utterance form and meaning. Our theory also substantiates the observation that iconic gesture on its own doesn't receive a coherent interpretation. Its form produces a very underspecified semantic representation; this must be resolved by reasoning about how it is coherently related to its context. Finally, we exploited discourse structure and dynamic semantics to account for co-reference across speech and gesture and across sequences of gestures.

One major advantage of our approach is that all aspects of our framework are already established for modelling purely linguistic discourse, and consequently we demonstrate that existing mechanisms for representing language can be exploited to model gesture as well. We showed that they suffice for a wide variety of spontaneous and naturally occurring coverbal gestures, ranging from simple deictic ones to much more complex iconic ones with metaphorical interpretations. Furthermore, our model is sufficiently formal but flexible that one can articulate specific hypotheses that can then guide empirical investigations that deepen our understanding of the phenomena.

Ultimately, we hope that the empirical and theoretical enquiries that this work enables will support a broader perspective on the form, meaning and use of nonverbal communication. This will require describing the organisation of gesture in relationship to speech, where theoretical and empirical work must interact to characterise ambiguities of form and describe how they are resolved in context. It also requires further research into other kinds of communicative action. For example, we believe that formalisms for modelling intonation and focus—e.g., Steedman (2000)—offer a useful starting point for an analysis of beat gestures. We have also ignored body posture and facial expressions. But as Krahmer and Swerts (in press) demonstrate, they can interact in complex ways not only with speech but also with hand gestures. Finally, we have focussed here almost entirely on contributions from single speakers. But in conversation social aspects of meaning are important: we need to explore how gestures affect and are affected by grounding and disputes, for instance. This is another place where empirical research such as (Emmorey, Tversky and Taylor, 2000) and formal methods such as (Asher and Lascarides, 2008) have been pursued independently and can benefit from being brought into rapport.

Authors' Addresses

Alex Lascarides,	Matthew Stone,
School of Informatics,	Computer Science,
University of Edinburgh,	Rutgers University,
10, Crichton Street,	110 Frelinghuysen Road,
Edinburgh, EH8 9AB,	Piscataway NJ 08854-8019,
Scotland, UK.	USA.
`alex@inf.ed.ac.uk`	`matthew.stone@rutgers.edu`

Acknowledgements. This work has been presented at several workshops and conferences: the Workshop on Embodied Communication (Bielefeld, March 2006); Constraints in Discourse (Maynooth, June 2006); the Pragmatics Workshop (Cambridge, September 2006), Brandial (Potsdam, September 2006); the Workshop on Dynamic Semantics (Oslo, September 2006); Dialogue Matters (London, February 2008); the Rutgers Semantics Workshop (New Jersey, November 2008). We would like to thank the participants of these for their very useful comments and feedback. Finally, we would like to thank the many individuals who have

influenced this work through discussion and feedback: Nicholas Asher, Susan Brennan and her colleagues at Stony Brook, Justine Cassell, Herb Clark, Susan Duncan, Jacob Eisenstein, Dan Flickinger, Jerry Hobbs, Michael Johnston, Adam Kendon, Hannes Rieser and his colleagues at Bielefeld, Candy Sidner, and Rich Thomason. We also owe a special thanks to Jean Carletta, who provided support in our search of the AMI corpus, including writing search scripts that helped us to find what we were looking for. Finally, we would like to thank two anonymous reviewers for this journal, for their very detailed and thoughtful comments on earlier drafts of this paper, and the editors Anna Szabolcsi and Bart Geurts. Any mistakes that remain are our own.

Much of this research was done while Matthew Stone held a fellowship in Edinburgh, funded by the Leverhulme Trust. We would also like to thank the following grants from the National Science Foundation (NSF) for their financial support: HLC-0308121, CCF-0541185, HSD-0624191.

References

Asher, N., and A. Lascarides. 1995. "Metaphor in Discourse." In *Proceedings of the AAAI Spring Symposium Series: Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity and Generativity*, 3–7. Stanford.

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

Asher, N. and A. Lascarides. 2008. "Commitments, Beliefs and Intentions in Dialogue." In *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue (Londial)*, 35–42. London.

Barker, C. 2002. "The Dynamics of Vagueness." *Linguistics and Philosophy*, 25, 1, 1–36.

Benz, A. G. Jäger, and R. van Rooij, eds. 2005. *Game Theory and Pragmatics*. Basingstoke, UK: Palgrave Macmillan.

Bittner, M.ȧ2001. "Surface composition as bridging." *Journal of Semantics*, 18, 127–177.

Buchwald, A. O. Schwartz, A. Seidl, and P. Smolensky. 2002. "Recoverability Optimality Theory: Discourse Anaphora in a Bidirectional Framework." In *Proceedings of the International Workshop on the Semantics and Pragmatics of Dialogue (EDILOG)*, 37–44. Edinburgh.

Carletta, J. 2007. "Unleashing the Killer Corpus: Experiences in Creating the Multi-Everything AMI Meeting Corpus." *Language Resources and Evaluation Journal*, 41, 2, 181–190.

Carston, R. 2002. *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford: Blackwell.

Cassell, J. 2001. "Embodied Conversational Agents: Representation and Intelligence in User Interface." *AI Magazine*, 22, 3, 67–83.

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

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

Clark, H. 1996. *Using Language*. Cambridge, England: Cambridge University Press.

Copestake, A. 2003. "Report on the Design of RMRS." Tech. Rep. EU Deliverable for Project number IST-2001-37836, WP1a, Computer Laboratory, University of Cambridge.

Copestake, A., and E. J. Briscoe. 1995. "Semi-Productive Polysemy and Sense Extension." *Journal of Semantics*, 12, 1, 15–67.

Copestake, A., D. Flickinger, I. A. Sag, and C. Pollard. 1999. "Minimal Recursion Semantics: An Introduction." Available from `http://www-csli.stanford.edu/~aac`.

Copestake, A., A. Lascarides, and D. Flickinger. 2001. "An Algebra for Semantic Construction in Constraint-based Grammars." In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL/EACL 2001)*, 132–139. Toulouse.

Cumming, S. 2007. *Proper Nouns*. Ph.D. thesis, Rutgers University.

Egg, M. A. Koller, and J. Niehren. 2001. "The Constraint Language for Lambda Structures." *Journal of Logic, Language, and Information*, 10, 457–485.

van Eijck, J. and H. Kamp. 1997. "Representing Discourse in Context." In Johan van Benthem and Alice ter Meulen, eds., *Handbook of Logic and Linguistics*, 179–237. Amsterdam: Elsevier.

Ekman, P. and W.V. Friesen. 1969. "The repertoire of nonverbal behavior: Categories, origins, usage, and coding." *semiotica*, 1, 1, 49–98.

Emmorey, K. B. 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.

Engle, R. 2000. *Toward a Theory of Multimodal Communication: Combining Speech, Gestures, Diagrams and Demonstrations in Structural Explanations*. Ph.D. thesis, Stanford University.

Farkas, D. 2002. "Specificity Distinctions." *Journal of Semantics*, 19, 1–31.

Fauconnier, G. 1997. *Mappings in Thought and Language*. Cambridge, UK: Cambridge University Press.

Gibbs, R.W. 1994. *The Poetics of Mind: Figurative Thought, Language, and Understanding*. Cambridge, UK: Cambridge University Press.

Ginzburg, J. and R. Cooper. 2004. "Clarification, Ellipsis and the Nature of Contextual Updates in Dialogue." *Linguistics and Philosophy*, 27, 297–366.

Glucksberg, S. and M.S. McGlone. 2001. *Understanding Figurative Language: From Metaphors to Idioms*. Oxford: Oxford University Press.

Goffman, E. 1963. *Behavior in Public Places*. The Free Press.

Goldin-Meadow, S. 2003. *Hearing Gesture: The Gestures we Produce when we Talk*. Cambridge, Mass: Harvard University Press.

Grice, H. P. 1975. "Logic and Conversation." In P. Cole and J. L. Morgan, eds., *Syntax and Semantics Volume 3: Speech Acts*, 41–58. New York: Academic Press.

Groenendijk, J., and M. Stokhof. 1991. "Dynamic Predicate Logic." *Linguistics and Philosophy*, 14, 39–100.

Grosz, B., A. Joshi, and S. Weinstein. 1995. "Centering: A Framework for Modelling the Local Coherence of Discourse." *Computational Linguistics*, 21, 2, 203–226.

Haviland, J. 2000. "Pointing, gesture spaces, and mental maps." In David McNeill, ed., *Language and Gesture*, 13–46. New York: Cambridge University Press.

Hays, E. and S. Bayer. 2001. "Metaphoric Generalization through Sort Coercion." In *Proceedings of the the 29th annual meeting on Association for Computational Linguistics (ACL)*, 222–228. Berkeley, California.

Heim, I. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachussetts.

Hobbs, J. R., M. Stickel, D. Appelt, and P. Martin. 1993. "Interpretation as Abduction." *Artificial Intelligence*, 63, 1–2, 69–142.

Johnston, M.i̇1998. "Unification-based Multimodal Parsing." In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and International Conference in Computational Linguistics (ACL/COLING)*. Montreal, Canada.

Johnston, M., P. R. Cohen, D. McGee, J. Pittman, S. L. Oviatt, and I. Smith. 1997. "Unification-based multimodal integration." In *ACL/EACL 97: Proceedings of the Annual Meeting of the Assocation for Computational Linguistics*. Madrid.

Kamp, H. 1981. "A Theory of Truth and Semantic Representation." In J. Groenendijk, T. Janssen, and M. Stokhof, eds., *Formal Methods in the Study of Language*, 277–322. Amsterdam: Mathematisch Centrum.

Kamp, H., and U. Reyle. 1993. *From Discourse to the Lexicon: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht, NL: Kluwer Academic Publishers.

Kendon, A. 1972. "Some Relationships between Body Motion and Speech: An Analsyis of an Example." In W. Siegman and B. Pope, eds., *Studies in Dyadic Communication*, 177–210. Oxford: Pergamon Press.

Kendon, A. 1978. "Differential perception and attentional frame: two problems for investigation." *Semiotica*, 24, 305–315.

Kendon, A. 2004. *Gesture: Visible Action as Utterance*. Cambridge, UK: Cambridge University Press.

Koller, A., K. Mehlhorn, and J. Niehren. 2000. "A Polynomial-time Fragment of Dominance Constraints." In *Proceeedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL2000)*. Hong Kong.

Kopp, S. P. Tepper, and J. Cassell. 2004. "Towards Integrated Microplanning of Language and Iconic Gesture for Multimodal Output." In *Proceedings of ICMI*. State College, PA.

Kortmann, B. 1991. *Free Adjuncts and Absolutes in English: Problems in Control and Interpretation*. Abingdon, UK: Routledge.

Krahmer, E. and M. Swerts. in press. "The Effects of Visual Beats on Prosodic Prominence: Acoustic Analyses, Auditory Perception and Visual Perception." To appear in *Journal of Memory and Language*.

Kranstedt, A. A. Lüking, T. Pfeiffer, H. Rieser, and I. Wachsmith. 2006. "Deixis: How to Determine Demonstrated Objects Using a Pointing Cone." In *Gestures in Human-Computer Interaction and Simulation*, 300–311. Berlin: Springer.

Kyburg, A. and M. Morreau. 2000. "Fitting words: Vague words in context." *Linguistics and Philosophy*, 23, 6, 577–597.

Lakoff, G., and M. Johnson. 1981. *Metaphors We Live by*. Chicago: University of Chicago Press.

Lascarides, A. and M. Stone. 2006. "Formal Semantics for Iconic Gesture." In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (Brandial)*. Potsdam.

Lascarides, A. and M. Stone. in press. "Discourse Coherence and Gesture Interpretation." To appear in *Gesture*.

Lewis, D. 1969. *Convention: A Philosophical Study*. Cambridge, Mass: Harvard University Press.

Lücking, A. H. Rieser, and M. Staudacher. 2006. "SDRT and Multi-Modal Situated Communication." In *Proceedings of BRANDIAL*. Potsdam.

Mann, W. C., and S. A. Thompson. 1987. "Rhetorical Structure Theory: A Framework for the Analysis of Texts." *International Pragmatics Association Papers in Pragmatics*, 1, 79–105.

McNeill, D. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.

McNeill, D. 2005. *Gesture and Thought*. Chicago: University of Chicago Press.

Nunberg, G.̇1978. *The Pragmatics of Reference*. Bloomington, Indiana: Indiana University Linguistics Club.

Oviatt, S. A. DeAngeli, and K. Kuhn. 1997. "Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction." In *Proceedings of the Conference on Human Factors in Computing Systems: CHI '97*. Los Angeles.

Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge, Mass: MIT Press.

Quek, F. D. McNeill, R. Bryll, S. Duncan, X. Ma, C. Kirbas, K. McCullough, and R. Ansari. 2002. "Multimodal Human Discourse: Gesture and Speech." *ACM Transactions on Computer-Human Interaction*, 9, 3, 171–193.

Reddy, M.J. 1993. "The conduit metaphor: A case of frame conflict in our language about language." In Andrew Ortony, ed., *Metaphor and Thought*, 164–201. New York: Cambridge University Press.

Schank, R.C. and R. Abelson. 1977. *Scripts, Plans, Goals, and Understanding*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Schlangen, D., and A. Lascarides. 2002. "Resolving Fragments using Discourse Information." In *Proceedings of the 6th International Workshop on the Semantics and Pragmatics of Dialogue (Edilog)*. Edinburgh.

So, W. S. Kita, and S. Goldin-Meadow. in press. "Using the Hands to Keep Track of Who Does What to Whom." To appear in *Cognitive Science*.

Sowa, T. and I. Wachsmuth. 2000. "Coverbal Iconic Gestures for Object Descriptions in Virtual Environments: an Emprical Study." In *Post-Proceedings of the Conference of Gestures: Meaning and Use*. Portugal.

Steedman, M. 2000. *The Syntactic Process*. Cambridge, Mass: MIT Press.

Stern, J. 2000. *Metaphor in Context*. Cambridge, Mass: MIT Press.

Talmy, L. 1996. "Fictive motion in language and "ception"." In Paul Bloom, Mary Peterson, Lynn Nadel, and Merrill Garrett, eds., *Language and Spacea*, 211–276. Cambridge, Mass: MIT Press.

Walker, M. 1993. *Informational redundancy and resource bounds in dialogue*. Ph.D. thesis, Department of Computer & Information Science, University of Pennsylvania.

Williamson, T. 1994. *Vagueness*. London: Routledge.