A Formal Semantic Analysis of Gesture

Alex Lascarides
School of Informatics,
University of Edinburgh

Matthew Stone
Computer Science,
Rutgers University

Abstract
This paper presents a formal semantics for productive iconic and deictic gestures that speakers use in tandem with speech in face-to-face conversation. In contrast to conventionalised, emblematic gestures, the meaning that productive gestures derive from their form is incomplete, and their content can be determined only through links to context, including previous gestures and simultaneous speech. Our model describes both the meaning revealed by form and the more specific interpretation in context. We build on established models of discourse to capture key insights from the descriptive literature on gesture—for instance, that synchronous speech and iconic gesture express a single thought. Our theory exploits dynamic semantics and rhetorical connections among communicative actions to describe how context contributes to the interpretation of gesture, including constraints on co-reference across speech and gesture, and across sequences of gesture.

1 Introduction
Gestures form an integrated part of face-to-face conversation. Gestures like the ones in utterance (1), taken from a corpus studied in Kopp et al. (2004), complement the words with precise spatial information.

(1) And [Norris]₁ is exactly across from the library.]₂
First: The left hand is 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 right hand is extended to the extreme upper right periphery.

Utterance (1) is visualised in Figure 1. Intuitively, the speaker’s gestures depict the spatial layout of the scene. It concludes a direction-giving episode in the dialogue (see example (13) in Section 2.3), in which the speaker has already created a virtual map of part of the university campus. The gesture on the left of Figure 1 repositions her left hand at the location initially established for Norris Hall. Similarly, the gesture on the right resumes the earlier demonstration of the library’s location. These two gestures demonstrate the spatial relationship of the buildings that’s asserted in the speech. Depiction is thus an essential part of the speaker’s communicative act.

The iconic gesture in (2) uses space more abstractly:

(2) 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 (i.e., vertically outwards) in clockwise circles (as viewed from left). The movement occurs below the mouth, where the previous gesture was performed.
Utterance (2) is part of a lecture about speech errors. In context, this gesture seems to show the iterative processes that cause low-level phonological errors, slipping beneath everyone’s awareness. As with utterance (1), the demonstration in (2) fits organically into the speaker’s overall message. In previous utterances to (2), the speaker used both words and gestures to show that anecdotal methods for studying speech errors are biased towards noticeable errors like Spoonerisms. Those errors were depicted with the hand emerging upward from the mouth into a prominent space between the speaker and his audience. The contrast set up by the quite different position of the gesture in (2) (below the mouth, nearer the speaker) helps to derive an interpretation of this gesture as depicting low-level phonological errors as less noticeable; and the fact that they are less noticeable explains why anecdotal methods would not naturally detect them. In other words, the content of the gesture in (2) serves as an explanation for why the content of its simultaneous speech is true.

The speaker’s movements in (2) are compatible with many different schemes for depicting the world on the hands. The speaker is just moving his fist in a circle, after all. The more specific understanding that emerges in context reflects a licensed figuration—namely, the process that causes speech errors being demonstrated by a fist carrying an error from the mouth—and the coherent relationships between speech and gesture—here, the recognition that the content of the gesture explains that of the speech.

Descriptive analyses of gesture 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, Bavelas and Chovil, 2000, Engle, 2000, Kendon, 2004). In Lascarides and Stone (2007), we argued that this observation can be encapsulated in terms of DISCOURSE COHERENCE. In other words, specific interpretations of multimodal utterances are derived via constrained inference which reconstructs how speech and gesture are rhetorically connected with relations like Background and Explanation. This parallels the constrained inference involved in recognising how sentences in discourse are coherent (Grosz and Sidner, 1986, Hobbs et al., 1993, Asher and Lascarides, 2003, Webber et al., 2003). In this paper we build
on this perspective. In formalising the principles of coherence that guide the interpretation of gesture, we go beyond previous work—whether descriptive (McNeill, 1992, Kendon, 2004), psychological (Lozano and Tversky, 2004), or applied to embodied agents (Cassell, 2001, Kopp et al., 2004)—by drawing on precisely-specified semantic models of discourse. At the same time, the new data afforded by gesture calls for extensions to earlier theories of purely linguistic discourse, resulting in more general and deeper models of semantics and pragmatics.

The analysis proceeds first by characterising coherent interpretations of multi-modal discourse (Section 2), then by characterising the semantic representations of multi-modal utterances as revealed by just their form (Section 3), and finally characterising how the former is derived from the latter via constrained inferences that exploit the discourse context (Section 4). Specifically in Section 2, we use dynamic semantics to capture the evolving structure of salient objects and spatial relationships that can be used to interpret gesture. A segmented structure organised by rhetorical connections among linguistic utterances and gestures constrains not only what communicative actions can coherently be performed now, but also the anaphoric dependencies across speech and gesture, and sequences of gesture.

In Section 3, we introduce underspecified semantic representations of gesture, as revealed by just their form. Underspecification allows us to formalise, very abstractly, what it is that a gesture must convey, given its particular pattern of shape and movement. Conversely, the alternative ways this underspecification can be resolved to a coherent (and specific) logical form give a discipline for systematising the role of context and inference in determining gesture interpretation. In Section 4, we describe how to formally model this: we introduce a glue logic for composing the logical form of discourse from the contents of the individual clauses and gestures as revealed by their form. This logic features default rules for inferring rhetorical relations—or, equivalently, the types of communicative actions that the speaker(s) performed—and as a byproduct of reasoning about the rhetorical connections, underspecified aspects of meaning are disambiguated to the pragmatically preferred, specific values.
These formal resources parallel those required for interpreting spoken discourse, as developed by Asher and Lascarides (2003) \textit{inter alia}. But gestures come with distinctive dynamics in their references to objects and space. They also connect to their contexts through distinctive rhetorical relationships, crucial for establishing and continuing depictions of spatialised situations.

The contribution of our work is to meet the challenge, implicit in descriptive work on non-verbal communication, of handling gesture within a theoretical framework that’s continuous with and complementary to purely linguistic theories. The semantic and pragmatic architecture we use here instantiates a general theory of discourse interpretation, Segmented Discourse Representation Theory (SDRT, Asher and Lascarides (2003)). We see both theoretical and practical advantages in modelling speech and gesture in a unified model. From a theoretical perspective, a uniform account of implicature is important for substantiating the descriptive claim that speech and gesture convey an integrated message. Indeed, there is empirical evidence that humans integrate both speech and gesture into a common representation that is available for expression in any modality (Cassell et al., 1999). From a practical perspective, a single pragmatic framework promises to greatly simplify the interface between the modules which process language and those which process gesture in any multimedia system (see also Kopp et al. (2004)). We look forward to implementing our approach in future work.

2 The Logical Form of Embodied Communication

Following Lascarides and Stone (2007), we assume that the logical form (LF) of an embodied discourse, like the logical form of purely spoken discourse, makes explicit the illocutionary content that the speaker is committed to in the conversation. This is distinguished from a more general pragmatic interpretation, which would also explicate such effects as politeness, indirection, and hinting. We also focus in this paper on LF as a representation of speaker meaning. We have reported elsewhere (DeVault and Stone, 2006, Thomason et al., 2006, Lascarides and Asher, 2007) on how the present framework is part of a model of the broader social reality of meaning—including grounding, disputes and misunderstandings. And we leave to future work the issue of how the content of gestures can be grounded or disputed. Finally, we note that a complete theory of embodied utterances must address how and why interlocutors choose to make the moves they do. But this requires a model of interlocutors’ psychology, as well as the relationships between embodied actions and their context specific meanings. Logical form is thus not the whole story in modelling utterance production, but it is an essential part.

In our LFs we use Rhetorical Relations to characterise the speech acts that agents perform, and Dynamic Semantics to define these acts’ effects on content. This section describes these constructs and shows how we use them to formalise the interpretation of gesture. The section culminates in the presentation of a language $\mathcal{L}_{base}$ that is based on that of SDRT (Asher and Lascarides, 2003).

Let’s start with the fact that gestures by their nature receive a vague interpretation, even in context. As is standard (Williamson, 1994, Kyburg and Morreau, 2000, Barker, 2002), our LFs will accommodate vague meanings through formulae that depend on the dynamic context of conversation. To take a well-studied case, the vague scalar predicate \textit{tall} appeals to the context to provide a specific standard $h$ for height. The LF relates \textit{tall} to $h$ as follows: \[ \text{tall}(x) \leftrightarrow \text{height}(x) > h. \] Such representations are precise, given a precise context. However,
interlocutors don’t typically nail down such a precise context. In contexts where the denotation of \( h \) is not determined uniquely, the LF \( \text{height}(x) > h \) will be vague about the property it ascribes to \( x \). Interlocutors can work to resolve these vagaries through their utterances, insofar as it is necessary for the purposes of the conversation.

The spatial information in gesture is also vague. It depends on an understood correspondence between the physical space in which the utterance takes place and the space where the objects and events described are actually (or metaphorically) located. For example, the spatial information conveyed in (1) depends on the correspondence between the speaker’s gesture space and the university campus. In our formalisation, this correspondence is a feature of the context, just like the height standard \( h \) for the vague predicate \( \text{tall} \). Interlocutors’ information will not fix one exact correspondence between real and virtual space—a range of possible correspondences can be at play in the conversation. Accordingly, the content associated with the interpretation of the gesture in context will fit a range of real-world layouts. It may leave open, for example, the relative positions of certain buildings on the campus.

Another source of vagueness in gesture interpretation is the metaphors that map spatial form to abstract concepts, as in (2). In formal semantics, metaphors in language can be fruitfully modeled in terms of context-dependent reference (Stern, 2000, Glucksberg and McGlone, 2001, Carston, 2002). In these models, a metaphor involves the use of a term to designate an unconventional property or relation—one related by a suitable metaphorical similarity to its usual reference. Take a verbal or gestural metaphor which shows something \textit{carrying something else along with it}, as in the figuration of the fist in (2). Rather than evoking the literal relation between a human agent and a physical object held in the hand, the metaphor evokes a different relation \textit{carry*} which characterises a process and its causal consequences. This indirect reference depends on the conventional referent \textit{carry} and a GENERATING PRINCIPLE taken from context that maps the conventional carrying activity into the domain of causal interactions. As revealed within cognitive linguistics (Lakoff and Johnson, 1981, Gibbs, 1994, Fauconnier, 1997), these mappings are typically richly-structured but flexible and open-ended. Thus, interlocutors’ information will no more fix a unique way to understand the metaphor than they will fix a standard for height. So metaphorical interpretations—though represented precisely in terms of context-dependent property reference—will be vague.

The discourse context can clearly make the interpretation of an utterance more precise than it would be otherwise. Following Lascarides and Stone (2006), we will use RHETORICAL RELATIONS to encode this interplay between context and interpretation: when one elaborates or explains (parts of) what has come before, the illocutionary effects of such speech acts constrain the semantics of the utterances. They may even place particular constraints on the parameters for interpreting vague language (Kyburg and Morreau, 2000, Barker, 2002). Eventually, interlocutors know enough from these constraints that—in Clark and Shaefer’s (1989) terms—they understand each other well enough for the purposes of the conversation.

Vague interpretations—which can be made ‘sufficiently precise’ while not necessarily unique—are to be distinguished from \textit{ambiguities} that need to be resolved to have a usable meaning at all. As is standard in the linguistics literature (e.g., Egg et al. (2001), Copetake et al. (2005)), we use semantic underspecification to represent semantic ambiguity (see Section 3). That is, the grammar composes a \textit{partial description} of the form of the LF of an utterance: for instance, the compositional semantic content of a pronoun will stipulate that the LF must include an equality between the variable that denotes the pronoun’s referent and some other variable that’s present in the LF of the context, but it will not stipulate \textit{which} contextual variable this is. Thus the partial description is compatible with several LFs: one
for each available variable in the LF of the context. Such partial descriptions are a useful for ensuring that the grammar neither under-determines nor over-determines content that’s revealed by form. And while the compositional semantics of the vague scalar predicate tall features no underspecification with the source of vagueness residing entirely in its dynamic interpretation in context, the ambiguity arising from anaphora must be resolved at LF.

2.1 Spatial Reference

Spatial reference is a fundamental source of the iconicity on which gesture meaning depends. The places, shapes and paths in space that we demonstrate in gesture can depict the gesture’s meaning. Spatial references are also essential to deictic reference to physical objects, for gestures that locate objects in a virtual space, and for gestures where movement traces paths that show static surfaces or objects in motion in described scenes.3

To formalise spatial reference, we offer three innovations to our logical language. First, we add symbols for places and paths through real space. Second, we add variables that map that physical space to other spaces, so as to capture how gestures establish virtual or metaphorical space. And third, we introduce new predicates that record the propositional information that gestures offer in locating entities in real and virtual spaces. This section presents each of these innovations in turn.

Before undertaking this presentation, however, we note that we do not treat spatial reference as the only source of iconicity in gesture interpretation. For example consider the depiction of a waiter carrying a tray in example (3).

(3) A waiter walked in.

The right arm is extended out to the right from the shoulder, with forearm held vertical, palm facing mostly upwards, and the hand in a loose open ASL 5 shape.

The speaker’s gesture mimics the prototypical posture in which a waiter carries a tray of food. Following the descriptive literature on gesture, we represent the form of such gestures with qualitative features, mirroring the English description we offered in (3)—the arm is held up and to the right, the palm points up, the hand is open etc. (see Section 3 for details). And we model the iconicity of this gesture by allowing each of these elements of gesture morphology to contribute a corresponding predication to the logical form. In this case, the position of the arm shows the waiter lifting the tray overhead; the horizontal palm shows the tray balanced in the waiter’s hand; the flat hand shape shows the contact and support between hand and tray. A formal representation of this interpretation is given in (4) (where w co-references with the argument of the predicate waiter in the logical form of the speech):

(4)  waiters-food-tray(t) \land right-hand(h,w) \land lift(e_1,w,t) \land overhead(t,w) \land balances(e_2,h,t) \land supports(e_3,h,t) \land contacts(e_4,h,t)

Iconicity here is captured by the relationship between the form of gestural elements—such as the pose of the arm and hand—and the predications that they contribute to logical form. Our formalism does not treat this iconicity as mediated by spatial reference. A qualitative representation of the form of the gesture and its meaning, as in (4), is necessary to recover the relationships of causality and choice that make the gesture a portrayal of the waiter’s action.

3Much the same applies to other forms of iconic representation. A line in a drawing, for example, shows from its placement the location of a corresponding feature in the depicted scene.
Merely locating the waiter’s body in space does not accomplish this. In fact, our intuition is that the speaker in (3) is not demonstrating a specific region in physical space as part of mimicking the waiter’s action.

The metaphorical gesture (2) receives a similar qualitative treatment. For example, the iterative motion of the hand is understood as a metaphorical depiction of the continuous operation of the processes responsible for speech errors. We do not account for this metaphorical meaning in terms of the speaker’s reference to the particular circular path in space that his hand follows. Instead, the meaning of this gesture stems from the speaker’s mimicry of an indefinitely repeated action plus generative principles of metaphorical meaning-transfer.

The gestures of (1), however, do invoke spatial reference in interpretation. This is because intuitively, the physical spatial reference in the gestures’ forms signal the (relative) locations of Norris Hall and the library in interpretation. To regiment the description of space in gesture, we begin by adding to the model a spatiotemporal locality \( L \subset \mathbb{R}^4 \) 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., \( [\vec{p}]^M = I^M(\vec{p}) \subset L^M \). Whenever we need to formalise the place or path in physical space designated by the speaker with gesture, we use a suitable constant \( \vec{p} \). For instance, the interpretations of the gestures in (1) use a spatial reference \( \vec{p}_n \) — denoting a place in front of the speaker’s shoulder — to locate Norris Hall, and a spatial reference \( \vec{p}_l \) — denoting a place up and to her right — to locate the library. The contrast is now evident between the use of space in (1) and the mimicry in (3). Whereas (3) characterises an action in terms of its underlying causal relationships, as portrayed through qualitative aspects of movement, (1) expresses intrinsically spatial information and does so 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 (5) — a description of the library from the discourse preceding (1) — the speaker designates the trajectory followed by the hand:

(5) 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.*

Here the motion of the hand sweeps out part of a cylindrical surface, which we might formalise as \( \vec{p}_s \). That surface is meant to represent the cylindrical exterior of the library building — a fact that must be captured in the logical form of the gesture by featuring \( \vec{p}_s \).

A further crosscutting distinction is whether the speaker indicates the location of the hand itself, as in (1) and (5), or uses the hand as a pointer 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 (6):

(6) [That] is a beautiful city.

*The speaker raises their right hand in an ASL 1-index hand shape, directing the index finger along a line to the region in space where a travel poster depicting Florence hangs on the wall.*

It seems clear that the speaker aims to call attention to the spatial location of the poster \( \vec{p}_p \), not the spatial location of the finger. Such distal reference is also available for dynamic gestures that trace extended paths. A speaker on the south observation deck of the Empire State Building might use (7) this way, for example.
The L train runs from 8th Avenue east into Brooklyn. The speaker takes up an ASL 1-index hand shape with their right hand, pointing first down and to the right in time with the utterance of “8th Avenue”, then scanning horizontally towards the left and finally slightly upward on “east into Brooklyn”.

We might naturally formalise the speaker as using the hand to call attention to the actual path $\vec{p}_L$ of the subway line. Further examples of the variety of spatial reference in gesture are provided by Johnston et al. (1997), Johnston (1998b), Lücking et al. (2006).

Utterances (6) and (7) also contrast with (1) and (5) in whether they link up with the real places or establish a virtual space that merely models the real world. In (6) the speaker locates the actual poster; in (7) he locates the actual path of the L. In (1) and (5), however, the speaker is NOT actually pointing at the library—she marks the IMAGINED locations of campus buildings in the space in front of her. The information that (6) and (7) give about the real world can therefore be characterised directly in terms of the real-world spatial regions $\vec{p}_p$ and $\vec{p}_L$ that the speaker’s gestures designate. By contrast, the information that utterances (1) and (5) give about the world 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. In (1) the speaker has set up a VIRTUAL SPACE where any reference $\vec{p}$ in front of her semantically contributes information about a corresponding place $v_c(\vec{p})$ in the university campus. The relationship between the real positions of the speaker’s hands during her demonstrations $\vec{p}_n$ and $\vec{p}_l$ 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 is at play in (5) when the speaker characterises the actual shape of the library facade $v_s(\vec{p}_s)$ in terms of an iconic reference to the curved path $\vec{p}_s$.

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 $v_1, v_2, \ldots$ of variables over transformations from subsets of $L$ to subsets of $L$. Introducing such variables simplifies the relationship between the form of a gesture and its semantics considerably. We do not have to assume that gesture is AMBIGUOUS between reference to physical space vs. virtual space. Rather, we assume that gesture always refers to physical space and always requires an interpreter to reconstruct the underlying relationship between this physical space and the situation that the speaker is describing (e.g., the gestures in (6) and (7) 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 transformation variables 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 2.4 for details). To respect the iconicity, the possible values for a mapping $v$ must be tightly constrained. While we would expect mappings that accomplish translation, rotation, and scaling by positive factors, we would not expect to find mappings that do mirroring transformations. At the same time (given their origin in human cognition and bodily action), we would not expect them always 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, $f(v) \in T$. We leave a precise characterisation of these mappings $T$ to future work.
While a mapping variable \( v \) is always at play when the semantics of gesture invokes physical spatial reference, its value may not be uniquely determined, even in context. Consider (7) again. Following Talmy (1996), the physical region \( \vec{p}_L \) that’s designated by the arm movement may depict in fast forward a single event of “fictive motion” that follows one imagined \( L \) train as it negotiates its path across New York City. We can formalise this interpretation of the gesture in terms of \( v_I(\vec{p}_L) \). But we might prefer to interpret the speaker’s spatial reference as covering a generic interval during which the \( L \) train operates (starting in 1928 and continuing into the indefinite future). Formally, this means introducing a mapping \( v_E \) that takes reference to each point in space at a particular time as a proxy for a reference to that point in space throughout the longer generic interval. The gesture is then interpreted in terms of \( v_E(\vec{p}_L) \). Utterances like (1) and (5) also afford alternative analyses as either exemplifying a statement at a particular time, or extending their spatiotemporal reference virtually throughout a wider interval. Such alternatives can only be resolved by a more comprehensive analysis of speakers’ patterns of spatial reference and spatial meaning in future work.

We complete an initial formalisation of the semantics of spatial reference in gesture by introducing two new predicates: \( \text{loc} \) and \( \text{classify} \) respectively describe the literal and metaphorical use of space to locate an entity. \( \text{loc} \) is a 3-place predicate taking an eventuality, an individual (or event) and a place as arguments, and \( \text{loc}(e, x, \vec{p}) \) 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, (8a) represents the interpretation of the gestures in (1).

\[
\begin{align*}
\text{a.} & \quad \text{loc}(e_1, n, v_c(\vec{p}_n)) \land \text{loc}(e_2, l, v_c(\vec{p})) \\
\text{b.} & \quad \text{loc}(e_3, f, v_s(\vec{p}_s)) \land \text{facade-front}(l, f) \\
\text{c.} & \quad \text{loc}(e_4, r, v_I(\vec{p}_p)) \\
\text{d.} & \quad \text{loc}(e_5, t, v_E(\vec{p}_L))
\end{align*}
\]

In words, \( e_1 \) is the state of \( n \), the discourse referent for Norris Hall that’s been introduced in speech, being contained in the spatiotemporal image \( v_c(\vec{p}_n) \) on the speaker’s virtual map of a designated 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 \( v_c(\vec{p}) \) of the designated point \( \vec{p} \) further up and to the right. (8b) states that the front \( f \) of the facade 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 (5). The predication \( \text{facade-front}(l, f) \) is not contributed by an explicit element of gesture morphology but is, as we describe in more detail in Section 2.2, the result of the BRIDGING inference that’s always required to link the referent portrayed in the gesture, \( f \) here, with available referents from the linguistic context (in this case, the library). The logical form (8c) of the deictic gesture in (6) locates the specific poster \( r \) at the location \( \vec{p}_p \) where the speaker is pointing. The deferred reference from \( r \) to the city that is depicted in the poster is obtained via pragmatics: context resolves to a specific value an underspecified relation between the deictic referent (the poster) and the referent of the pronoun \( \text{that} \) (the city shown in the poster), this underspecified relation being a part of the compositional semantics of the multimodal act (see Section 3.2 for details). Finally, according to (8d), the gesture of (7) indicates the generic transit of \( t \), the discourse referent for the \( L \) subway line, on a generic temporal-extension \( v_E(\vec{p}_L) \) of the real-world path through Manhattan and Brooklyn \( \vec{p}_L \) indicated by the speaker.

Speakers can also use positions in space as a proxy for other abstract predications. Such metaphorical uses are formalised through the predicate \( \text{classify} \). A representative illustration of metaphorical spatial gesture and its formalisation can be found in the conduit metaphor

9
for communication (Reddy, 1993), where information contributed by a particular dialogue agent is metaphorically located with that agent in space. Utterance (9) features this:

(9) 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.

It 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 (9) and the adjunct adverbial 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 $\text{classify}(e, u, v_m(\vec{p}_i))$ 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, and is not just a matter of qualitative features of the performance of the gesture that underpin a metaphorical resemblance (as in (2)). The logical form for the gesture in (9), then, will use the formula (10), where $u_*$ denotes an utterance whose content entails We have this one ball, and $\vec{p}_S$ denotes Susan’s location.

\begin{equation}
\text{classify}(e_6, u_*, v_m(\vec{p}_S))
\end{equation}

### 2.2 Dynamic Semantics

In this section, we motivate and explain our use of dynamic semantics to formalise how context contributes to the interpretation of gesture. We adopt just a few key principles. To start, gesture can introduce referents that are related to the entities made salient by simultaneous speech. We have illustrated this already through (3). The tray portrayed in the gesture is not mentioned explicitly in the utterance; indeed, reference to the tray seems incoherent without the support of context, as shown in (11a).

\begin{enumerate}
\item a. ??John walked in.  
    \emph{Gesture as in (3)}.  
\item b. He arrived at our table.  
    \emph{The speaker lowers his right hand from overhead to lower right in a wide scooping motion that keeps the hand oriented palm upward throughout}.  
\item c. He put it down beside our table.  
\end{enumerate}

The requirement to resolve reference in gesture using a natural BRIDGING connection to a salient referent from speech mirrors the interpretation of definite noun phrases in discourse (Clark, 1977, Chierchia, 1995) and constitutes a constraint on the logical form of gesture.
Second, entities depicted in prior gestures can inform the interpretation of subsequent ones. (11b), for instance, is a coherent subsequent utterance to (3), and it depicts the waiter lowering the tray and bringing it to rest. The tray remains unmentioned in the accompanying utterances, and indeed there is not even specific textual evidence in (11b) to support a bridging inference to the tray. So reference to it in (11b) must be achieved through its connection to the gesture in (3). Finally, objects that are introduced in gesture are not brought to prominence in such a way that they license the use of a pronoun in subsequent speech; see the anomalous use of *it* in (11c), an incoherent continuation of (3) followed by (11b).

The same constraints seem to apply to gestures that get their meaning through spatial reference rather than through mimicry. Example (6) shows that an entity evoked by deictic gesture (in this case, a poster) need not be linguistically explicit. But the city and the poster are linked with a natural bridging relation—namely, *shown*. And here too, the gesture does not license the use of a subsequent pronoun to refer to the poster: we cannot coherently follow (6) with (12). Rather, we must repeat the deictic gesture to the poster while uttering *it*, and/or use a new referring expression such as the demonstrative noun phrase *that poster*.

(12) ??I got it there.

We formalise this with a very simple dynamic semantics (van Eijk and Kamp, 1997), where context is represented as a partial variable assignment function, and content is treated in an extensional first-order way. As a discourse is interpreted the model remains fixed, but the formulae’s truth definitions change the input assignment function 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 into a new output 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 us them model anaphoric dependencies between spoken utterances and gesture and between sequences of gestures.4

Our key innovation is to implement the distinction between entities evoked explicitly in speech and those depicted only in gesture. The entities evoked in speech constitute a primary context of referents, which further communicative action in both speech and gesture can exploit so as to resolve anaphoric aspects of their meaning. Gestures, by contrast, offer a backgrounded or circumscribed set of referents, which are not necessarily available to interpret anaphora in speech (see (11c) and (12)), but are available for anaphoric links to subsequent gestures (see (11b). To model this distinction, we take our cue from Bittner’s (2001) treatment of the distinction between foregrounded and backgrounded referents in centering morphology and Asher and McCready’s (2006) analysis of modal discourse. Both proposals involve splitting the context into two separate assignment functions to distinguish referents with different status in the context. We also represent a context as a pair of variable assignment functions, rather than just one. The first element of the pair records the entities that can be used to interpret pronouns and other anaphors in speech. The second element records the entities that can serve as the basis for referential depiction in gesture. To ensure that indefinites which are introduced in speech can be antecedents to anaphoric elements in gesture,

---

4Our technical developments are straightforwardly compatible with more complex dynamic semantics (e.g., van den Berg (1996), Brasoveanu (2006a,b), Cumming (2007)). However, our very simple contexts are already sufficient to explore the content of gesture in examples such as (1) and (2).
existential quantifiers in the logic will trigger an update to both input variable assignment functions. Meanwhile, to delimit the scope of gesture, we introduce a MODAL OPERATOR \([G]\) to the logical language. This operator cordons off the context-updating operations that happen within its scope so that they apply only to the component of the context that represents the entities available for reference in gesture (see Section 2.4 for details).

2.3 Rhetorical Relations

We adapt SDRT (Asher and Lascarides, 2003) to represent rhetorical relations in the interpretation of embodied discourse. SDRT’s 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 include (binary) rhetorical relations among labels. Thus LFS in SDRT impose a hierarchical structure on labels: a label \(\pi_1\) IMMEDIATELY OUTSCOPES \(\pi_2\) if \(\pi_1\)’s formula includes a literal \(R(\pi, \pi_2)\) or \(R(\pi_2, \pi)\) for some \(R\) and \(\pi\). For situated utterances, we treat both the individual clauses in spoken utterances and the gestures that accompany them as units of discourse, and so they each receive a label.

Each rhetorical relation represents a type of (relational) speech act (Asher and Lascarides, 2003). Examples include Narration, the strategy of following a description of one eventuality with the description of another that is in contingent succession and that forms a common topic with the first eventuality; Background, a strategy that is like that of Narration save that the eventualities temporally overlap; and Contrast, the strategy of presenting related information about two entities, using parallelism of syntax and semantics to call attention to a relevant difference between the entities. The inventory also includes METATALK relations, that relate utterances at the level of the speech acts rather than content. For instance, you might follow \emph{Chris is impulsive} with \emph{I have to admit it}—an explanation of why you said Chris is impulsive, not an explanation of why Chris actually is impulsive. These metatalk relations are symbolised with a subscript star—\(Explanation^*\) for the case above.

To extend the account to gesture, we highlight an additional set of connections which specify the distinctive ways embodied communicative actions connect together. There are three such relations. First, Depiction is the strategy of using a gesture to visualise exactly the information conveyed in the simultaneous 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 redundant information in purely spoken discourse (Walker, 1993)—it is helpful, not marked.

Second, Overlay relates one gesture to another when the latter continues to develop a portrayal of the same virtual space. Example (1), being the culmination of an extended description that includes embodied utterances (13a) and (13b), illustrates this:

(13) a. Norris is like up here—
   \emph{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.
   \emph{After returning the right hand to rest, the right hand is re-extended now to the extreme upper right periphery, with palm held left.}
In the discourse (13) followed by (1), the speaker evokes the same virtual space, by designating the same physical positions when naming the buildings. Introducing Overlay into the inventory of rhetorical connections marks the intuition that commonalities in the use of space marks the coherent use of gesture. Here, a logical form of the discourse that features Overlay connections between the successive gestures in (13a), (13b), and the two gestures in (1) captures the correct spatial aspects of the discourse's content.

One might wonder why these relationships among mappings from physical to virtual space aren’t treated as part of the semantics of existing rhetorical relations, such as Narration or Background. But the distinct interpretations of (14) vs. (15) motivate including Overlay as a primitive relation in the inventory:

(14) \[ \pi_1: \text{Max fell.} \]
    \[ \pi_3: \text{John pushed him.} \]

(15) \[ \pi_1: \text{Max fell.} \]
    \[ \pi_2: \text{The speaker’s right hand, with flat palm, starts with the palm facing left (from the speaker’s perspective) and the hand placed near the centre of the speaker’s torso. The hand then moves in an arc down and to the left, ending near the left of the speaker’s torso with the palm facing downwards. Intuitively, this depicts Max’s fall.} \]
    \[ \pi_3: \text{John pushed him.} \]
    \[ \pi_4: \text{The speaker starts by placing both hands in the location where his right hand was at the end of the prior gesture (i.e., down and to the left of his torso). His hands, with flat palms facing towards his interlocutor, are pushed outward towards the the interlocutor, thereby depicting John pushing.} \]

(14') \[ \pi_0: \text{Explanation}(\pi_1, \pi_3) \]

The intuitive interpretation of (14) is given by the logical form (14') (for simplicity we have omitted the formulae representing the content of the clauses). In words, the content of the entire discourse \( \pi_0 \) is that John’s pushing Max explains why he fell. This contrasts with the preferred rhetorical connection \( \text{Narration}(\pi_1, \pi_3) \) between the same spoken utterances in (15), which are now accompanied by the gestures described. The interpretation of the gestures must effect this change to the content of the speech, and the Overlay relation can explain this. \( \text{Overlay}(\pi_2, \pi_4) \) entails that the mapping of the physical location of the hands to objects in \( \pi_4 \) continues those mappings from the gesture \( \pi_2 \). Therefore, together with \( \text{Depiction}(\pi_1, \pi_2) \) and \( \text{Depiction}(\pi_3, \pi_4) \), \( \text{Overlay}(\pi_2, \pi_4) \) entails that the location of John pushing Max is the place where Max fell to, rather than where he fell from. This is inconsistent with \( \text{Explanation}(\pi_1, \pi_3) \). Therefore by including Overlay as a primitive relation in the inventory, the glue logic for composing the LF of discourse can model why \( \text{Narration}(\pi_1, \pi_3) \) rather than \( \text{Explanation}(\pi_1, \pi_3) \) is inferred—it is the only way of getting a coherent interpretation of both the the gesture and the speech.

Our third new relation for connecting gestures to their context is Replication. \( \text{Replication}(\pi_1, \pi_2) \) means that both \( \pi_1 \) and \( \pi_2 \) are gestures, and \( \pi_1 \) and \( \pi_2 \) use the body in the same way to depict the same entities. We’ve seen an example of Replication in the discourse consisting of (3) followed by (11b). The Replication relation linking the first gesture to the second formalises their interpretive connection—that the second gesture continues to mimic the waiter’s actions. Introducing Replication as a primitive relation in the inventory is necessary in order to account for data concerning co-reference of NPs in simultaneous speech.
that would not be salient if the gestures are absent. For instance, while the pragmatically preferred interpretation of (16) implies that the three people who walked in are the same as the three people who walked out, the interpretation of the purely linguistic discourse where the gestures aren’t performed at all, while compatible with this co-reference, but does not imply it.

(16)  
(a) Just as a terrible comic came onto the stage  
(b) three people walked in.  
   *The right hand is placed near the speaker’s central torso, with a flat palm facing towards the speaker, and the hand moves out towards his interlocutor.*  
(c) But immediately after the comic started his act, three people walked out.  
   *The right hand, with the same shape, starts where the prior gesture ended. It moves towards the speaker’s central torso, ending where the first gesture began.*

As we have already seen, rhetorical connections among units of discourse create discourse segments. For instance, the label $π_0$ in (14′) represents the discourse segment consisting of the spoken clauses $π_1$ and $π_3$. While units of discourse are often *continuous* segments, they needn’t be (see Asher and Lascarides (2003) for many such examples). The interpretation of gesture in terms of rhetorical relations likewise structures discourse in more flexible ways, since, as we have seen, gestures like those in (15) will bear a rhetorical relation both to simultaneous speech and to previous gestures. In all cases, however, the graph over labels determined by the OUTSCOPE relation cannot contain directed cycles, and must consist of a single root—i.e., the unique segment that is the entire discourse.

Each rhetorical relation symbol receives a semantic interpretation that spells out its truth conditions in terms of those of its arguments. For instance, a discourse unit that is formed by connecting together smaller units with a VERIDICAL rhetorical relation entails the content of the smaller units, as interpreted in dynamic succession, and goes on to add a set of conditions that encode the particular illocutionary effects of the particular rhetorical connection. For example, $\text{Narration}(π_1, π_2)$ transforms an input context $C$ into an output one $C'$ only if $K_{π_1} \land K_{π_2} \land \varphi_{\text{Narration}(π_1, π_2)}$ also does this, where $\land$ is dynamic conjunction and $K_{π_1}$ and $K_{π_2}$ are the contents associated with the labels $π_1$ and $π_2$ respectively. The formula $\varphi_{\text{Narration}(π_1, π_2)}$ is a test on its input context and ensures that the individuals are in the same place in time and space at the end of the eventuality $e_{π_1}$ that’s described by $K_{π_1}$ as they are at the beginning of $e_{π_2}$ (and so $e_{π_1}$ precedes $e_{π_2}$). The formalisation of our new rhetorical relations involving gesture is straightforward in this setting; see Section 2.4. 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 to gesture 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 of communication (Lascarides and Stone, 2007).

The structure induced by the labels and their rhetorical connections impose constraints and preferences over interpretations. For example, SDRT gives a characterisation of ATTACHMENT that limits the rhetorical links interpreters should consider for a new utterance, based on the pattern of links in preceding discourse. SDRT also restricts the AVAILABILITY of possible referents as antecedents to anaphoric expressions in an utterance, based on the rhetorical connections that tie it to preceding discourse. Both of these ingredients of the theory carry over to embodied discourse (Lascarides and Stone, 2007).
Anticipating the arguments of Section 3, we assume that the grammar distinguishes deictic gestures from iconic ones. For the former, a grammar construction rule will combine the deictic gesture with its synchronous syntactic spoken constituent, and its semantic effect is to introduce an underspecified predicate symbol that relates the referent of the deictic gesture to that of the spoken unit. This underspecified relation resolves to a specific value when constructing the logical form for the discourse (e.g., shown for (6)). Similarly, a grammar construction rule combines iconic gesture with a syntactic constituent in the spoken unit (often a clause), and its semantic effect is to introduce an underspecified rhetorical connection between the labels \( \pi_g \) and \( \pi_s \) of the gesture part and the spoken part respectively. Thus for iconic gesture, constructing the logical form of the discourse involves achieving (at least) four, logically co-dependent tasks. First, one resolves the underspecified content of the gesture to specific values. Second, the underspecified rhetorical connection between \( \pi_g \) and \( \pi_s \) is resolved to one of a set of constrained values (e.g., Depiction). Third, one identifies a label for this rhetorical connection: if it’s a new label which in turn attaches to some available label in the discourse context then \( \pi_g \) and \( \pi_s \) start a new discourse segment; and if it’s an existing label, then \( \pi_g \) and \( \pi_s \) continue an existing discourse segment. And finally, one computes whether \( \pi_g \) and/or \( \pi_s \) are also directly rhetorically connected to other labels in the context.

As we mentioned, the discourse structure of the context imposes constraints on all of these decisions. For instance, SDRT’s theory of spoken discourse assumes roughly that labels in the context can connect to new information only if they are on the “right frontier”. More precisely, the available labels in the context for connections to new ones are (i) the last label \( \pi_l \) that was added, and (ii) any label that dominates \( \pi_l \) via a sequence of the outscopes relation and/or subordinating relations (e.g., Elaboration, Explanation, Background are subordinating, while Narration is not; for a complete list of subordinating relations see Asher and Lascarides (2003)). This defines the right frontier of the discourse structure when it is visualised as a graph, with outscoping and subordinating relations depicted with arcs in downward trajectories. However, extending SDRT to handle embodied discourse introduces a complication, because it means that the last label is not unique: while a spoken discourse imposes a linear order on its minimal discourse units (one can only say one clause at a time); now our minimal units may be performed simultaneously, since one can gesture and speak at the same time.

As proposed in Lascarides and Stone (2007), we aim to analyse embodied discourse with minimal changes to the semantics/pragmatic interface for modelling purely spoken discourse. We therefore explore here the working hypothesis that new attachments remain limited to the right frontier, the only difference being that instead of one last label there are two: the label \( \pi_l^g \) for the last minimal spoken unit, and the label \( \pi_l^g \) for its synchronous gesture (if there was one). Since there are two last labels there are now two right frontiers generated from them. The ‘spoken’ right-frontier \( \Pi^s \) is the set of all labels that dominate \( \pi_l^s \) via outscopes and subordinating relations, and the ‘gesture’ frontier \( \Pi^g \) is the set of all labels that dominate \( \pi_l^g \). Thus the available labels are now \( \Pi^s \cup \Pi^g \) (note that \( \Pi^s \cap \Pi^g \) is non-empty, including at least the unique root \( \pi_0 \)). This definition says that 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.

\[5\] In fact, the presence of the structural relations Parallel and Contrast and discourse subordination relax this right frontier constraint, but we ignore this here.
Given this revised definition of attachment, antecedents to anaphora can be controlled as in Asher and Lascarides (2003). Specifically, if \( \pi_2 \) labels a formula \( K_{\pi_2} \) which contains an anaphoric condition \( a \), then antecedents to \( a \) must be either accessible to \( a \) in \( K_{\pi_2} \) (where accessibility is defined as in Kamp and Reyle (1993)) or accessible to all parts of \( K_{\pi_1} \) where \( R(\pi_1, \pi_2) \) for some rhetorical relation \( R \). Simply put, the antecedent to an anaphoric condition must be in the same utterance, or in one to which it is rhetorically connected.

Let’s now examine how these discourse structural constraints impact on the interpretation of a particular example, namely (15). The logical form of the first utterance is arguably \( \pi_0 : Depiction(\pi_1, \pi_2) \), for this captures the intuition that the (resolved) content \( K_{\pi_2} \) of the gesture depicts the content of the spoken utterance (i.e., that Max fell). We treat \( Depiction \) as a coordinating relation, since intuitively the description of the scenario by one argument is not at a more fine-grained level than that of the other (unlike \( Elaboration \)). Nevertheless, \( \pi_1 \) and \( \pi_2 \) are both last (since they were simultaneous and part of the last communicative act in the discourse), and this means that \( \pi_3 \) and \( \pi_4 \), as well as being rhetorically connected to each other (as demanded by the grammar; see Section 3), can also connect to any of the available labels \( \pi_0, \pi_1 \) and \( \pi_2 \). We already argued for a connection \( Overlay(\pi_2, \pi_4) \) for this example. In addition the utterance \( \pi_3 \) contains a pronoun, which intuitively resolves to Max, and therefore \( \pi_3 \) must connect to \( \pi_1 \) to make this antecedent available. As we said earlier the relation is \( Narration \)—this is compatible with the \( Overlay \) between the gestures. So the logical form (15’) for this discourse respects the anaphoric dependencies among the units, the constraints imposed by the grammar, the ‘right-frontier’ constraints on which parts of the context can relate to new information, and the appropriate contextually-resolved interpretation:

\[
\text{(15')} \quad \pi_0 : Depiction(\pi_1, \pi_2) \land Narration(\pi_1, \pi_3) \land Overlay(\pi_2, \pi_4) \land Depiction(\pi_3, \pi_4)
\]

Thus the available labels now are \( \pi_3 \) and \( \pi_4 \) (because they are last), \( \pi_2 \) (because it attaches to \( \pi_4 \) with a subordinating relation) and \( \pi_0 \) (because it dominates available labels). But \( \pi_1 \) is not available. This correctly predicts that continuing the discourse with the purely spoken utterance \( It \ hurt \), where \( it \) is intended to co-refer with the fall, is anomalous. On the other hand, we predict that uttering \( That \ hurt \) and repeating the original gesture that depicted falling is acceptable: the gesture can attach to the available ‘falling’ gesture \( \pi_2 \) with \( Replication \), and the pronoun \( that \) can be treated as a demonstrative, and an the \( Elaboration \) relation between \( that \ hurt \) and the gesture resolves \( that \) to the falling.

### 2.4 Summary: Formalism and Examples

We now complete our presentation of the logical form for embodied discourse by giving formally precise definitions that encapsulate the ideas from prior sections. We start with the syntax of the language \( \mathcal{L}_{base} \) for expressing logical forms, which is based on that of SDRT. It is extended to include spatial expressions (see Section 2.1), and the two last labels (for speech and its synchronous gesture; see Section 2.3). The dynamic semantics of \( \mathcal{L}_{base} \) 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 the possible antecedents to anaphora in speech and gesture respectively (see Section 2.2). We will then illustrate the formalism via logical forms for our sample discourses.

**Definition 1 Vocabulary and Terms**

The following vocabulary provides the syntactic atoms of the language:
• A set $P$ of predicate symbols ($P_1, P_2, \ldots$) each with a specified sort giving its arity and the type of term expected in each argument position (individual, eventuality, place, and so on);
• A set $R$ of (2-place) rhetorical relation symbols over labels (e.g., Contrast, Explanation, Narration, Overlay, \ldots);
• Individual variables ($x_1, x_2, y, z \ldots$);
• Eventuality variables ($e_1, e_2 \ldots$);
• Constants for spatiotemporal regions ($\vec{p}_1, \vec{p}_2 \ldots$);
• Variables over mappings from one spatiotemporal region to another ($v_1, v_2 \ldots$);
• The boolean operators ($\wedge$, $\vee$, $\neg$, $\rightarrow$);
• 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;
• 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 for discourse is a Segmented Discourse Representation Structure (SDRS). This is constructed from SDRS-formulae, which in turn are defined recursively from the vocabulary and terms in Definition 1:

Definition 2 SDRS-Formulae
The SDRS-formulae is the set of expressions which satisfy the following:

1. If $P \in P$ is an $n$-place predicate and $i_1, i_2, \ldots, i_n$ are terms of the appropriate sort for $P$, then $P(i_1, \ldots, i_n)$ is an SDRS-formula.
2. If $R \in R$ is a rhetorical relation and $\pi_1$ and $\pi_2$ are labels, then $R(\pi_1, \pi_2)$ is an SDRS-formula.
3. If $\phi$ is an SDRS-formula, then so are $\exists u\phi$ and $\forall u\phi$, where $u$ is an individual, eventuality or mapping variable.
4. If $\phi$ and $\psi$ are SDRS-formulae, then so are $\phi \land \psi$, $\phi \lor \psi$, $\phi \rightarrow \psi$ and $\neg \phi$.

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: $(A, F, last)$, where:

• $A$ is a set of labels;
• $F$ is a mapping from the labels in $A$ to SDRS-formulae; and

17
• last is either (a) a single label \( \pi \in A \) (intuitively, this is the last communicative act performed, which is within a single modality); or (b) last is a binary set of labels \( \{ \pi_s, \pi_g \} \subseteq A \), where \( \pi_s \) labels the content of a token speech utterance, and \( \pi_g \) the content of a token gesture (intuitively, this is the last communicative act, which is multi-modal).

We say that \( \pi \) immediately outscopes \( \pi' \) iff \( F(\pi) \) contains as a literal either \( R(\pi'', \pi') \) or \( R(\pi', \pi'') \) for some \( R \) and \( \pi'' \). The relation \( \succ \) that is the transitive closure of immediately outscopes then satisfies the following two constraints: it forms a well-founded partial order; and it has a unique root (that is, there is a unique \( \pi_0 \in A \) such that \( \forall \pi \in A, \pi_0 \succeq \pi \)).

The constraint that there is a unique root ensures that an SDRS is 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 \( \pi_4 \) in (15), for instance). When there is no confusion we may omit last from an SDRS, making it \( \langle A, F \rangle \). As shown in the SDRS (15'), we may also stipulate the mapping \( F \) with the notation \( \pi : \phi \). For instance, \( \pi_0 : R(\pi_1, \pi_2) \) is a notational variant of \( F(\pi_0) = R(\pi_1, \pi_2) \). We also often gloss the content \( F(\pi) \) as \( K_\pi \).

We now turn to the model theory.

**Definition 4  Model**

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

- \( D \) is a domain of entities, partitioned into eventualities (\( D_E \)) and individuals (\( D_I \)).
- \( L \subset \mathcal{R}^4 \) is a spatiotemporal locality.
- \( T \) is a set of mappings from \( L \) to \( L \) which are all geometric isomorphisms but not inversions. I.e., where \( t \in T \) there is no ‘reference circle’ centered at \( \vec{p} \) with radius \( r \) such that \( t \) maps coordinates inside the circle to outside it and vice versa (a mapping to a mirror image is a special case of an inversion, where \( r \) is infinite). So mappings in \( T \) can expand or contract distances between coordinates and rotate space, but they cannot map coordinates to their mirror image along some plane (or point in time).
- \( I \) is an interpretation function that maps non-logical constants to denotations of appropriate type (e.g., \( I(\vec{p}) \subseteq L \) and for a 2-place predicate \( P \in \mathcal{P} \) taking an eventuality and individual as arguments \( I(P) \subseteq D_E \times D_I \)).

Note that \( I \) does NOT assign denotations to the rhetorical relations. We’ll return to their interpretation shortly.

The semantics of an SDRS-formula \( \phi \) relative to a model \( M \) will specify a context-change potential that characterises exactly when \( \phi \) relates an input context to an output one. A context is a pair of partial variable assignment functions \( \langle f, g \rangle \); these define values for individuals 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 formulae save the existential quantifier and conjunction impose tests on the input context. The existential quantifier \( \exists x \) extends the input
variable functions \(\langle f, g \rangle\) to be defined for \(x\), and dynamic conjunction is composition. Hence \(\exists x \phi\) is equivalent to \(\exists x \land \phi\). Our innovation for gesture is to add a modality \([\mathcal{G}]\) to ensure that all formulae in its scope act as tests or updates only on the second input variable assignment function \(g\) in the input context \(\langle f, g \rangle\), but leave \(f\) unchanged. This means that an indefinite that’s within the scope of gesture can act as an antecedent to a subsequent anaphoric condition that’s also within the scope of gesture, but it cannot act as an antecedent to an anaphoric condition that’s introduced in speech (for its value won’t be defined by \(f\) in the input context \(\langle f, g \rangle\)).

**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, \(\langle f, g \rangle[i] = f(i)\).
3. Where \(v(p)\) is a spatial term, \(\langle f, g \rangle[v(p)] = f(I(v(p)))\).
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 \(\langle f, g \rangle[i_1]^M, \ldots, \langle f, g \rangle[i_n]^M \in I(P^n)\)
5. \(\langle f, g \rangle[\exists x]^M \langle f', g' \rangle\) iff:
   - (a) \(\text{dom}(f') = \text{dom}(f) \cup \{x\}\), and \(\forall y \in \text{dom}(f), f'(y) = f(y)\) (i.e., \(f \subseteq_x f'\));
   - (b) \(\text{dom}(g') = \text{dom}(g) \cup \{x\}\) and \(\forall y \in \text{dom}(g), g'(y) = g(y)\) (i.e., \(g \subseteq_x g'\));
   - (c) \(f'(x) = g'(x)\).
6. \(\langle f, g \rangle[\phi \land \psi]^M \langle f', g' \rangle\) iff \(\langle f, g \rangle[\phi]^M \circ [\psi]^M \langle f', g' \rangle\).
7. \(\langle f, g \rangle[\phi \lor \psi]^M \langle f', g' \rangle\) iff \(\langle f, g \rangle = \langle f', g' \rangle\) and there is an \(\langle f'', g'' \rangle\) such that either \(\langle f, g \rangle[\phi]^M \langle f'', g'' \rangle\) or \(\langle f, g \rangle[\psi]^M \langle f'', g'' \rangle\)
8. \(\langle f, g \rangle[\neg \phi]^M \langle f', g' \rangle\) iff \(\langle f, g \rangle = \langle f', g' \rangle\) and there is no \(\langle f'', g'' \rangle\) such that \(\langle f, g \rangle[\phi]^M \langle f'', g'' \rangle\)
9. \(\langle f, g \rangle[\phi \rightarrow \psi]^M \langle f', g' \rangle\) iff \(\langle f, g \rangle = \langle f', g' \rangle\) and for all \(\langle f'', g'' \rangle\) such that \(\langle f, g \rangle[\phi]^M \langle f'', g'' \rangle\) there is a \(\langle f''', g''' \rangle\) such that \(\langle f'', g'' \rangle[\psi]^M \langle f''', g''' \rangle\)
10. \(\langle f, g \rangle[[\mathcal{G}]\phi]^M \langle f', g' \rangle\) iff \(f = f'\) and \(\exists g''(g, g)[\phi]^M \langle g', g'' \rangle\)

Finally, we address the semantics of rhetorical relations. Unlike the predicate symbols in \(\mathcal{P}\), these do not impose tests on the input context, but rather they transform it. This reflects their status as speech acts; like actions generally they change the context. We emphasise veridical relations, as shown in Definition 6:

**Definition 6**  
Semantic Schema for Rhetorical Relations

Let \(R\) be a verbal rhetorical relation (i.e., **Narration**, **Background**, **Elaboration**, **Explanation**, **Contrast**, **Parallel**, **Depiction**, **Overlay**, **Replication**), and let \(\pi_1\) and \(\pi_2\) be labels. Then:

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

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 \(K_{\pi_1}\) followed by \(K_{\pi_2}\) followed by some particular illocutionary effects \(\varphi_{R(\pi_1, \pi_2)}\) also do this. The meaning postulates that define the illocutionary effects \(\varphi_{R(\pi_1, \pi_2)}\) for various relations \(R\) then forms a major component of SDRT. For instance, the meaning postulate for
\( \varphi_{\text{Narration}}(\pi_1, \pi_2) \) stipulates that individuals are in the same spatio-temporal location at the end of the first described event \( e_\alpha \) as they are at the start of the second described event \( e_\beta \), and so \( e_\alpha \) temporally precedes \( e_\beta \) (we assume \textit{prestate} and \textit{poststate} are functions that map an eventuality to the spatiotemporal regions in \( L \) of its prestate and poststate respectively).\(^6\)

- **Meaning Postulate for Narration**
  \[
  \varphi_{\text{Narration}}(\alpha, \beta) \rightarrow \text{overlap}(\text{poststate}(e_\alpha), \text{prestate}(e_\beta))
  \]

So, for instance, representing (17) as (17') ensures that its dynamic interpretation captures the following: John went out the door, and then from the other side of the door he turned right (for simplicity, we have omitted from (17') the semantic representations of the clauses \( \pi_1 \) and \( \pi_2 \)).

(17) \( \pi_1 \). John went out the door.
\( \pi_2 \). He turned right.

(17') \( \pi_0 : \text{Narration}(\pi_1, \pi_2) \)

Rhetorical relations that are already a part of SDRT—like Explanation and Narration—can now relate contents of gesture. On the other hand, as we’ll see in Section 3, some rhetorical relations are excluded from relating gestures (e.g., Disjunction). We also argued earlier for the introduction of 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, \( \text{Depiction}(\pi_1, \pi_2) \) holds only if \( \pi_1 \) labels the content of a spoken unit and \( \pi_2 \) labels the content of a gesture, and \( K_{\pi_1} \) and \( K_{\pi_2} \) are nonmonotonically equivalent. We represent the modality in which content is performed with the predicate symbols \textit{speech} and \textit{gesture}, as shown below, and assume that the grammar determines the modality for labels of minimal units (see Section 3), and for larger units \( \pi \), \textit{speech}(\pi) holds if and only if \( \forall \pi' \) where \( \pi \geq \pi' \), \textit{speech}(\pi') holds (similarly for \textit{gesture}).

- **Meaning Postulate for Depiction:**
  \[
  \varphi_{\text{Depiction}}(\pi_1, \pi_2) \rightarrow (\text{speech}(\pi_1) \land \text{gesture}(\pi_2) \land (K_{\pi_1} > K_{\pi_2}) \land (K_{\pi_2} > K_{\pi_1}))
  \]

In words, \( \phi > \psi \) means \textit{If \( \phi \) then normally \( \psi \)}. See Asher and Lascarides (2003) for its interpretation (this requires a modal model theory, and so we don’t give details here).

\( \text{Overlay}(\pi_1, \pi_2) \) holds only if \( \pi_1 \) and \( \pi_2 \) label the contents of gesture, and \( K_{\pi_2} \) continues to develop a portrayal of the same virtual space as \( K_{\pi_1} \):

- **Meaning Postulate for Overlay:**
  \[
  \varphi_{\text{Overlay}}(\pi_1, \pi_2) \rightarrow (\text{gesture}(\pi_1) \land \text{gesture}(\pi_2) \land \exists v \exists P \exists P'(\left((K_{\pi_1} \rightarrow P(v)) \land (K_{\pi_2} \rightarrow P'(v))\right))
  \]

Finally, \( \varphi_{\text{Replication}}(\pi_1, \pi_2) \) holds only if \( \pi_1 \) and \( \pi_2 \) are gestures, and they depict common entities in the same way. More formally, there is a partial isomorphic mapping \( \mu \) from the constructors in \( K_{\pi_1} \) to those in \( K_{\pi_2} \) such that for all constructors \( c \) from \( K_{\pi_1} \), \( e \) and \( \mu(e) \) are semantically similar. We forego a formal definition of semantic similarity here.

We are now in a position to define how SDRSS are interpreted with respect to an input context \( (f, g) \) and a model \( M \).

\(^6\)In fact, this axiom is stated here in slightly simplified form; for details, see Asher and Lascarides (2003).
Definition 7  The Dynamic Interpretation of an SDRS

Let $S = \langle A, F, \text{last} \rangle$ be an SDRS, and let $\pi_0 \in A$ be the unique root label of $S$ (by Definition 3 $\pi_0$ exists). Then:

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

In words, the CCP of an SDRS is that of the SDRS-formula associated with its (unique) root label $\pi_0$. If the SDRS features only veridical rhetorical relations, then the SDRS transforms an input context into an output one only if the contents of each speech and gesture act performed in the discourse also do this.

As discussed in Section 2.3, we minimise the changes to SDRT’s original notion of availability for purely spoken discourse (see Definition 8 from Asher and Lascarides (2003)), ensuring that the available labels in embodied discourse are those which are on the right frontier of at least one last label (see Definition 9).

Definition 8  SDRT’s Original Availability

Let $S = \langle A, F, \text{last} \rangle$ be an SDRS. Where $\pi \in A$, we say that $\pi > \pi'$ if either $\pi$ immediately outscopes $\pi'$ or there is a label $\pi''$ such that $F(\pi'')$ contains as a 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 the available labels of $S$ are the labels $\pi \in A$ such that $\pi >^* \text{last}$.

Definition 9  Availability for Multimodal Discourse

Let $S = \langle A, F, \text{last} \rangle$ be an SDRS for multimodal discourse (and so by Definition 3, last is a unary or binary set of labels). Then the available labels of $S$ are the labels $\pi \in A$ such that $\pi >^* l$, where $l \in \text{last}$.

SDRT’s existing constraints on interpreting anaphora remain unchanged, in keeping with our strategy for minimising changes that are engendered by gesture:

Definition 10  Antecedents to Anaphora

Suppose that $K_\beta$ contains an anaphoric condition $\varphi$. Then the available antecedents to the anaphoric condition are the discourse referents that are:

1. in $K_\beta$ and DRS-accessible to $\varphi$ (see Kamp and Reyle (1993) for a definition of DRS-accessibility)

2. in $K_\alpha$, DRS-accessible to any condition in $K_\alpha$, and there is a condition $R(\alpha, \gamma)$ in the SDRS such that $\gamma = \beta$ or outscopes($\gamma, \beta$).

In other words, an antecedent must be in the same utterance as the anaphoric expression, or in one to which it is rhetorically connected.

To illustrate this formalism, we give the precise semantics for a number of the key discourses that have guided its development. We start with the embodied discourse (3) followed by (11b), repeated here as (18). The SDRS for (18) is shown in (18′)—with some simplifications, since we have ignored tense and presuppositions and also omitted existential quantification over events:

---

7This ignores the complexities rendered by Contrast and Parallel (Asher and Lascarides, 2003).
(18a) a. \( \pi_1 \): A waiter walked in.
\( \pi_2 \): The right arm is extended out to the right from the shoulder, with forearm held vertical, palm facing upwards, and the hand in an open ASL 5 shape.

b. \( \pi_3 \): He arrived at our table.
\( \pi_4 \): The speaker lowers his right hand from overhead to lower right in a wide scooping motion that keeps the hand oriented palm upward throughout.

This logical form is the result of a highly complex inference process that resolves underspecified aspects of meaning that are revealed by form, both in speech and gesture. In fact, resolving the underspecified content and resolving the rhetorical connections are logically co-dependent tasks. We will discuss such inferential processes in Section 4. Here, we focus only on the dynamic semantic interpretation of (18a). First, sentence \( \pi_1 \) introduces the waiter and his entry while its gesture \( \pi_2 \) depicts his action of carrying his tray overhead. The referential connection between the gesture and the speech (i.e., the use in \( K_{\pi_2} \) of the variable \( w \), which is bound by an existential quantifier in \( K_{\pi_1} \)) is licensed by the fact that they are rhetorically connected—this connection makes \( w \) available as an antecedent for resolving the anaphoric conditions in the content of the gesture (i.e., introducing any individual in gesture is constrained to be bridging related to an available antecedent in speech, as discussed earlier). It thus supports interpreting the right hand as depicting the waiter’s tray, a plausible bridging reference in this context. \( Background(\pi_1, \pi_2) \) entails that carrying the tray and walking temporally overlap. The logical form of (18a), consisting of \( \pi_0 : Background(\pi_1, \pi_2) \), ensures that the available labels for subsequent rhetorical connections are \( \pi_1 \) and \( \pi_2 \) (because they are both last) and \( \pi_0 \) (because it outscreens available labels). Thus the connections for \( \pi_3 \) and \( \pi_4 \) that are shown in (18a) respect constraints on availability given in Definition 9.

The rhetorical connections of \( \pi_3 \) and \( \pi_4 \) to the context also have semantic consequences. \( Narration(\pi_1, \pi_3) \) entails that he walked in and then he arrived at our table (and their rhetorical connection licenses interpreting the pronoun he as the waiter); \( Narration(\pi_3, \pi_4) \) entails that he put down the tray after arriving at the table, and he did so at the table; and \( Replication(\pi_2, \pi_4) \) entails that the use of physical space in both gestures is the same, ensuring that both gestures depict the imitation of the waiter’s embodied action.

We now focus on examples that illustrate the possibilities for coherence between speech and gesture within a single utterance, starting with the deictic gesture in (6):

(6) [That] is a beautiful city.

The speaker raises their right hand in an ASL 1-index handshape, directing the index finger along a line to the region in space where a travel poster depicting Florence hangs on the wall.

(6') \( \pi_0 : \exists c(city(c) \land beautiful(c)) \land \[G] \exists p(\text{shown}(p,c) \land \text{loc}(e,p,v_I(\vec{p}))) \)
The compositional semantics of (6) introduces an underspecified anaphoric condition for *that* in the speech, and an underspecified condition *deictic_rel* for relating the denotation of *that* and the denotation of the deictic gesture. Resolving *deictic_rel* (to *shown*), identifying the denotation of *that* (to the city *c* that’s shown in the poster *p*), and identifying the mapping of physical space set up by the deictic gesture to what is denoted by it (to *v_I*) are all logically co-dependent tasks. Note that like iconic gestures, individuals that are introduced by deictic gestures but not explicitly referred to in synchronous speech, even by a pronoun like *that* or *he*, are embedded within the gesture modality (here, this generalisation about semantic scope ensures that \([G]\) has scope over the quantifier that binds the poster *p*). This explains the anomaly of continuing with (12), *I got it there*.

In (7), the words give a symbolic description of the path of the L train *t*, and the gesture shows that trajectory, by indicating it in physical space:

\(7\) The L train runs from 8th Avenue east into Brooklyn.

The speaker takes up an ASL 1-index hand shape with their right hand, pointing first down and to the right in time with the utterance of “8th Avenue”, then scanning horizontally towards the left and finally slightly upward on “east into Brooklyn”.

\(7'\) \(\pi_1 : \exists tabd(ltrain(t) \land runs(e, t) \land from(e, a) \land eighth-ave(a) \land direction(e, d) \land east(d) \land into(e, b) \land brooklyn(b))\)

\(\pi_2 : [G](loc(e', t, v_E(\vec{p}_L)))\)

\(\pi_0 : Depiction(\pi_1, \pi_2)\)

The logical form \((9')\) of example \((9)\) formalises the metaphorical use of spatial reference. We give the interpretation of speech in two segments. One \((\pi_1)\) gives the description of the content: we have one ball. The other \((\pi_2)\) elaborates the speaker’s act in providing this information: it’s something Susan has already said.

\(9\) 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.

\(9'\) \(\pi_1 : \exists wb(we(w) \land have(e, w, b) \land one(b) \land ball(b))\)

\(\pi_2 : \exists us(susan(s) \land said(e', u, s))\)

\(\pi_3 : [G]classify(e'', u, v_m(\vec{p}_1))\)

\(\pi : Depiction(\pi_2, \pi_3)\)

\(\pi_0 : Elaboration_*(\pi_1, \pi)\)

The gesture \((\pi_3)\) 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 \(K_{\pi_1}\).

Finally, in (2), the metaphorical gesture explains the content of the spoken utterance; it depicts that low-level phonological errors are sustained—carried along with them—by continuous subconscious processes:

\(2\) So there are these very low level phonological errors that tend to not get reported.

In the gesture used here, the hand is held in a fist with the thumb to the side (as
in an ASL A) and moves iteratively in the sagittal plane (i.e., vertically outwards) in clockwise circles (as viewed from left). The movement occurs below the mouth, where the previous gesture was performed.

(2’)

\[ \pi_1: \exists y (\text{low-level}(y) \land \text{phonological}(y) \land \text{errors}(y) \land \text{go-unreported}(e, y)) \]

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

3 Underspecified Meaning for Gesture

We intend to derive underspecified semantic representations of iconic and deictic gestures from their form. To do this we re-use devices from linguistics (e.g., Muskens (1996) and others). On such approaches, the grammar does not explicitly construct the logical form. Instead, it builds a partial description of it, leaving open multiple alternatives. The underspecified elements in the description must be resolved pragmatically to identify a specific logical form. More formally, formulae in \( \mathcal{L}_{ulf} \) are partial descriptions of the formulae in the language \( \mathcal{L}_{base} \) of fully-specific logical forms. Each model \( M \) for \( \mathcal{L}_{ulf} \) corresponds to a unique formula in \( \mathcal{L}_{base} \), and \( M \) satisfies \( \phi \in \mathcal{L}_{ulf} \) if and only if \( \phi \) (partially) describes the unique formula corresponding to \( M \). So when \( \phi \) is satisfied by more than one model, it supports more than one alternative specific interpretation.

Semantic underspecification languages are typically able to express partial information about semantic scope (e.g., Alshawi and Crouch (1992)). Formalisms which also underspecify the content of anaphora, ellipsis and lexical senses have also been proposed (e.g., Koller et al. (2000), Poesio (1996), Asher and Lascarides (2003)). 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. We will show that the iconic meaning of gesture constrains, but doesn’t fully determine, all these aspects of interpretation. Consequently we will adopt RMRS as the underlying semantic formalism \( \mathcal{L}_{ulf} \).

3.1 The Form of Gesture

A sentence’s syntax determines its predicate argument structure in semantics. In contrast, gestures lack hierarchical syntax (McNeill, 1992). Rather, they are ‘multidimensional’: 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 along those paths are all aspects of a gesture’s form that potentially reveal things about its meaning. But unlike language, they do not combine in a hierarchical way, and consequently predicate argument structure is left underspecified.\(^8\)

For example, a fist-shaped hand (ASL-A) can denote a holds relation of some kind (either literally or metaphorically), or it can denote an object. In (2), the fist shape contributes a metaphorical holding or carrying relation, depicting something metaphorically carrying or ‘supporting’ speech errors (one conceives of the speech errors as being ‘held inside’ the hand, while the hand itself denotes the thing that sustains or causes the speech errors). In a neo-Davidsonian representation, therefore, the fist shape contributes a three-place predicate symbol \( \text{sustain}(e, x, y) \) (where \( x \) denotes the processes that cause the speech errors \( y \)) to the

\(^8\)This makes compositional semantics a misnomer for gesture, but we will continue to use this term to distinguish meaning derived from form vs. interpretation in context.
contextually-determined logical form (see (2')). In (19), however, the same fist shape denotes an object—namely, a marker point on the wheel—and it does not depict that something is contained within that object.

(19) The mouse ran round the wheel for a few minutes.

Gesture is as illustrated in Figure 2.

Thus the fist shape contributes a one-place predicate symbol marker-point(x) to the contextually-resolved logical form. We need the underspecified content of this hand shape ASL-A in iconic gesture to support both of these possible resolved interpretations. Hence the it must underspecify the arity of the predications it contributes to interpretation, and the sort of arguments they qualify (an event for sustain, but an individual for marker-point).

In formal terms, each feature of a gesture—such as hand shape—potentially resolves to a complex conjunction of elementary predications. This is also true of words, particularly for so-called nonce uses such as (20) (Nunberg, 1995).

(20) The hamburger is getting impatient.

Like the hand-shape and other features of iconic gesture, convention imposes constraints on what a word can mean, but context can embellish it in an unlimited number of ways. Thus researchers in lexical semantics advocate representing the meanings of words as contextually-determined conjunctions of elementary predications, thereby allowing for an unlimited number of meaning shifts in context (Pustejovsky, 1995). However, unlike a hand shape, the lexical subcategorisation of a word is fully specified by form, and so although the value of the predicate symbol (or conjunctions thereof) that it contributes to logical form aren’t specified by syntax, the arity of the complex expression is fixed by it.

With this in mind, let’s now specify the form and underspecified meaning of iconic gesture. We’ll illustrate the idea with the gesture in Figure 2. Following Kopp et al. (2004), we represent the multi-dimensionality of its form with a feature structure: each of the six features (hand shape, finger-direction, palm-direction, trajectory, movement direction, and location) take particular values which correspond to the physical features of the gesture. The six features thus combine via unification or ‘conjunction’. The form of the gesture in Figure 2 is shown in (21):

(21)  

\[
\begin{align*}
\text{iconic-gesture} & \\
\text{right-hand-shape} : \ asl-a & \\
\text{right-finger-direction} : \ down & \\
\text{right-palm-direction} : \ left & \\
\text{right-trajectory} : \ sagittal-circle & \\
\text{right-movement-direction} : \ \{\text{iterative, clockwise}\} & \\
\text{right-location} : \ central-right & 
\end{align*}
\]

Alternatively, one could conceive of marker-point as a 2-place relation (x is a marker-point on y), rather than assuming, as we have done here, that making x a marker point on the wheel y is expressed as marker-point(x) ∧ part-of(x, y) ∧ wheel(y). But our point about the different arities of predications in alternative, resolved interpretations of the hand shape still stands.

If the hand shape changes during the performance of the gesture—e.g., the two fingers move to depict walking legs—then we assume that the corresponding value is more complex, consisting of a sequence of values. This will then affect the underspecified predications in semantics, but we forego details here.
Some values are expressed as *sets*. For example, the iterative clockwise trajectory of the movement is composed of the set of two values. This allows us to capture generalisations over clockwise trajectories on the one hand (iterative or not), and iterative trajectories on the other although we do not explore this further here.

Unlike Kopp et al. (2004), we have typed the gesture as being of type *iconic-gesture*. This allows us to capture generalisations about which features are appropriate for which types of gesture. So, for example, the feature *trajectory* is not always relevant for deictic gesture. Thus, we might represent the form of the deictic gesture shown on the right hand in (1) with the feature structure (22):

```
(22)

  deictic-gesture
  right-hand-shape : loose-asl-5-thumb-open
  right-finger-direction : forward
  right-palm-direction : up-left
  right-location : central-upper-right
```

Furthermore, the form of deictic gesture must serve to identify the region $\vec{p}$ in space that is denoted by it; but spatial reference isn’t always a part of the interpretation of iconic gesture (e.g., see (2)). Type distinctions are useful for capturing such differences in the form-meaning relation (Pollard and Sag, 1994). One can also quite naturally extend the type hierarchy of gestures to include further subtypes of *deictic-gesture* and further subtypes of *iconic-gesture*, which would provide more specific information about form and meaning for particular gestures. But we forego details of this here.

### 3.2 The Meaning of Gesture

We need to map the forms of gestures to their meaning. We’ll start with iconic gesture, using (21) as an illustrative example. Its underspecified meaning must constitute a partial description of all its possible interpretations, abstracting over all discourse contexts. It must underspecify the values of predicate symbols and the arity of their arguments, semantic dependencies and semantic scope. As we said, RMRS (Copestake, 2003) can do this. It is a generalisation of Minimal Recursion Semantics (MRS; Copestake et al. (2005)). To illustrate with a (toy) example, sentence (23a) features a semantic scope ambiguity that syntax doesn’t resolve. Its MRS (23b) reflects this: (i) each predication is labelled ($l_1$, $l_4$, etc); (ii) scopal arguments are indicated by holes (e.g., $h_2$ and $h_3$ in the quantifier *every*); (iii) constraints on scope are expressed by the qeq conditions (roughly, $h =_q l$ means that only zero or more quantifiers lie between the scopal argument $h$ and the predication labelled by $l$ in any fully-resolved logical form described by the MRS); and (iv) these scope constraints allow for more than one way of ‘plugging’ the holes with labels.
(23)  a. Every black cat loved some dog.
    b. \[l_1 : \text{every}_q(x, h_1, h_2)\]
      \[l_2 : \text{black}_a(e_1, x)\]
      \[l_2 : \text{cat}_n I(x)\]
      \[l_3 : \text{loved}_v I(e_2, x, y)\]
      \[l_4 : \text{some}_q(y, h_3, h_4)\]
      \[l_5 : \text{dog}_n I(y)\]
      \[h_2 = q l_2, h_3 = q l_5\]
    c. \[l_1 : a_1 \text{every}_q(x), \text{RESTR}(a_1, h_1), \text{BODY}(a_1, h_2)\]
      \[l_{21} : a_{21} : \text{black}_a(e_1), \text{ARG1}(a_{21}, x_1)\]
      \[l_{22} : a_{22} : \text{cat}_n I(x_2)\]
      \[l_3 : a_3 : \text{loved}_v I(e_2), \text{ARG1}(a_3, x_3), \text{ARG2}(a_3, y_1)\]
      \[l_4 : a_4 : \text{some}_q(y), \text{RESTR}(a_4, h_3), \text{BODY}(a_4, h_4)\]
      \[l_5 : a_5 : \text{dog}_n I(y_2)\]
      \[h_2 = q l_2, h_3 = q l_5\]
      \[x = x_1, x = x_2, x = x_3, x_1 = x_2, x_2 = x_3, y = y_1, y = y_2, y_1 = y_2, l_{21} = l_{22}\]

Intersective modification is achieved by sharing labels across multiple predications (e.g., \(l_2\) in (23b)), meaning that in any fully specific LF \(\text{black}(e_1, x)\) and \(\text{cat}(x)\) are connected with \(\land\). Sharing labels provides a more compact description. Observe also that mrs adopts naming conventions for the predicate symbols, based on the word lemmas and their tags. We will shortly adopt similar naming conventions in representing the compositional content of gesture.

The mrs (23b) is a notational variant of the rmrs (23c): 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 POS tagger: one can build its semantic component simply by deriving lexical predicate symbols from the word lemmas and their tags and ensuring that the predication for each word receives a unique label and argument, as given in (24):

(24)  a. Every_AT1 black_JJ cat_NN1 loved_VVD some_DD dog_NN1
    b. \[l_1 : a_1 : \text{every}_q(x),\]
      \[l_{21} : a_{21} : \text{black}_a(e_1),\]
      \[l_{22} : a_{22} : \text{cat}_n I(x_2)\]
      \[l_3 : a_3 : \text{loved}_v I(e_2),\]
      \[l_4 : a_4 : \text{some}_q(y),\]
      \[l_5 : a_5 : \text{dog}_n I(y_2)\]

Semantic relations, sense tags and the arity of the predicates are missing from (24b) because the POS tagger doesn't reveal information about syntactic constituency, word sense or lexical subcategorisation. But the rmrss (23c) and (24b) are entirely compatible, the former being more specific than the latter.

With this brief overview of rmrs in place, we'll now use it to represent the content of gesture as derived from its form. The form of gesture underspecifies the predicate symbols in its logical form and their arity (as well as all semantic dependencies and scope), but the
possible interpretations of the gesture’s features are constrained by the *iconicity*: each of the attribute-value elements may convey a specific, analogous piece of descriptive content. One cannot, for example, interpret the circular motion in Figure 2 as denoting a square concept.

With **RMR**S we can formalise this principle in straightforward steps that are analogous to those used in the semantic component of a POS tagger. Each attribute-value element yields an *elementary predication* that must be resolved to a particular formula in the LF of gesture in context. Following the design of the semantic component to the POS tagger above, we introduce a convention that reads this predication directly off the feature structure, as in (25):

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

Here, \( l_1 \) is a uniquely indexed 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 \( \text{hand\_shape\_asl-a} \) underspecifies reference to a property that the entity \( i_1 \) has and that can be depicted through the gesture’s fist shape.

We treat attributes with set values (e.g., the movement-direction attribute in (21)) like intersective modification in language (e.g., see the analysis of *black cat* in (23b)). This captures the intuition that the different aspects of the direction of movement in (21) must all depict properties of the same thing in interpretation. Thus the RMR representation of the iconic gesture (21) is (26) (where we have substituted the relevant variable and label equalities that are rendered by intersective modification into the relevant predications):

(26) \( l_1 : a_1 : \text{right\_hand\_shape\_asl-a}(i_1), \)
\( l_2 : a_2 : \text{right\_finger\_dir\_down}(i_2), \)
\( l_3 : a_3 : \text{right\_palm\_dir\_left}(i_3), \)
\( l_4 : a_4 : \text{right\_traj\_sagittal-circle}(i_4), \)
\( l_5 : a_{51} : \text{right\_move\_dir\_iterative}(i_5), \)
\( l_5 : a_{52} : \text{right\_move\_dir\_clockwise}(i_5), \)
\( l_6 : a_6 : \text{right\_loc\_central-right}(i_6) \)

We observed in Section 2.2 that any individual that is introduced in gesture must be bridging related to an available antecedent introduced in speech. In line with the compositional semantics of definite descriptions proposed by Chierchia (1995), we include bridging constraints in the logical form of gesture. These underspecified bridging relations can be added to the RMR that’s produced by the grammar. We assume this addition occurs outside the grammar, however, since bridging relations don’t affect semantic composition from syntax. Rather, they impose constraints on discourse update, stipulating that a particular relation to a particular available antecedent must be found as part of the process of constructing a coherent logical form of the discourse. Thus we assume the following monotonic schema as part of the glue logic for discourse update, where using the notation from Kamp and Reyle (1993), \( R =? \) and \( x =? \) mean that the values of the relation \( R \) and antecedent \( x \) are unknown:

- **Individuals in Gesture:**
  \[ [G](l : a : P(i) \land \text{individual}(i)) \rightarrow [G](l : a' : R(x) \land ARG1(a', i) \land R =? \land x =?) \]

\[^{11}\text{This suggests an analogy that the form of an iconic gesture is treated like a bag of words. Kopp et al. (2004) liken it to a bag of morphemes, on the grounds that the resolved interpretations of the features can’t be finitely enumerated. But word senses can’t be enumerated either (Pustejovsky, 1995); hence our analogy with words is also legitimate.}\]
In words, this rule stipulates that if an individual variable $i$ is part of the logical form of gesture, then it must be related to an available antecedent $x$ with a (licensed) bridging relation $R$ (and this bridging relation will be joined via conjunction to the original predication over the individual variable). The form of the gesture doesn’t identify the antecedent or the bridging relation; rather, these values will be determined by other information. Naturally, $R$ may resolve to equality, in which case we get co-reference between the individual in the gesture and the individual mentioned in context. But $R$ could also resolve to values other than equality (e.g., the relation between a waiter and his tray in (3)).

The introduction of the modal operator $[\mathcal{G}]$ into the base language $\mathcal{L}_{base}$ of fully-specific logical forms means that the grammar must introduce it into an appropriate scopal position. This is easy: one simply extends the form-meaning relation to include this modal operator, outscoping all the predications that are introduced by the values of our six features of gesture. In other words, the revised (underspecified) logical form of (21) is (27):

\begin{align}
(27) & \quad \begin{align*}
  l_0 &: a_0 : [\mathcal{G}](h), \\
  l_1 &: a_1 : \text{right hand shape asl-a}(i_1), \\
  l_2 &: a_2 : \text{right finger dir down}(i_2), \\
  l_3 &: a_3 : \text{right palm dir left}(i_3), \\
  l_4 &: a_4 : \text{right traj sagittal circle}(i_4), \\
  l_5 &: a_{51} : \text{right move dir iterative}(i_5), \\
  l_5 &: a_{52} : \text{right move dir clockwise}(i_5), \\
  l_6 &: a_6 : \text{right loc central-right}(i_6), \\
  h & \geq l_j, \text{ for } 1 \leq j \leq 6
\end{align*}
\end{align}

The label constraints are expressed using $\geq$ rather than $=_{q}$; this is because the predications may resolve to scopal modifiers. Furthermore, if the context-resolved interpretation of a gesture introduces a new individual that is not bound to any individual in the synchronous speech, then the quantifier that binds this individual must be outscoped by $[\mathcal{G}]$ so that availability (see Definition 9) and semantic interpretation (see Definition 5) constrain anaphoric dependencies correctly—i.e., the individual introduced in the gesture cannot be an antecedent to a pronoun in subsequent speech. This scopal constraint can also be expressed as part of discourse update (although we forego formal details here).

Following Kopp et al. (2004), we do not believe that the specific interpretations can be finitely enumerated. We therefore represent the possible resolutions of the underspecified predicates via a hierarchy of increasingly specific properties in logical form, as in Figure 3. While some of the leaves in this hierarchy correspond to fully specific interpretations, the creative use of metaphor makes interpretation open-ended (Lakoff and Johnson, 1981). Therefore, some of the hierarchy’s leaves correspond to more vague interpretations, and we envisage that either the speaker and hearer sometimes settle on a coherent but vague interpretation, or additional logical axioms will resolve a vague interpretation to a more specific one in the particular discourse context.\textsuperscript{12} We assume that the hierarchy in Figure 3 encapsulates a set of meaning postulates which reside outside the grammar, since they have no impact on the composition of the rmrs from the gesture’s form.

Let’s examine this hierarchy in a bit more detail. To capture the metaphorical contribution of the fist to the depiction of the process in (2), we resolve right_hand_shape_asl-a as depicting

\textsuperscript{12}We don’t give an exhaustive account of the meaning postulates for gestural predicate symbols here, since we believe that this must rest on a better understanding of metaphor, and this is a matter for future research.
Figure 3: Part of the hierarchy of underspecified and fully-specified predications for hand_shape_asl-a.

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). At the same time, we can capture the contribution of the fist to the depiction of (19) by resolving right_hand_shape_asl-a as depicting something held, in particular a marker-point x indicating a designated location on the mouse’s spinning wheel. Finally, all underspecified predications are resolvable to ⊤—the formula representing validity—since they might not contribute to meaning in context (e.g., the clockwise motion in (2)). 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, and thus the gesture accompanying (19) could have been performed with a flat hand instead of a fist.

Crucially, Figure 3 reflects the fact that, like all dimensions of iconic gesture, the fist shape doesn’t fully determine how many entities are related by its specific semantic interpretation. The predications in Figure 3 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 predicate 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 predicate, and therefore l : a : marker-point(x), ARG1(a, y) is unsatisfiable. One can also underspecify the position of a variable in a predication: ARGN(a, x) means that x is an argument to the (unique) predicate whose anchor is a, but its argument position is unspecified (so ARGI(a, x) ⊢ ARGN(a, x)).

The divergent resolutions of the same gesture in different contexts highlight how we capture insights about gesture meaning from previous work: we represent gesture meaning compositionally and iconically, yet in an underspecified form that requires context to resolve. You can compare predications like hand_shape_asl-a to Kopp et al.’s (2004) image description features—an abstract representation, distinct from form and content, that captures gesture
meaning. By using rmrs, we can reinterpret these representations as analogous, both formally and substantively, to the underspecified meanings that computational semanticists already use to represent speech. In particular, as we show in Section 4, we can therefore build reasoning mechanisms that combine information from speech and gesture to derive a single, overall coherent logical form of discourse.

To account for the integrated interpretations of language and gesture, we assume that the inventory of predicates in the fully-resolved logical forms for language and for speech are the same. Thus the fully-specific predicate symbols (or conjunctions thereof) that the predicate symbols in the rmrs of iconic gesture can resolve to are the same as those used to represent linguistic semantics. For example, one meaning postulate will stipulate that traj\_sagittal\_circle(x) can resolve to _circle\_n\_1(x), the predicate symbol associated with one sense of the noun circle. This allows us to fully integrate the contextually-determined content of both gesture and speech; they are interpreted by the same model and the same contexts of evaluation. This is crucial for a coherent semantic interpretation of rhetorical relations that connect speech and gesture (see Section 2.4).

Let’s now consider the compositional semantics of deictic gesture. In words, (28) states that deictic gesture denotes an individual which is spatially located at the coordinates v(\vec{p}), where \vec{p} is the physical point in space-time that is 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).

\[(28)\]
\[
l_0 : [G](h_1),
\]
\[
l_1 : a_1 : \text{deictic}\_q(x), \text{RESTR}(a_1, h_2), \text{BODY}(a_1, h_3)
\]
\[
l_2 : a_2 : \text{loc}(e), \text{ARG1}(a_2, x), \text{ARG2}(a_2, v(\vec{p}))
\]
\[
h_2 = q \downarrow l_2
\]

We gave the full dynamic semantics of the predicate loc in Section 2.1. The compositional semantics for gesture in (28) does not specify any semantic relation between the the individual x that’s introduced by the deictic gesture and the individual y that’s introduced by its synchronous linguistic phrase. But clearly, the meaning of a deictic gesture combined with the relevant phrase should convey some semantic relationship between them. We argued in Section 1 that this semantic relationship is underspecified. For example, the relationship could be equality (e.g., when saying the man and pointing at a spatio-temporal region \vec{p} where context establishes that a man is physically located); or it could be ‘depicted’ or shown (e.g., (6)). In Section 3.3, we motivate introducing the underspecified relation deictic\_rel(y, x) in the semantic component of a construction rule in the grammar, which joins deictic gesture to its synchronous phrase in speech. (28) also underspecifies the semantic scope of the quantifier deictic\_q and [G]. As discussed earlier, in order to capture anaphoric dependencies correctly we assume that discourse update imposes a constraint that [G] outscopes deictic\_q if deictic\_rel resolves to something other than equality.

3.3 Combining Speech and Gesture in the Grammar

Like Kopp et al. (2004), we believe that gesture should combine with synchronous speech within the grammar, producing a single derivation tree. An alternative would be to combine

\[\text{This is strictly speaking an idealisation; some concepts can be expressed easily in language but are hard to depict in gesture and vice versa. However, all we need technically is the predicate symbols for both language and gesture to be interpreted by the same function in the model.}\]
them entirely as a byproduct of discourse update. But Kopp et al. (2004) argue for combining speech and gesture in the grammar on the grounds that it smoothes the transition between sentence realisation and gesture formation in generation. And our motivation for doing so stems from considerations involving interpretation.

First, we assume, as is standard in dynamic semantic theories, that any aspect of form that conveys meaning should be encoded in the grammar, so that discourse update can continue to function with access only to the compositional semantic representations of situated utterances, and no access to their form directly. Here, we argue that synchrony is an aspect of form that conveys meaning, and consequently its form meaning relation should be encoded in the grammar. Consider iconic gesture first. We observed in Section 2.3 that the gesture must be rhetorically connected to its synchronous speech. In fact, certain rhetorical relations are excluded for connecting them; e.g., Contrast. To see this, contrast (29a) with (29b).

(29) a. The cable is made of highly stretchable material.
   *Both hands in a fist shape, in contact, and at the centre of the speaker’s torso, move apart in a horizontal line.*

b. The cable is made of steel.
   *Gesture is the same as in (29a).*

c. The cable is made of steel. It is about 20cm long.

d. The cable is made of steel. But it has elastic properties.

The gesture in (29a) depicts the elastic properties of the cable. But in (29b), the same gesture provides Background information to the proposition that the cable is made of steel, providing its length, making its interpretation equivalent to that of (29c). Crucially, it’s not possible to interpret the gesture in (29b) in the same way as it’s interpreted in (29a) in spite of their same form. In other words, one cannot utter (29b) and simultaneously gesture that the cable is elastic or stretchable. Our conjecture is that this interpretation is blocked because synchrony excludes Contrast, and interpreting the gesture as depicting elasticity creates a situation where the content of the gesture violates expectations stemming from the content of the speech, which demands a Contrast relation for connecting them, as in (29d). A gesture and its synchronous utterance also cannot be connected with Disjunction: that is, the message conveyed cannot be that the content conveyed by the utterance is true or the content conveyed by the gesture is true.

If we are right that certain rhetorical relations are always excluded from connecting a gesture to its synchronous speech, then synchrony is an aspect of form that conveys semantic information that’s similar to that conveyed by a sense-ambiguous discourse connective in language: they both introduce rhetorical connections between their syntactic complements, but do not fully specify the value of the relation. This therefore suggests that synchrony and its contribution to compositional semantics should be encoded in the grammar, just as it is for discourse connectives.

Our second motivation for assuming that speech and gesture are combined in a single derivation in the grammar comes from evidence that the very concept of synchrony itself should be defined in terms of the syntactic constituency of the utterance, and not just in terms of time. Consider sentence (30):

---

14These remarks are restricted to co-verbal gesture, since one clearly can utter *The cable is made of elastic but* and then finish the communicative act by performing a gesture depicting elastic properties.
The cable, which is made of steel, abruptly and unexpectedly broke.

It is perfectly acceptable to perform a gesture that is cotemporal with uttering the word *abruptly*, where the gesture depicts the cable abruptly breaking (e.g., the two hands are in a fist shape and touching, as if holding a cable, and then suddenly pulled apart). But one cannot perform the gesture before one utters one of the pre-head modifiers to *break*; e.g., one could not perform the gesture at the same time as *steel* is uttered. In terms of the timeline, uttering *steel* occurs a few milliseconds before *abruptly*, and yet gesturing the breakage at the same time as the former word is ill-formed, and at the same time as the latter word is well-formed (so long as making *abruptly* the focus is coherent). Thus constraints on synchrony seem to be governed by syntactic constituency, among other things. Encoding synchrony in the grammar would allow us to express such well-formedness constraints.

For deictic gesture, it’s just as vital to record the semantic contribution of synchrony. As with iconic gesture, identifying the phrase in speech that a deictic gesture is synchronous with involves a complex mix of information involving syntactic constituency, intonation and time of performance (Bellik, 1995, 1997, Oviatt, 1999, Kühnlein et al., 2002). Specifying in detail the constraints on synchrony is beyond the scope of this paper. But nevertheless one thing is clear: synchrony is a dimension of *form* that contributes to meaning—it identifies which individual in speech is in an (underspecified but restricted) semantic relation to the individual denoted by the deictic gesture.

While we don’t give details of the grammar here, we make some general assumptions about its features. Following Johnston (1998a), we assume that all syntactic constituents in the grammar are stamped with a representation of *when* the constituent is performed relative to other communicative acts in the derivation tree. In addition, the modality of delivery is already marked via the type system (e.g., *np* and *vp* are subtypes of *speech* and *iconic_gesture* and *deictic_gesture* are subtypes of *gesture*). Constraints on synchrony will ultimately be defined in terms of when the phrase is uttered and the gesture performed, and the syntactic properties of the phrase (including information about its pitch accents, boundary tones and syntactic dependencies). Previous research shows that the constraints on synchrony are complex and controversial (Schegloff, 1984, Oviatt et al., 1999, Sowa and Wachsmuth, 2000, Quek et al., 2002). We will therefore simply assume that whatever the constraints on synchrony, they can be expressed in a constraint-based grammar (e.g., see Cohen et al. (1997), Johnston (1998a,b)).

While we remain neutral about constraints on the *form* of synchrony, we make specific assumptions about what it contributes to *meaning*. First, we assume that the construction rule for combining iconic gesture \(I\) and a syntactic constituent \(S\) from speech contributes its own semantics. Namely, it introduces (31) to the RMRS of the mother node, where \(h_s\) is the top-most label of the content of the ‘speech’ daughter, and similarly \(h_g\) is the top-most label of the gesture (e.g., \(h_g = l_0\) if the gesture were (21) with corresponding RMRS (27)):

\[
(31) \quad l : a : iconic\_rel(h_s), ARG1(a, h_g)
\]

The underspecified predicate *iconic\_rel* must then be resolved via discourse update to a specific rhetorical relation, where *iconic\_rel* \((\alpha, \beta)\) entails \(¬\text{Contrast}(\alpha, \beta)\) and \(¬\text{Disjunction}(\alpha, \beta)\).

Similarly, we assume that the construction rule that combines a deictic gesture with an NP contributes its own semantics. Namely, it introduces the labelled predication (32), where \(l_2\) is the same label as the label of the spatial condition \(loc(e, x, v(\vec{p}))\) introduced by the RMRS of the deictic gesture on its semantic index \(x\) (see (28)), and \(y\) is the semantic index of the NP.

33
(32) \[ l_2 : a_{21} : \text{deictic}_\text{rel}(x), \text{ARG1}(a_{21}, y) \]

So, for instance, the rmrs for a multimodal constituent consisting of the NP the dog combined with a deictic gesture is the following:

(33) \[ l_0 : [G](h_1), \]
\[ l_1 : a_1 : \text{deictic}_\text{q}(x), \text{RESTR}(a_1, h_2), \text{BODY}(a_1, h_3) \]
\[ l_2 : a_2 : at(x), \text{ARG1}(a_2, v(\vec{p})) \]
\[ l_2 : a_{21} : \text{deictic}_\text{rel}(x), \text{ARG1}(a_{21}, y) \]
\[ l_3 : a_3 : _\text{the}_\text{q}(y), \text{RESTR}(a_3, h_4), \text{BODY}(a_3, h_5), \]
\[ l_4 : a_4 : \text{dog}(y) \]
\[ h_2 = q l_2, h_4 = q l_4 \]

Deictic 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.

4 Establishing Coherence through Default Inference

SDRT models 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. So far, SDRT’s discourse update has been used to interpret anaphora and elided constructions in speech. Here, we indicate how discourse update can resolve the underspecified meaning of both speech and gesture.

Discourse update in SDRT involves nonmonotonic reasoning with both linguistic information and non-linguistic information. It is defined in a constraint-based way: Where \( \phi \in L_{ulf} \) represents the (underspecified) old content of the context, which is to be updated with the (underspecified) new content \( \psi \in L_{ulf} \) of the current utterance, the result of update will be a formula \( \chi \in L_{ulf} \) that is entailed by both \( \phi, \psi \) and the consequences of a nonmonotonic logic—known as the GLUE LOGIC—when \( \phi \) and \( \psi \) are premises in it.

The glue logic axioms encapsulate inferences about which type of 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 \( \chi \). Formally, the glue logic axioms typically have the following form, where \( A > B \) can be read as If \( A \) then normally \( B \).\(^{15}\)

- **Glue Logic Schema:** \( (\lambda : ?(\alpha, \beta) \land \text{some stuff}) > \lambda : R(\alpha, \beta) \)

In words, this says: if \( \beta \) is to be connected to \( \alpha \) with a rhetorical relation, and the result is to appear as part of the logical scope labelled \( \lambda \) (in other words \( F(\lambda) \) includes \( R(\alpha, \beta) \) as a conjunct for some value of \( R \)), but we don’t know what the value of that relation is yet, and moreover “some stuff” holds of the content labelled by \( \alpha \) and \( \beta \), then normally the rhetorical relation is \( R \). The “some stuff” is derived from the (underspecified) logical forms (expressed in \( L_{ulf} \)) that \( \alpha \) and \( \beta \) label, and the rules are justified either on the basis of underlying linguistic knowledge, world knowledge, or knowledge of the cognitive states of the conversational participants.

\(^{15}\)Note that for the sake of simplicity, we sometimes use the notational variant \( l : P(x, y, z, \ldots) \) instead of \( l : a : P(x), \text{ARG1}(a, y), \text{ARG2}(a, z), \ldots \).
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 $\alpha$ lead to events of the sort described by $\beta$:

- **Narration**: $\lambda : ?(\alpha, \beta) \land \text{occasion}(\alpha, \beta) > \lambda : \text{Narration}(\alpha, \beta)$.

Schank’s (1977) scripts attempted to capture information about which eventualities occasion which others. We assume that this type of information is encoded in default rules for inferring occasion. For example, we assume that the underspecified logical forms of the clauses in (17) verify the antecedents of a default axiom whose consequence is $\text{occasion}(\alpha, \beta)$, yielding the logical form (17'):

(17) $\pi_1$. John went out the door.
    $\pi_2$. He turned right.

(17') $\pi_0 : \text{Narration}(\pi_1, \pi_2)$

Indeed, the glue logic axioms that do this should be neutral about about sentence mood, so that they also predict that the imperatives in (34) are connected by Narration:

(34) Go out the door. Turn right.

Thus by the semantics of Narration, the logical form for (34) 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. This is exactly the semantics we desire for the example (35) involving gesture, which is taken from the same dialogue as utterance (1):

(35) 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 logical form of (35): (a) the utterance and gesture are related by Narration; and (b) the content of the gesture resolves to turn right in this context. We’ll see how shortly.

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

- **Explanation**: $(\lambda : ?(\alpha, \beta) \land \text{cause}_D(\beta, \alpha)) > \lambda : \text{Explanation}(\alpha, \beta)$

Note that $\text{cause}_D(\beta, \alpha)$ means that there’s evidence in the discourse that $\beta$ caused $\alpha$, not that it actually did cause $\alpha$; the latter causal relation would be inferred if Explanation is inferred. The formula $\text{cause}_D(\beta, \alpha)$ 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 $\alpha$ and $\beta$. We assume that there will be such a monotonic axiom for inferring $\text{cause}_D(\pi_2, \pi_1)$ for the discourse (36), which bears obvious analogies to the embodied utterance (2).

\footnote{We did not give formal details of 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.}
There are low-level phonological errors which tend not to get reported.

They are caused by processes which are continuous and subconscious.

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 one can infer a particular rhetorical relation and from its semantics infer how the underspecified conditions of the utterance or gesture are resolved. Alternatively, there are cases where the premises for inferring rhetorical relations are not satisfied by the underspecified compositional semantics. In this case, all is not lost: one can resolve the underspecified content so as to support 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 of the rhetorical relations one can infer from it (see Asher and Lascarides (2003) for details).

Let’s illustrate the inference from underspecified content to complete LF with a specific example.

(19) The mouse ran round the wheel for a few minutes.

Gesture is as illustrated in Figure 2.

(37)

\[ l'_{1}:_{\text{the.q}}(x,h_{1},h_{2}), \]
\[ l'_{2}:_{\text{mouse}}(x'), \]
\[ l'_{3}:_{\text{run}}(e',x'), \]
\[ l'_{4}:_{\text{on}}(e,y), \]
\[ l'_{5}:_{\text{for}}(e,z), \]
\[ l'_{6}:_{\text{the.q}}(y,h_{3},h_{4}), \]
\[ l'_{7}:_{\text{wheel}}(y), \]
\[ l'_{9}:_{\text{a-few.q}}(z,h_{5},h_{6}), \]
\[ l'_{10}:_{\text{minutes}}(z), \]
\[ h_{1}=q l'_{2}, h_{3}=q l'_{7}, h_{5}=q l'_{10} \]

As described in Section 3.2, the grammar yields (27) for the content of the gesture, (37) for the content of the speech, and the construction rule that combines them contributes the predication \( l : \text{iconic\_rel}(h_{a},l_{0}) \), where \( h_{a} \) is in a qeq condition to the label \( l'_{5} \) of \( _{\text{run}}(e',x') \), and \( l_{0} \) labels the scopal modifier \([G]\) in (27). Producing a fully specific LF from this therefore involves, among other things, resolving the underspecified predications in (27), and the underspecified predicate \( \text{iconic\_rel} \) must resolve to a rhetorical relation that’s licensed by it (so not \( \text{Contrast} \) or \( \text{Disjunction} \)).

We introduce a glue logic axiom which captures the following intuition: if two propositions are rhetorically related somehow, and they both describe movement events involving the same participant that can occur simultaneously, then there is evidence in the discourse that these events are in a subtype relation (following Asher and Lascarides (2003), we assume a notational convention where \( e_{\alpha} \) and \( e_{\beta} \) are respectively the semantic indices of \( \alpha \) and \( \beta \)):

- **Movement Subtypes:**

  \[ (\lambda :? (\alpha,\beta) \land l_{1}:a_{1}:\text{movement}(e_{\alpha}) \land l_{1}:a_{2}:\text{pred}(i) \land \text{ARGn}(a_{2},x) \land \]
  \[ l'_{1}:a'_{1}:\text{movement}(e_{\beta}) \land l'_{1}:a'_{2}:\text{pred}'(i') \land \text{ARGn}(a'_{2},x) \land \]
  \[ \text{temporally-compatible}(e_{\alpha},e_{\beta}) \rightarrow \text{Subtype}_{D}(\alpha,\beta) \]
Note the rule underspecifies whether the participant $x$ is realised as an argument to the movement predicate symbol or via the equivalent of intersective modification (such an individual can therefore be realised via a PP that modifies the verb phrase; e.g., the wheel is in this kind of dependency to \textit{ran} in (19)).

The predication \textit{Subtype}_D(\alpha, \beta) does \textit{not} entail that the events actually are in a subtype relation, only that there is evidence in the discourse for this. Asher and Lascarides (2003) use this predicate to infer \textit{Elaboration}:

- \textit{Elaboration}: $(\lambda :? (\alpha, \beta) \wedge \text{Subtype}_D(\alpha, \beta)) > \lambda : \text{Elaboration}(\alpha, \beta)$

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

Now let’s return to the situated utterance (19). Its underspecified content on its own is insufficient for inferring a rhetorical relation, for although the physical form of the gesture features (physical) movement, some of the specific interpretations that this movement licenses do not entail physical movement (e.g., the movement could be metaphorical, or indeed the movement predications could resolve to \textit{T} as explained in Section 3). However, one of the possible resolved meanings of the gesture is one which satisfies the axiom \textit{Movement Subtypes}, as shown in (38):

(38) \[
\text{[G]} \exists x w z (\text{marker-point}(x) \wedge \text{part-of}(x, y) \wedge \text{move}(e, x) \wedge \text{path}(e, w) \wedge \text{sagittal-circle}(w) \wedge \text{direction}(w, z) \wedge \text{clockwise}(z) \wedge \text{iterative}(z))
\]

This mirrors our earlier discussion about resolving the fist to a \textit{marker-point}. The bridging inference between this individual and the wheel \textit{y} from (37), as required by the axiom \textit{Individuals in Gesture}, is supported to the fact that the speech and gesture are rhetorically connected (by a relation licensed by \textit{iconic} \textit{rel}), making the wheel available (see Definition 10). Similarly, the bridging constraints on the path \textit{w} and direction \textit{z} are also bridging related to the wheel, as required.

If the gesture is interpreted this way, then Modus Ponens on \textit{Movement Subtypes} followed by Defeasible Modus Ponens on \textit{Elaboration} would lead to an inference of a particular rhetorical relation between the utterance and the gesture. Suppose that this is the \textit{only} possible resolved interpretation of the gesture that leads to an inference about which rhetorical relation connects the utterance and the gesture, or it leads to the rhetorical connections of the highest quality. Then discourse update in SDRT forces this specific interpretation (see Asher and Lascarides (2003) for formal details). Thus the gesture provides more information about the movement described in the utterance: the wheel is in a vertical plane (and fixed at a central point), and moves in a clockwise direction several times.

The analysis of (2) is similar to that of (19). However, the specific interpretation of the gesture in Figure 2, in which the underspecificed predications are interpreted as movement, cannot satisfy \textit{Movement Subtypes} this time, because the sentence is not about movement. So another specific interpretation is needed to support a particular rhetorical connection between the speech and gesture. As an alternative, the underspecified content of the gesture can resolve to denote a continuous, subconscious process which causes the phonological errors mentioned in (2), as shown in (39) where \textit{y} is the low-level phonological errors introduced in the speech:

(39) \[
\text{[G]} \exists x (\text{continuous}(x) \wedge \text{below-awareness}(x) \wedge \text{process}(x) \wedge \text{sustain}(e', x, y))
\]

The process \textit{x} meets the bridging constraints articulated in \textit{Individuals in Gesture} through the predicate \textit{sustain}, a possible resolved interpretation of the hand’s fist shape (see
Figure 3). This particular interpretation, like the interpretation of the sentences in (36), satisfies the antecedent of an axiom whose consequent is \( \text{Cause}_D(\beta, \alpha) \) (we forego formal details of the axiom here). And so by Defeasible Modus Ponens on Explanation, an Explanation relation is inferred between the utterance and its accompanying gesture. If this is the specific interpretation which maximises the quality of the connection between the constituents, then discourse update dictates that the logical form of the discourse resolves the interpretations this way.

Now consider (35):

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

While the gesture has movement features in its form, it cannot be interpreted as a physical movement which is temporally compatible with the (physical) movement of walking out the door. This is because if the gesture is interpreted as a physical movement then it must be a turning right (as this is the direction of the hand), and one cannot go out the door and turn right at the same time. However, a glue logic axiom for inferring occasion\((\alpha, \beta)\) is satisfied when the underspecified content of the gesture is resolved to depict physical movement (see earlier discussions of (17) and (34)). Thus there is a specific interpretation of the gesture from which one can nonmonotonically infer a Narration relation. Assuming that this specific interpretation is the one which leads to an inference of a rhetorical connection with maximum quality, discourse update constructs a logical form of the situated utterance which encapsulates this interpretation.

Now consider (29a) vs. (29b):

(29) a. The cable is made of highly stretchable material.
   Both hands in a fist shape, in contact, and at the centre of the speaker’s torso, move apart in a horizontal line.

b. The cable is made of steel.
   Gesture is the same as in (29a).

As discussed earlier, the gesture in (29a) depicts elasticity, while in (29b), it depicts length. For SDRRT to predict these interpretations, discourse update must predict that in (29a), interpreting the gesture as conveying information about length provides a less coherent interpretation than interpreting the gesture as conveying elasticity; it must predict the reverse for (29b).

So let’s examine (29a) first. As before, the underspecified content of the gesture on its own fails to satisfy the antecedent to any glue logic axiom. And so one must entertain alternative ways of resolving it to a specific interpretation, and explore which of these yields a rhetorical connection of maximal quality. Two glue logic axioms are at play for this. The first one specifies that when \( \alpha \) and \( \beta \) both describe states of the same individual, they are by default connected with Background (see Asher and Lascarides (2003) for the formal rule). The other rule, for inferring Depiction, is as follows, where \( \mathcal{K}_\alpha \) and \( \mathcal{K}_\beta \) are a gloss for the content that \( \alpha \) and \( \beta \) label:

- **Depiction:**
  \[
  (\lambda :?(\alpha, \beta) \land \text{speech}(\alpha) \land \text{gesture}(\beta) \land (\mathcal{K}_\alpha > \mathcal{K}_\beta) \land (\mathcal{K}_\beta > \mathcal{K}_\alpha)) > \lambda : \text{Depiction}(\alpha, \beta)
  \]
This rule is a particular instantiation of general rule in the glue logic, that the necessary consequences of a rhetorical connection are normally sufficient for inferring it.

Now, if one resolves the gesture in (29a) to an interpretation where it depicts elasticity, then the antecedents to both glue logic axioms are verified, yielding Background and Depiction. Moreover, because Depiction entails that they express nonmonotonically equivalent content, the common topic that is demanded by Background can be the content of the speech. If, on the other hand, one resolves the gesture to an interpretation where it depicts length, then the glue logic axioms support an inference to Background, but not to Depiction. This results in a logical form with fewer rhetorical connections, and also a more general common topic for the Background relation (something that’s roughly equivalent to “general properties of the cable”), which in turn must be added to the logical form. Asher and Lascarides (2003) argue that, all else being equal, an SDRS that includes fewer constituent labels and more relations is more coherent than one which contains more labels (as would be demanded by the introduction of a linguistically implicit common topic) and fewer rhetorical connections. Thus the pragmatically preferred interpretation of the gesture in (29a) is one which yields the Depiction and Background relations: namely, the interpretation where it depicts elasticity.

Now contrast this with the same gesture in (29b). If the cable is interpreted as conveying that it’s elastic, then as well as satisfying a glue logic axiom for inferring a Background connection, world knowledge dictates that the elasticity violates an expectation arising from its being made of steel. As such, the antecedent of a monotonic glue-logic axiom is satisfied, whose consequent stipulates (a) that the constituents are connected with Contrast, and (b) an explicit linguistic cue that indicates the contrast, such as intonation or the cue phrase but, is present. Asher and Lascarides (2003) argue that this monotonic axiom is needed to explain why (29d) is acceptable while (29e) is anomalous:

(29)   d. The cable is made of steel. But it has elastic properties.
   e. The cable is made of steel. ??It has elastic properties.

But inferring Contrast is inconsistent with the underspecified relation iconic_rel that already connects the speech and gesture; moreover inferring the explicit linguistic cue for contrast is inconsistent with the possible specific ways of resolving the gesture (for none of the predications can resolve to a cue phrase like but). So, for the situated utterance to be coherent, another interpretation of the gesture must be used. And as we mentioned, interpreting the gesture as conveying length achieves this, by supporting an inference to a Background relation.

Finally, let’s examine deictic gesture: discourse update must support inferences which resolve the underspecified relation deictic_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**: \((\text{deictic}_\text{rel}(x, y) \land \text{loc}(e, y, \vec{p}) \land P(x) \land P(y)) > x = y\)

The axiom Co-reference stipulates that if \(x\) and \(y\) are related by deictic_rel, and moreover, the individual \(y\) that is physically located at \(\vec{p}\) shares \(x\)’s properties, then normally the individuals \(x\) and \(y\) are equal. Other default axioms than Co-Reference can be articulated for inferring shown\((x, y)\) rather than \(x = y\) in the interpretation of (6).
5 Conclusion

We have provided a formal semantic analysis of iconic and deictic gesture which captures several compelling features from the descriptive literature. For instance, it predicts that speech and gesture together form a ‘single thought’. This is achieved by three features in our analysis. First, the logical forms of speech and gesture are represented in the same language and interpreted against the same models and indices of evaluation. Secondly, we use rhetorical relations to connect the logical form of iconic gesture to that of its synchronous speech. Finally, following Kopp et al. (2004), we integrated the analysis of speech and gesture in the grammar.

Our theory also models the observation that iconic gesture on its own doesn’t receive a coherent interpretation: this is achieved by assigning a very underspecified content to iconic gesture as revealed by its form. The analysis then models how reasoning about the rhetorical connection between gesture and its context and reasoning about how its underspecified content is resolved to a specific interpretation are logically co-dependent tasks. Finally, we exploited discourse structure and the dynamics in dynamic semantics to account for dependencies on co-reference across speech and gesture and among different gestures in the discourse.

One virtue in our analysis is that we demonstrate that existing mechanisms for representing the content of language can be exploited to model gesture as well. However, there is much future work that needs to be done. We have focussed here almost entirely on single speakers. But of course in conversation several agents contribute utterances and gestures, and further research is needed on how discourse update can model their interpretation. As part of this, we must encode how gestures affect and are affected by grounding (Clark, 1996). Luckily, there is a lot of research from cognitive psychology to draw on, providing a wealth of rich data that one can aim to model formally (e.g., Emmorey et al. (2000)). We would also like to examine other types of gesture, particularly what McNeill (1992) calls beats. Beats share a lot of semantic properties with intonation and focus in language, and we believe that recent advances in modelling the link between focus and intonation (e.g., Steedman (2000)) will provide a useful starting point for providing a formal model of beats.

Authors’ Addresses

Alex Lascarides, Matthew Stone,
School of Informatics, Computer Science,
University of Edinburgh, Rutgers University,
2, Buccleuch Place, 110 Frelinghuysen Road,
Edinburgh, EH8 9LW, Piscataway NJ 08854-8019,
Scotland, UK.
alex@inf.ed.ac.uk matthew.stone@rutgers.edu

References


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


