

# Aligning Speech and Co-speech Gesture in a Constraint-based Grammar



# Aligning Speech and Co-speech Gesture in a Constraint-based Grammar

*Katya Alahverdzhieva*[1]*, Alex Lascarides*[1]*, and Dan Flickinger*[2]
[1] School of Informatics, University of Edinburgh, UK
[2] Center for the Study of Language and Information, Stanford University, USA

## ABSTRACT

This paper concerns the form-meaning mapping of communicative actions consisting of speech and improvised co-speech gestures. Based on the findings of previous cognitive and computational approaches, we advance a new theory in which this form-meaning mapping is analysed in a constraint-based grammar. Motivated by observations in naturally occurring examples, we propose several construction rules, which use linguistic form, gesture form and their relative timing to constrain the derivation of a single speech-gesture syntax tree, from which a meaning representation can be composed via standard methods for semantic composition. The paper further reports on implementing these speech-gesture construction rules within the English Resource Grammar (Flickinger 2000). Since gestural form often underspecifies its meaning, the logical formulae that are composed via syntax are underspecified so that current models of the semantics/pragmatics interface support the range of possible interpretations of the speech-gesture act in its context of use.

*Keywords:*
*co-speech gesture, constraint-based grammar, compositional semantics, underspecification*

## 1 INTRODUCTION

In face to face conversation, people exchange information via a range of meaningful and visibly accessible communication channels (Goffman 1963); in particular they use "visible bodily actions" (Kendon

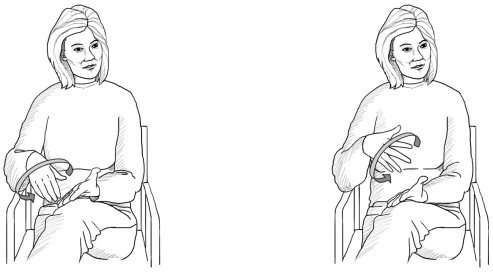

Figure 1: Gesture Depicting Mixing Mud, example (1)

2004). For instance, in utterance (1),[1] extracted from a conversation where the speaker is describing installing a drywall (Loehr 2004),[2] the speaker performs a circular movement with the right hand over the left palm (see Figure 1) along with the spoken utterance. Both the speech and the hand movement are relevant for the conveyed meaning of mixing mud, and both are produced and perceived as a coherent idea unit (McNeill 1992).

(1) So he mixes [$_N$mud] …

In this article, we analyse signals like (1), in which the hand is spontaneously used to convey meaning in tandem with speech. In the literature, these hand signals are known as *co-speech gesture*, *co-verbal gesture* or *gesticulation* (e.g., Kendon 1972). In *depicting/referential* gestures, the form of the hands visually characterises a salient feature of the referent. The depiction could be *iconic* (McNeill 1992) (e.g., in

[1] We adopt the following conventions in utterance transcriptions: the part of the speech signal that is simultaneous with the expressive phase of the gesture, the so-called stroke, is underlined. We include words that start or end at midpoint in relation to the gesture phase boundaries. The pitch accented words are shown in square brackets with the accent type in the left corner: PN (pre-nuclear), NN (non-nuclear) and N (nuclear).

[2] For this and for all subsequent examples that are cited as Loehr (2004), we are grateful to Daniel Loehr who kindly provided us with an annotated corpus of speech and co-speech gesture. We used this corpus to study depicting gestures.

(1) the hands perform a rotating movement to depict the mud being mixed), or *metaphoric* (McNeill 1992) (e.g., a rotating hand while saying "This was a long, boring process" can designate an iterative process). In *deixis/pointing* gestures, the hand points to a region in space so as to identify the referent's location in Euclidean space. The pointing can be *concrete* (McNeill 1992), as when pointing to something that's physically present in the communicative situation. It can also be *abstract* (McNeill 1992): the referent is a virtually created object in the gesture space just in front of the speaker, and its location in the gesture space constrains its physical location; e.g., a speaker, while describing her apartment that's on the other side of town, extends her right hand to the right periphery while saying "The bedroom is on the right". Formless flicks of the hand, beating the time along with the rhythm of the speech are known as *beats*. The current analysis focusses on depicting and pointing co-speech gestures.

We adhere to current theories of gesture (Cassell *et al.* 1999; Lascarides and Stone 2009a; Pfeiffer *et al.* 2013), in that we assume that co-speech gesture can affect the truth-conditional content of the speech-and-gesture action. Both deictic gestures and iconic representations say something about the world and as such they have propositional content; this extends to pictorial representations as well (Abusch 2014; Grzankowski 2015).

Our paper contributes to the existing approaches to integrating the contents of speech of co-speech gesture in a single semantic unit (McNeill 1992; Kendon 2004; Bavelas and Chovil 2006; Engle 2000; Giorgolo 2012) in that we explore the coordination patterns of the two modalities, we formalise them within an integrated grammar and we spell out the gesture's semantic contributions to the proposition that is conveyed by the speech-gesture action. The main challenges are twofold: on the one hand, the gesture signal is massively ambiguous (Lascarides and Stone 2009a); on the other, the speech-gesture integration is not a free-for-all, in that the *form* of the speech-gesture action rules out certain interpretations of it, whatever its context of use. To illustrate gesture's ambiguity, consider again the hand movement in (1). Taken out of its speech context, this gesture could be a depiction of a circular movement (e.g., the turning of a wheel), or it could refer to the object being rotated (e.g., the wheel itself), or it could refer to an iterative process. It is only via context that gesture receives a spe-

cific meaning: the content conveyed by the rotating movement while saying "He mixes mud" is distinct from that while saying "It's a huge, long boring process".

The form of a deictic gesture is also imprecise on the region pointed out by the hand and what is being designated (Kühnlein *et al.* 2002): when pointing in the direction of a book with an extended index finger, does the deictic gesture identify the physical object book, the book's content, or the location of the book—e.g., the table?

This ambiguity notwithstanding, the form of the gesture, abstracted away from its context of use, conveys some meaning, no matter how incomplete it might be. A depicting gesture, by the definition of iconicity, must support a perceptual resemblance between the gesture's form and its denotation (Kendon 2004; Kopp *et al.* 2007): i.e., the gesture's movement, hand shape etc. visualise qualitative characteristics of the referent. Deixis, on the other hand, indexes spatial reference in Euclidean space by projecting the hand to a region that is proximal or distal in relation to the speaker's location (e.g., Levinson 1983). Through deictic gestures, people anchor the referents in their utterances to the physical context (Kaplan 1989). This difference between depicting gestures and deictic gestures is accounted for in how we model the form-meaning mapping, and we also support the analysis of gestures that are *both* deictic and depictive simultaneously (and so inherit the characteristics of both gestural types).

### *Outline*

This article is structured as follows: in Section 2, we discuss the ambiguous form-meaning mappings of the speech-and-gesture signal, assuming a coherence-based pragmatic theory. In Section 3, we introduce examples to motivate a grammar-based approach to co-speech gesture. We then proceed with a discussion of related work and our distinct contribution (Section 4). In Section 5, we discuss how to formally represent gesture form and map this form to (underspecified) meaning. In Section 6, we propose domain-independent grammar rules which are based on the empirically extracted generalisations. Section 7 reports on the grammar implementation and evaluation.

## 2 AMBIGUOUS FORM–MEANING MAPPING

There is a right balance to be struck between constraining the mapping from form to meaning, while ensuring that existing pragmatic theories will support inferring the context-specific interpretations from the underspecified meanings derived only from form. The aim of this section is to use examples of speech-gesture actions to motivate one way of striking that balance. We first introduce an existing coherence-based model of pragmatics, which we assume underlies the inferences from the meaning that's derived from form alone to a preferred pragmatic interpretation in context. We then use this to motivate speech-gesture attachment ambiguities by illustrating how each syntax tree supports a different interpretation of the speech-and-gesture action, given the assumed pragmatics model. We also argue that licensed attachments are constrained, despite the multiple ways co-speech gestures can relate to speech.

### 2.1 *Pragmatic Theory Background*

In this paper, we assume a coherence-based model of the semantics/pragmatics interface as discussed in the literature of discourse interpretation (e.g., Hobbs 1985, Kehler 2002). The main principle of a coherence-based pragmatic theory is that discourse content is dependent on *coherence relations*—e.g., Elaboration, Explanation, Contrast, Contiguity—which link the meaning of its segments together. Identifying coherence relations is a defeasible process, informed by the compositional and lexical semantics of the units and contextual information such as real-world knowledge.

For instance, the pragmatic interpretation of the discourse in (2) involves the following contents: Max fell, John pushed Max, and the latter explains the former (so the pushing caused the falling and hence preceded it).

(2) Max fell. John pushed him.

Using the notation of Segmented Discourse Representation Theory (SDRT, Asher and Lascarides 2003), as shown in (3), this is represented as a rooted hierarchical set of labels—each label corresponds to a discourse segment—with each label associated with some content: $\pi_1$ is associated with the content that the event $e_1$ of Max $m$ falling hap-

pened before now; segment $\pi_2$ with the content that the event $e_2$ of John $j$ pushing $x$, where $x$ is identical to $m$, happened before now; and the (root) segment $\pi_0$ stipulates that $\pi_2$ explains $\pi_1$ (in other words, the content of $\pi_2$ explains why the content of $\pi_1$ is true).

(3) $\pi_0 : \mathit{Explanation}(\pi_1, \pi_2)$
$\pi_1 : \mathit{fall}(e_1, m) \wedge e_1 < \mathit{now}$
$\pi_2 : \mathit{push}(e_2, j, x) \wedge x = m \wedge e_2 < \mathit{now}$

The linguistic grammar doesn't identify the antecedent $m$ to the pronoun $x$. Rather, "him" introduces an *underspecified* equality condition between the newly introduced referent $x$ and some antecedent—written $x =?$. Generally, (disambiguated) linguistic form yields an Underspecified Logical Form (ULF), because syntax on its own does not fully resolve all semantic and anaphoric ambiguities. Similarly, the grammar does not introduce the Explanation relation between the segments. Rather, identifying this coherence relation and the antecedent $m$ to $x$ (thereby replacing $x =?$ with $x = m$ in the logical form of the discourse) is achieved via commonsense reasoning, using the ULFs of the clauses as premises. Moreover, the assumption that $\pi_2$ is coherently related to $\pi_1$ is what makes $m$ an available antecedent for $x$.

Following Lascarides and Stone (2009a), we assume that gestures are elementary discourse units (that is, segments at the leaves of the hierarchical discourse structure); so interpreting gesture involves inferring coherence relation(s) between it and other speech units and gesture units. Furthermore, Lascarides and Stone (2009a) stipulate that co-speech gesture *must* be coherently related to its synchronous speech, and it *can* be related to other units as well. The main aim of this paper is to model this *necessary* connection between co-speech gesture and its synchronous speech. In line with theories of dynamic semantics and discourse interpretation (Hobbs 1985; Kehler 2002; Asher and Lascarides 2003), we further assume that there are constraints on which antecedents are available for resolving the anaphoric elements of the current discourse unit. In speech-only discourse, antecedents to anaphora in the discourse unit $\pi$ must be introduced in $\pi$ itself or in a unit $\pi'$ that $\pi$ is coherently related to. Following Lascarides and Stone (2009a), we carry over these constraints to gesture: i.e., *all* individuals that are a part of the pragmatic interpretation of a gesture behave like anaphoric expressions—they must bind via a bridging relation to

an available antecedent (Asher and Lascarides 1998). Thus inferring a pragmatic interpretation of gesture is dependent on inferring how it coherently connects to available speech unit(s).

The meaning representations that we derive from the form of a sentence with co-speech gesture must respect the above constraints on interpretation. To achieve this, we make the choices of speech and gesture integration—which we formally express by attachments in the syntax tree—determine the speech phrase that the gesture is coherently related to. This in turn affects which referents, introduced in speech, are available antecedents for resolving the underspecified gesture meaning (given just its form).

Lascarides and Stone (2009a) observe additional constraints on antecedents for resolving gesture interpretation; constraints that we assume here. Specifically, they claim that the antecedent for resolving gesture can be introduced by a gesture or a linguistic discourse unit, but antecedents for resolving linguistic anaphora cannot be introduced by depicting gestures. This doesn't apply to deixis: a linguistic anaphor can co-refer with a referent that's pointed at. For instance, when a person points at a knife and says "It's sharp", it is perfectly acceptable for "it" to refer to the knife introduced by the deictic gesture. In contrast, when a person says "He cut the cake" and makes a 'cutting' gesture with a vertically flat palm to depict the instrument used for cutting, it is rather unnatural to continue this discourse with "It was sharp" where "it" refers to the knife introduced by the iconic gesture.

By drawing on standard methods from formal linguistics, our goal is to make the analysis of a discourse featuring co-speech gestures compatible with the analysis of purely linguistic discourse. Given the fact that we are adopting a coherence-based theory, the pragmatic interpretation of co-speech gesture is dependent on the content of the linguistic signal it is coherently related to. With this in mind, we introduce the notion of *speech-gesture alignment* to roughly designate: (i) that speech and gesture are coherently related; and (ii) that resolving the (underspecified) semantics of gesture to a specific interpretation and inferring a coherence relation are logically co-dependent tasks. We shall refine the notion of alignment in Section 3.3 after a discussion of how linguistic form and gestural form, including their relative timings, constrain the alignment configurations. In the next section, we illustrate the various ways in which a gesture can be interpreted

in context.

### 2.2 *Ambiguous Form-Meaning Mapping*

Syntactic attachment ambiguities and semantic scope ambiguities are ubiquitous in grammars. For instance there is the non-unique choice for attaching the PP in "John saw the man with the telescope". And there's the non-unique semantic scope of the quantifier in "every dog probably did not walk"—"probably" semantically outsopes the negation, which outscopes "walk", but the quantifier "every man" may outscope "probably", or have narrow scope to "probably" but outscope the negation, or have narrow scope to the negation. Most grammars have to handle semantic scope ambiguity in the absence of syntactic ambiguity.[3] So syntax derives a ULF that underspecifies semantic scope.

We will now argue that the range of plausible pragmatic interpretations of co-speech gesture can likewise be analysed via a non-unique choice of attachment of the co-speech gesture to speech and a non-unique way of resolving scope in the ULF that gets composed via such attachments. In essence, these sources of ambiguity familiar from linguistics can also capture ambiguities in co-speech gestures. In Section 3.1, we will then argue that not only *can* one model co-speech gesture ambiguity this way, but one *should*.

We use a slight modification of example (1), namely (4), to discuss the ambiguous form-meaning mapping of depicting gestures. Its plausible pragmatic interpretations are presented in SDRT notation, except that we ignore tense and presupposition, and (following the English Resource Grammar (ERG, Flickinger 2000)), events are not existentially bound.

(4) John mixes mud
*Same gesture as in* (1).

Intuitively, one of the possible denotations of the circular hand movement is paraphrasable as "the mud is going round in horizontal circles". This interpretation is regimented in the LF in (5), which features an Elaboration relation between the speech content $mud(x)$

[3] For instance, CCG (Steedman 2000) and Montague Grammar (Montague 1988).

(labelled $\pi_s$) and the gesture content labelled $\pi_g$—a horizontal rotating event $e'$ over a substance $x'$ that is made equal to the 'mud' referent $x$ introduced in $\pi_s$. So the speech-gesture action conveys "John mixes mud, (specifically) the mud that is going round". Like (2), this LF consists of a hierarchical structure of coherently related segments.

(5) $\pi_s : mud(x)$
$\pi_g : \exists x'(substance(x') \wedge rotate(e', x') \wedge horizontal\_motion(e'', e')$
$\wedge\, x = x')$
$\pi_0 : \exists x(john(j) \wedge mix(e, j, x) \wedge Elaboration(\pi_s, \pi_g))$

The constraints on anaphoric reference imposed by the discourse structure in (5) license using $x$ as an antecedent for specifying the content of $\pi_g$ (Asher and Lascarides 2003; Lascarides and Stone 2009b): $x$ is available because it's 'introduced' by the predication $mud(x)$—or more precisely, using HPSG terminology, $x$ is the semantic index of $mud(x)$ (its first argument which introduces a noun variable)—and $mud(x)$ is a part of $\pi_s$, to which $\pi_g$ is coherently related.

Further, this LF represents one way of resolving the underspecified semantic scope of the ULF that you would get by attaching the gesture to the NP "mud" in the syntax tree. Specifically, following the standard approach to semantic composition (Sag and Wasow 1999; Copestake *et al.* 2001), assume the semantic component of the construction rule that attaches gesture to a linguistic unit introduces an (underspecified) *coherence relation*—here resolved to Elaboration—between the gesture and the predications in that linguistic unit, but the ULF so derived underspecifies the relative scope of this (underspecified) coherence relation and the quantifiers in the linguistic unit. Then the ULF derived by attaching the gesture to the NP "mud" would force the coherence relation to outscope the predicate $mud(x)$ but it won't outscope the predicates $mixes(e, j, x)$ or $john(j)$. (5) is a fully specific logical form that is licensed by this ULF. Here, $\exists x$ *must* outscope the coherence relation because free occurrences of $x$ are forbidden (Copestake *et al.* 2005).

An alternative pragmatic interpretation of the co-speech gesture in (4) is that it depicts the event of mud going round as a *result* of the mixing. A formal rendition of this interpretation is given in (6).

(6) $\pi_s : \exists x(mud(x) \wedge mix(e, j, x))$

$\pi_g : \exists x'(substance(x') \wedge rotate(e', x') \wedge$
$\quad horizontal\_motion(e'', e') \wedge x = x' \wedge cause(e, e'))$
$\pi_0 : john(j) \wedge Result(\pi_g, \pi_s)$

Unlike (5), the gesture qualifies the event $e$ of mixing—$e$ is available because it's the semantic index of $mix(e, j, x)$, which is a part of $\pi_s$. Here, the speech content $\pi_s$ and the gesture content $\pi_g$ are coherently related via Result (rather than Elaboration): a rough linguistic paraphrase would be "By making it go round, John was mixing mud". I.e., the gesture here functions roughly like a free adjunct.

This interpretation can be derived by attaching the gesture to a linguistic unit whose timing is (again) not *equal* to the timing of the gesture (though they temporally overlap), and then resolving the ULF that results from this attachment to a fully specific logical form. Here, (6) can be derived from the ULF you get by attaching the gesture to the VP "mixes mud": this attachment forces $\pi_s$ to include the predication $mix(e, j, x)$. Consequently, the quantifier $\exists x$ can now have narrower scope than the coherence relation, as shown. This contrasts with attachment to the NP "mud": this attachment ruled out $mix(e, j, x)$, and hence also $\exists x$, from being within the scope of the coherence relation. Further, since the predication $john(j)$ in (6) isn't a part of $\pi_s$, $j$ is not available for resolving the content of $\pi_g$.

The particular linguistic grammar that we use in this paper to analyse co-speech gesture—specifically the ERG (Flickinger 2000)—makes the ULF generated by VP attachment the same as that derived by S attachment. For example, the adverbial in *Probably John mixed mud* and *John probably mixed mud* attaches to the S and VP nodes respectively, but in both cases the ULF forces the modal introduced by *probably* to outscope $mixes(e, j, x)$ and it *underspecifies* whether it also outscopes $john(j)$ and/or $mud(x)$, or not. Thus (6) is also derivable from the ULF you get by attaching the gesture to the S node. An alternative fully scoped form of this ULF corresponds to a further plausible interpretation of the gesture:

(7) $\pi_s : \exists x(john(j) \wedge mud(x) \wedge mix(e, j, x))$
$\pi_g : \exists x'(agent(j') \wedge substance(x') \wedge rotate(e', j', x') \wedge$
$\quad horizontal\_motion(e'', e') \wedge x = x' \wedge e = e' \wedge j = j')$
$\pi_0 : Depiction(\pi_s, \pi_g)$

Unlike (5) and (6), $john(j)$ is now outscoped by the coherence relation; so $j$ is available for resolving the content of $\pi_g$. As before, the choice of antecedents for specifying the content of $\pi_g$ interacts with the choice of coherence relation: here, the coherence relation is Depiction and the overall content is roughly paraphrasable as another free adjunct: "As he was making it go round, John was mixing mud".

The interpretations in (5), (6) and (7) all feature identity between a referent introduced by the co-speech gesture and a referent introduced by speech. However in (8) the gesture does not denote a salient property of the referents introduced in speech: instead, it qualifies the speech act of questioning (signalled by a rising intonation). A rough paraphrase of the meaning of the multimodal action in (8) would be "Are you telling me that John mixes mud?". Interpreting the gesture in this metaphorical way (see the LF in (9)), and inferring a Metatalk relation (Polanyi 1985) whose semantics is defined in terms of the *speech act* rather than the domain-level content, would be supported via an attachment of the co-speech gesture to the S node.

(8) John mixes mud?
*Speaker's right hand is vertically open with palm facing up. The speaker moves it forward to the frontal space.*

(9) $\pi_s : question(\exists x(john(j) \wedge mud(x) \wedge mix(e,j,x)))$
$\pi_g : question(tell(e',you,p) \wedge p = \pi_s)$
$\pi_0 : Metatalk(\pi_s, \pi_g)$

While the attachments we've proposed deviate from McNeill's (1992) claim that co-speech gesture is semantically related to its *temporally simultaneous* speech phrase, we remain agnostic about his claims (and those of others) about the underlying production processes—e.g., McNeill's claim that decisions about which contents are expressed in which channel stem from a single (complex) thought.

## 3 SPEECH–GESTURE ALIGNMENT AS SHOWN IN DATA

This section introduces examples of speech-gesture actions that illustrate that despite their ambiguities, speech-gesture alignment is jointly constrained by prosody, linguistic syntax and relative timing of speech

and co-speech gesture. This serves as qualitative evidence for: (a) encoding the constraints on speech-gesture alignment within a grammar (rather than entirely via pragmatics); and in particular (b) suitably constraining the application of construction rules of the kind we described in the prior section. The examples we use as evidence include both constructed examples (to illustrate our judgements about ill-formedness) and examples extracted from existing corpora.

### 3.1 *Speech-Gesture Alignment and Prosody*

We begin with the constructed example (10), which reflects intuitions of native speakers about multimodal grammaticality.

(10) * Your [$_N$mother] called.
*The speaker puts his hand to the ear to imitate holding a receiver.*

Intuitively, it seems anomalous to perform the gesture along the unaccented "called", even though the gesturing hand is shaped as holding a receiver and can thus be associated with calling. This anomaly would not arise if the gesture was performed along the whole utterance (or a part of it) which, importantly, includes the prosodically prominent element "mother": e.g., "mother called" or "your mother called". As suggested by Mark Steedman (personal communication), gestures exhibit contrastive properties in analogy to those conveyed by pitch accents. If this is so, then it's not surprising if a co-speech gesture is well-formed only if, unlike (10), it temporally overlaps with a contrastive component that's signalled via prosodic prominence (this is not to say that gesture performance is *driven* by prosody, but rather that their performances are mutually constraining). Further, a pragmatic interpretation where the gesture depicts calling must be sourced in a syntactic derivation where the gesture is aligned with a linguistic unit that includes "called"—prosody constrains the gesture to be aligned with a phrase that includes "mother", but the event of calling is available to its interpretation only if it aligns with a phrase that includes "called" as well. Thus, just like with purely linguistic discourse, considerations about plausible pragmatic interpretations can serve to resolve syntactic ambiguities that are licensed by the construction rules in the grammar. Further, this strong relationship in (10) between the performance of the gesture and prosody is in line with the empirical findings of Giorgolo and Verstraten (2008), who

Figure 2: Gesture Depicting “greasy”, example (11) (Kendon 2004)

isolated prosody as the parameter that influences the perception of multimodal well-formedness vs. multimodal ill-formedness.

Considering that form (here, prosody) constrains what part of the speech signal a co-speech gesture can align with, we define alignment as a constraint on grammaticality. Ungrammatical (and hence misaligned) speech and co-speech gestures comprise cases where the timing of co-speech gesture relative to the timing of speech does not validate *any* construction rule in the grammar by which speech and gesture may be combined; and our aim is to ensure that such constraints on the construction rules match native speakers’ judgements about ill-formedness.

### 3.2 *Speech-Gesture Alignment and Syntax*

To illustrate that linguistic syntax influences decisions about which phrase a co-speech gesture semantically aligns with, consider utterance (11), where the speaker is discussing new owners of a factory finding it filthy. Along with “greasy…”, the speaker’s hands spread out to the left and right periphery (Figure 2) so as to designate some spatial extent, some closed area being made greasy (Kendon 2004).

(11) First of all they made [pause 0.1 sec] everything
[$_N$* gre]asy in the whole room place.

Consider how moving the timing of this gesture affects its meaning. If the gesture onset was moved a few milliseconds earlier so that it happened along “made everything greasy” or if it was held further so as to span “made everything greasy in the whole room”, this would

not change the interpretation of it: it still designates an enclosed area that's greasy. This interpretation would also remain unchanged if the primary pitch accent were on "everything" rather than "greasy", and the gesture temporally coincided with "everything". However, the gesture cannot receive this interpretation if it temporally coincides only with the subject NP "they" (which in turn would need to be accented for the speech-gesture action to be well-formed): now it designates a spatial referent for "they" in the gestural space, and cannot qualify the spatial extent of greasiness. These variations suggest that a gesture that temporally coincides with "they" can only semantically align with "they", but a gesture temporally coinciding with any element in a VP can semantically align with the VP, sub-portions of the VP containing the temporally coinciding words, and with the whole clause.

A special class of deictic gestures behave differently with regards to the semantic effects of prosody and timing, however. In (12) from the annotated AMI corpus (Carletta 2007), the deictic gesture is performed along with the prominent "Thank you" but its denotation binds to that of the NP "the mouse". The alternative interpretation where the gesture signal and the speech signal are bound through a causal relationship—i.e., handing the mouse is the reason for thanking the addressee—is not possible, since it's clear in context that "Thank you" is related to what came in the *previous* discourse (i.e., projecting the presentation in slide show mode in response to the speaker's request).

(12) [$_N$Thank] you. [$_{NN}$I'll] take the [$_N$mouse]
*Speaker's right hand is loosely open, index finger is loosely extended, pointing at the computer mouse*

In (13) (again from the AMI corpus), the deixis happens along the nuclear accent "said", but it identifies the individual that resolves the pronoun "she" coming from speech.

(13) And a as she [$_N$said], it's an environmentally friendly uh material
*The speaker extends her arm with a loosely open palm towards the participant seated diagonally from the speaker.*

In these examples, the gesture would fail to map to the intended meaning if the grammar were to license attaching a co-speech gesture only to its temporally simultaneous linguistic phrase.

Based on Lascarides and Stone (2009a), we formalise the location of the pointing hand with the constant $\vec{c}$; this marks the physical location of the tip of the index finger. This combines with the features of the pointing hand—the hand shape, the orientation of the palm and fingers, and the hand movement—to determine the spatial region $\vec{p}$ that's designated by the gesture—e.g., a stroke with an extended index finger will make $\vec{p}$ a line (or a cone) that starts at $\vec{c}$ and continues in the direction of the index finger. Abstract deixis identifies referents that are not physically salient in the communicative situation. To account for this inequality between the gestured space and actual denotation, Lascarides and Stone (2009a) use the function $\nu$ to map the physical space $\vec{p}$ designated by the gesture to the space $\nu(\vec{p})$ it denotes (and they claim that the value of $\nu$ is pragmatically determined). Essentially, $\vec{p}$ is not equal to $\nu(\vec{p})$ in cases where the referent introduced in the gesture space is not physically present. Conversely, $\vec{p}$ equals $\nu(\vec{p})$ when the referent introduced by the gesture is at the physical coordinates identified in the gesture space.

With this in mind, we observed in all the annotated corpora we examined[4] that the temporal/prosodic mismatch occurred only in cases where the visible space $\vec{p}$ designated by the gesture was *equal* to the space $\nu(\vec{p})$ it denoted, i.e., the function $\nu$ that maps the space identified by gesture to the actually denoted space resolves to equality. So we shall capture this finding in the grammar via a construction rule that allows gesture to align with a spoken word that is not prosodically marked and/or that doesn't temporally overlap with the gesture, but only if the deictic referent is physically located at the exact coordinates identified by the pointing hand.

Bearing in mind that we are restricting our study and analysis to only those gestures that temporally overlap with speech (i.e., co-speech gestures), these examples provide evidence that their semantic alignment depends on the syntax and prosody of the speech signal, as well as the relative timing of the gesture and speech. This motivates encoding the constraints on alignment *within a grammar*, for

[4] To study depicting gestures, we used a 165-second collection of four recorded meetings, annotated for gesture events and intonation events in the ToBI framework (Loehr 2004). To study deictic gestures, we used two multimodal corpora: a 5.53 min recording from the Talkbank Data,[5] and observation IS1008c, speaker C from the AMI corpus (Carletta 2006).[6]

this is where information about syntactic constituency is expressed. The alternative approach would be to infer speech-gesture alignment at the pragmatic level, via the commonsense reasoning that resides there for inferring which discourse units are coherently connected to which other units. But this alternative is incompatible with existing and well-established assumptions about the interface between syntax, semantics and pragmatics. For instance, our discussion of example (11) showed that the temporal relationship between subject NP/VP boundary and the gesture profoundly affect the possible interpretations. To capture this fact, pragmatics would need access to the *syntax* of the speech. However, there is no formal model of pragmatics that supports that kind of architecture, without pragmatics being fully integrated into the grammar itself along the lines of Dynamic Syntax (Kempson *et al.* 2000). In contrast to the non-modular approach of Dynamic Syntax, we aim to maintain a conservative, well-established and modularised interface between syntax, semantics and pragmatics, so that implementations of our grammar can be supported by standard methods for computing discourse meanings (e.g., statistical discourse parsers, Afantenos *et al.* 2015).

Accordingly, we will develop a speech-gesture grammar using standard techniques for syntactic derivation and semantic composition, where the constraints on attaching co-speech gesture to a linguistic constituent are defined in terms of relative timing, prosody and linguistic syntax.

The examples we've discussed so far motivate allowing attachments of gesture to linguistic constituents whose timing is *not* identical to the timing of the gesture; we saw in Section 2.2 that making alignment equivalent to temporal simultaneity would under-generate the range of plausible pragmatic interpretations. Rather, the choices of attachment, and hence ultimately the choices of what the gesture means, are determined by the prosodic properties and constituent boundaries of the speech signal as well as relative timing.

### 3.3 *Speech-Gesture Alignment*

Given our assumptions about constrained inference in pragmatics, and also given our observations of how form affects the speech-gesture interaction, we now refine the notion of alignment as follows:

**Definition 3.1. Speech-Gesture Alignment.** *Our choice of which speech phrase a gesture (stroke) can align with is guided by the following factors:*

*i. the final interpretation of the gesture in specific context of use;*

*ii. the speech phrase whose content is semantically related to that of the gesture given the value of (i); and*

*iii. the syntactic structure that, with standard semantic composition rules, would yield a* ULF *supporting (i) and hence also (ii).*

The derivation of the single speech-gesture syntactic structure, which is constrained by the prosody of the temporally overlapping speech signal, is achieved within the grammar. This definition encompasses both form (introduced in clause (iii)) and meaning (all three clauses). We capture semantic alignment of speech and gesture via attachment in a single syntax derivation tree, because—as shown—syntax (among other things) governs semantic alignment. If there is a choice as to which phrase a co-speech gesture can align to, then this is modelled via a combination of structural—i.e., attachment—ambiguity and semantic scope ambiguity that's licensed by the ULF so-derived. The semantic effects of alignment are thus captured using standard methods of semantic composition on the derivation tree. Given the theory of pragmatics we aim to support, the construction rules combining speech and a depicting gesture introduce an (underspecified) semantic relation $vis\_rel(s, g)$ (visualising relation) between the content $g$ of the depicting gesture and the content $s$ of the speech constituent to which the gesture attaches, which captures the fact that speech and gesture are coherently connected (Lascarides and Stone 2009a). The (underspecified) relation that's introduced by the construction rules that combine deixis and speech is $deictic\_rel(s, g)$ (Lascarides and Stone 2009a). The resolution of these underspecified relations to a pragmatically preferred and specific value happens externally to the grammar at the semantics/pragmatics interface.[7] In

[7] Resolving the underspecified relations is a matter of commonsense reasoning which includes the underspecified semantics produced by the grammar, as well as real-world knowledge. A relation such *vis_rel* is a supertype of the more specific Depiction and Result.

Section 6 we discuss the formal framework and in Section 7 the implementation in HPSG.

## 4 PREVIOUS WORK AND CONTRIBUTION

This paper aims to demonstrate that informal observations about the relationship between speech-gesture form and meaning can be regimented formally, using standard techniques from linguistics. In particular, we use standard techniques for deriving logical form from a syntax tree within a grammar, while ensuring that the meaning representations so derived comply with the requirements imposed by existing formal models of pragmatics.

The idea of integrating speech and gesture within a grammar is by no means new, with several such proposals established over the past 20 years (see, *inter aliae*, Johnston 1998a,b, Kühnlein *et al.* 2002, Paggio and Navarretta 2009, Giorgolo and Asudeh 2011). Further, the "constituent structure" of gesture, as well as its syntactic function for the integration within the language, has also been a matter of research (see Fricke 2008, Müller and Ladewig 2013). And the construction of meaning across speech and gesture has been the subject of analysis within construction grammars (Steen 2013).

But there are a few main differences between this prior work and our approach. First, we claim that the speech phrase that gesture aligns with is not determined uniquely by when the gesture was performed. Whilst the TIME feature matters, we also constrain alignment via prosody and syntactic notions such as headedness. Further, in contrast to these prior grammars, we aim for a *domain independent* analysis, and so we must fully capture all linguistically licensed semantic alignments between speech and co-speech gesture, rather than only those that are plausible in the chosen domain of application. The other main difference lies in the semantic component of the grammar. In particular, we draw on recent advances in deriving an Underspecified Logical Formula (ULF), which allows the grammar developer to capture semantic ambiguity in the absence of syntactic ambiguity. The above grammatical approaches all assume that every semantic ambiguity corresponds to a syntactic ambiguity.

There are previous semantic analyses of gesture (Lücking *et al.* 2006b; Lascarides and Stone 2009a) that assume a grammar pro-

duces an underspecified meaning representation: these theories focus on how contextual information contributes to mapping the underspecified meaning that's derived from form into a fully specific and pragmatically preferred interpretation. Our work contributes to this by providing a grammar framework that produces the form-meaning mappings they assume. In doing so, we not only capture informal observations about gestural ambiguity, but our formal model uses well-established methods from linguistics to produce a meaning representation that is compliant with current models for multimodal processing at the semantics/pragmatics interface.

To achieve that, we perform two dependent tasks: first, we extract generalisations from the existing literature and from our own observations in annotated multimodal corpora about the syntactic and semantic well-formedness of speech-gesture signals; second, we use the extracted generalisations to define a precise grammar that models the form of the speech, the form of the gesture and the form of their combination, producing ULFs of speech and gesture using standard methods of syntactic derivation and semantic composition from linguistics. We also demonstrate that the grammar can be implemented by extending an existing linguistic grammar.

## 5 MAPPING GESTURE FORM TO MEANING

### 5.1 *Modelling Gesture Form*

One major difference between speech and gesture is how the meaning gets derived from the form of the signal. Gestures are 'global' and 'synthetic' (McNeill 1992), i.e., the meanings of the various features of a gesture's form—such as the direction of the movement, the hand shape, the location of the hands, etc—determine the meaning of the gesture as a whole. This is unlike the semantic compositionality via natural language syntax. Following previous work (Kopp *et al.* 2004, Lascarides and Stone 2006, Hahn and Rieser 2010 inter alia), we regiment this difference by using Typed Feature Structures (TFS) since they support a *non-hierarchical* representation of the distinct aspects of the gesture's form. The gesture type designates its category: e.g., *depict-literal* for literally depicting gestures (Figure 3) and *deictic-abstract* for abstract deixis (Figure 4), of the kind exhibited in (14):

| *depict-literal* | |
|---|---|
| HAND-SHAPE | bent |
| PALM-ORIENT | towards-down |
| FINGER-ORIENT | towards-down |
| HAND-LOCATION | lower-periphery |
| HAND-MOVEMENT | circular |

Figure 3: TFS representation of the form of the depicting gesture in (1)

| *deictic-abstract* | |
|---|---|
| HAND-SHAPE | flat |
| PALM-ORIENT | towards-centre |
| FINGER-ORIENT | away-body |
| HAND-MOVEMENT | down |
| HAND-LOCATION | $\vec{c}$ |

Figure 4: TFS representation of the form of the deictic gesture in (14)

(14) I [$_{PN}$enter] my [$_{N}$apartment]
*Speaker's hands are in centre, palms are open vertically, finger tips point upward; along with "enter" they move briskly downwards, after the downward move, the palms are still vertically open but this time the finger tips point forward.*

The feature-value pairs of a depicting gesture capture every aspect of the form of the hand that (potentially) contributes to its meaning: the hand shape, the orientation of the palm and fingers, the location of the hand relative to the speaker's torso and the hand movement. With deictic gestures, the shape of the hand determines the region of space that is identified by the pointing hand: e.g., an extended index finger identifies a line or a cone that starts from the tip of the index finger; with a vertical open hand, the designated region is a plane. Recording the form of the pointing hand is essential, because prior work shows that it is significant for interpreting its meaning in context (Kendon 2004): e.g., an extended index finger typically singles out an individuated object while a vertical open hand typically denotes a *class* of objects rather than an individuated object, or it serves a pragmatic function such as offering the floor or citing someone else's contribution to the discourse. The hand location of a deictic gesture is represented via the constant $\vec{c}$. This, combined with the deixis form features, determines the region $\vec{p}$ actually marked by the gesture.

## 5.2 *Modelling Meaning*

As we've already highlighted, a well-established method for handling cases where form does not fully determine meaning is semantic underspecification. All frameworks for semantic underspecification—e.g., Quasi-Logical Form (Alshawi 1992), Underspecified Discourse Representation Theory (Reyle 1993), the Constraint Language for Lambda Structures (Egg *et al.* 2001), Hole Semantics (Bos 2004), Minimal Recursion Semantics (Copestake *et al.* 2005), Regular Tree Grammars (Koller *et al.* 2008)—construct from a fully disambiguated form an abstract representation of meaning that can resolve to several distinct specific messages in context, rather than deriving those specific representations from syntax directly, and assuming a syntactic ambiguity for every semantic ambiguity. Technically, the ULF derived by syntax *partially describes* the form of a fully specific logical form, which in turn represents a context-specific interpretation which can be evaluated against a model or the actual situation at hand.

To map the form of the gesture to an underspecified meaning representation, we use the underspecification formalism of Robust Minimal Recursion Semantics (RMRS, Copestake 2007)—a factorised version of ERG's semantic framework, Minimal Recursion Semantics (MRS, Copestake *et al.* 2005). RMRS was originally developed to support the integration of deep and shallow processing. Modelling gesture is somewhat akin to shallow processing in that one has to handle the large degree of underspecificity.

To illustrate it, consider the MRS for "every dog chased some cat" in (15). Here, the semantic scope ambiguities are captured by the so called *qeq* ($=_q$) contraints which allow for two alternative fully scoped formulas.

(15) $l_1 : every(x_0, h_3, h_1)$
$l_{11} : dog(x_1)$
$l_2 : some(y_0, h_4, h_2)$
$l_{21} : cat(y_1)$
$l_3 : chase(e_1, x_2, y_3)$
$h_3 =_q l_{11}, h_4 =_q l_{21}$

While MRS underspecifies scope, it still requires a fully specified predicate-argument structure. However, neither shallow language

processors nor gestural form on their own can fully determine a unique predicate argument structure. Refining MRS to RMRS solves this. One simply produces a highly factorised representation of each elementary predication: each one is equipped with its own unique *anchor (a)*, which serves as a locus for specifying the predicate's arguments; equations (e.g., $x_0 = x_1 = x_2$) are also added to express unifiability between variables. So (16) is a notational variant of (15).

(16) $l_1 : a_1 : every(x_0), l_1 : a_1 : RSTR(h_3), l_1 : a_1 : BODY(h_1)$
$l_{11} : a_{11} : dog(x_1)$
$l_2 : a_2 : some(y_0), l_2 : a_2 : RSTR(h_4), l_2 : a_2 : BODY(h_2)$
$l_{21} : a_{21} : cat(y_1)$
$l_3 : a_3 : chase(e_1), l_3 : a_3 : ARG1(x_2), l_3 : a_3 : ARG2(y_3)$
$h_3 =_q l_{11}, h_4 =_q l_{21}$
$x_0 = x_1 = x_2$
$y_0 = y_1 = y_3$

For instance, a POS tagger would yield (17) instead of the more specific (16). (17) captures the semantic insight that, for example, knowing that the word *chase* is tagged as a verb, one knows that its semantic index is an event, but one does not know how many arguments the predicate symbol introduced by *chase* takes because the POS tagger lacks information about lexical subcategorisation.

(17) $l_1 : a_1 : every(x_0)$
$l_{11} : a_{11} : dog(x_1)$
$l_2 : a_2 : some(y_0)$
$l_{21} : a_{21} : cat(y_1)$
$l_3 : a_3 : chase(e_1)$

Semantic composition with RMRS follows the semantic algebra of Copestake *et al.* (2001): the predications and *qeq* on the mother are acummulated from those in the daughters and the semantic head daughter has its 'hook' (roughly equivalent to a $\lambda$-term) replaced by the semantic index of the non-head.

## 5.3 *Form-Meaning Mapping*

### 5.3.1 Depicting Gestures

Following Lascarides and Stone (2009a), mapping the form of a depicting gesture to its meaning involves mapping each feature value

pair in the TFS representing its form to an RMRS-based underspecified predication: the ULF of the gesture from Figure 3 is shown in (18).

(18) $l_0 : a_0 : [\mathscr{G}](h)$
$l_1 : a_1 : hand\_shape\_bent(i_1)$
$l_2 : a_2 : palm\_orient\_towards\_down(i_2)$
$l_3 : a_3 : finger\_orient\_towards\_down(i_3)$
$l_4 : a_4 : hand\_location\_lower\_periphery(i_4)$
$l_5 : a_5 : hand\_movement\_circular(i_5)$
$h =_q l_n$ *where* $1 \leq n \leq 5$

Each predicate has a label, an anchor, and a semantic index, as is standard in RMRS. Since a predication mapped from depicting gesture could resolve in context to an event $e$ or an individual $x$, its semantic index is a metavariable $i$ that generalises over $e$ or $x$. The predicate symbols underspecify the particular constructor and its arity in the LF. For instance, a feature-value pair like $\left[\text{HAND-MOVEMENT circular}\right]$ would map to $l_1 : a_1 : hand\_movement\_circular(i)$. Resolving these predicates happens outside the grammar as a byproduct of discourse processing (Lascarides and Stone 2009a). In particular, each underspecified predicate (such as $hand\_movement\_circular(i)$) is a root to a type hierarchy of increasingly specific predications of content. This is roughly analogous to constructing a specific lexical meaning out of a polysemous lexical entry (Copestake and Briscoe 1995), but here the type hierarchy captures constraints on interpretation that are imposed by the requirement for iconicity—i.e., a resemblance between the form of the gesture and its meaning. This type hierarchy is designed so that a circular hand movement can never resolve to, say, a rectangular concept. To illustrate the idea, in Section 2.2 we claimed that one of the interpretations of the circular hand movement in (1) was the mud being mixed. This is achieved by resolving $hand\_movement\_circular(i)$ to a conjunction of predications: $substance(x') \wedge rotate(e', x')$, which is a node in the type hierarchy that's rooted at $hand\_movement\_circular(i)$, and is featured in (5). In an alternative interpretation this hand movement is a depiction of the mixing event from the agent's viewpoint: i.e., the underspecified predicate $hand\_movement\_circular(i)$ can resolve to the three-place predicate $rotate(e', j', x')$, featured in (7).

Further, recall from Section 2.1 the constraint that an individual that is introduced in a depicting gesture can't be an antecedent to a

pronoun in speech. Lascarides and Stone (2009a) regiment this constraint by introducing the scopal operator $[\mathscr{G}]$: all predicates mapped from depicting gesture fall within its scope (via the scopal condition $h =_q l_n$), and the dynamic semantics Lascarides and Stone assign to $[\mathscr{G}]$ ensures that co-reference across the modalities is suitably constrained.

### 5.3.2 Deictic Gestures

The mapping of deixis form to a ULF captures the fact that deixis provides the spatial reference of an individual or event in the physical space $\vec{p}$ (the complete RMRS logical form mapped from the gesture in Figure 4 is shown in (19)). This is formalised by the two-place predicate $l_{21} : a_2 : sp\_ref(i_1)$ $l_{21} : a_2 : ARG1(\nu(\vec{p}))$ whose first argument is the underspecified variable $i_1$, and the second argument ARG1—linked through the anchor $a_2$—is the actually denoted space $\nu(\vec{p})$ with $\nu$ being the function that maps the gesture space to the space in denotation (recall discussion in Section 3.2). The ULF is only a partial description of the resolved LF: e.g., resolving the underspecified referent $i_1$ to an object $x$ and inferring a relation between the deixis denotation and the speech denotation is a matter of pragmatic reasoning. Note how in the prior interpretation of $hand\_movement\_circular(i)$ $i$ resolves to an individual $x$, but here it resolves to an event $e$.

To capture how the form of the pointing hand affects its meaning, we map each deixis feature-value pair to a two-place predicate, with the first argument being an event variable ($e_0 ... e_n$) and the second argument ARG1 being the referent identified by the pointing signal ($i_0 ... i_n$). This formalisation is similar to the treatment of non-scopal modification in the English Resource Grammar (ERG, Flickinger 2000): a deictic predication (as mapped from form) is a two-place predication whose second argument ARG1 is equated with the semantic index of the modified predication, obtained by equating $i_0 = i_1 = i_2 = i_3 = i_4 = i_5 = i_6$ and whose label is equated with the label of the modified predication, obtained via $l_{21} = l_{22} = l_{23} = l_{24} = l_{25} = l_{26}$. For consistency with ERG where individuals are all bound by quantifiers, we use the *deictic_q* quantifier to quantify over the spatial referent $i_1$.

(19) $l_1 : a_1 : deictic\_q(i_0)$ $l_1 : a_1 : RSTR(h_1)$ $l_1 : a_1 : BODY(h_2)$
$l_{21} : a_2 : sp\_ref(i_1)$ $l_{21} : a_2 : ARG1(\nu(\vec{p}))$
$l_{22} : a_3 : hand\_shape\_flat(e_0)$ $l_{22} : a_3 : ARG1(i_2)$

$l_{23} : a_4 : palm\_orient\_towards\_centre(e_1)\ l_{23} : a_4 : ARG1(i_3)$
$l_{24} : a_5 : finger\_orient\_away\_centre(e_2)\ l_{24} : a_5 : ARG1(i_4)$
$l_{25} : a_6 : hand\_movement\_down(e_3)\ l_{25} : a_6 : ARG1(i_5)$
$l_{26} : a_7 : hand\_location\_c(e_4)\ l_{26} : a_7 : ARG1(i_6)$
$h_1 =_q l_{21}$
$l_{21} = l_{22} = l_{23} = l_{24} = l_{25} = l_{26}$
$i_0 = i_1 = i_2 = i_3 = i_4 = i_5 = i_6$

## 6 GRAMMAR RULES FOR SPEECH AND GESTURE

In this section, we propose grammar construction rules that integrate the form of the gesture and the form of the speech signal into a single syntax tree that in turn provides the basis for deriving a ULF of the speech-gesture action. The construction rules license particular speech-gesture alignments, and constraints on their application make predictions about well-formedness, as motivated via the qualitative observations about speech-gesture data in Section 3.

### 6.1 *Prosodic Word and Gesture Alignment*

We begin with the straightforward case where gesture aligns with a single lexical item:

**Construction Rule 6.1. Situated Prosodic Word Constraint.** *A depicting or deictic gesture can attach to a spoken word* w *of a spoken utterance if (a.) there is an overlap between the temporal performance of the gesture stroke and* w*; and (b.)* w *bears a nuclear or a pre-nuclear pitch accent.*

We represent the mulitmodal rules as phrase structure rules equipped with the following information (Figure 5): the speech daughter S-DTR and the gesture daughter G-DTR each introduce a TIME feature, a SYNSEM|CAT feature which captures its syntacic category (note that for gestures, this information includes the form feature-value pairs, discussed in Section 5.1) and a SYNSEM|CONT feature which captures its (underspecified) semantic contribution. The speech daughter also introduces a PHON feature which captures the phonological information. The construction rule introduces a feature OVERLAP whose values are re-entrant with values in the temporal components of the daughters; and also a TIME feature which is the union of the

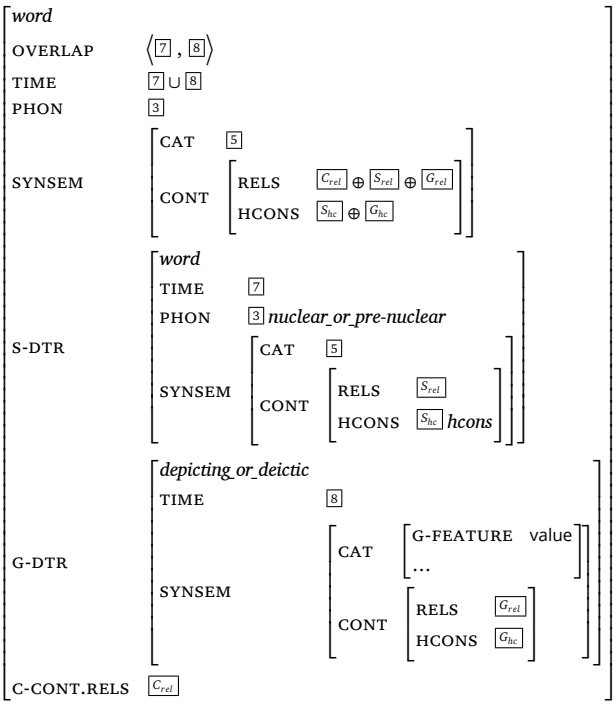


Figure 5: HPSG-based formalisation of the Situated Prosodic Word Constraint aligning gesture and a spoken word

speech daughter's value and the gesture daughter's value. In so doing, we follow previous work where timing is used as a constraint on the integration (Johnston *et al.* 1997). As it is standardly done in ERG, the semantic contribution of the construction rule is captured within C-CONT: here, a depicting gesture introduces an underspecified relation *vis_rel* between the main label of the gesture semantics and the main label of the semantics of the spoken phrase; the underspecified relation introduced by deixis is *deictic_rel* between the semantic index of the speech daughter and the semantic index of the gesture daughter. Multimodal integration happens via unification of these features.

Given the different form-meaning mappings of depicting vs. de-

ictic gestures, we will now provide separate analyses for both gesture types.

### 6.1.1 Situated Prosodic Word Constraint and Depicting Gesture

To illustrate how the Situated Prosodic Word Constraint works with depicting gestures, consider again example (1). The nuclear accent is on the rightmost word "mud", which licenses an attachment of the gesture to it using Construction Rule 6.1. The derivation, which attaches the gesture to "mud", is shown in Figure 6.

The prosodic PHON and syntactic CAT information of the speech head daughter gets propagated to the mother node. We do not propagate the gesture form features to the mother node since we do not need to access gesture form any further. The timing of the situated utterance is recorded in the mother's TIME value. This information is necessary in case the (situated) word aligns with another gesture.

The semantic composition follows the standard English Resource Grammar (ERG) process, namely: the individual semantic formulae are decorated with a global label ($h_1$) which demonstrates the derivation of a single LF. Each formula is also augmented with a hook containing the ltop label (equated to the label of the main predication) and the semantic index. The ltop of the predicate contributed by the speech daughter $l_6 : a_6 : \_mud\_n\_1(x_1)$ is $l_6$ and the index is $x_1$. The ltop of the gesture daughter is equated to the label of the $\mathscr{G}$ modality—$l_0$. Regarding the gesture semantic index, the gesture LF is too underspecified to know which of the semantic predications will resolve to the main variable and hence at this stage we have no information as to which is the semantic index of the formula. We therefore use $i_{1-5}$ as a shorter notation for a disjunction of co-indexations to reflect the fact that the underspecified variable $i_1 \ldots i_5$ of each gesture predicate could potentially resolve to the main variable: event $e$ or individual $x$.

Note that the semantic representation CONT of the situated utterance which features the underspecified relation *vis_rel* between the top label $l_6$ of the speech daughter and the top label $l_0$ of the gesture daughter to designate that the speech and gesture are coherently connected. In RMRS, labels denote the scopal position of an elementary predication. We therefore code the arguments of *vis_rel* as S-LBL and G-LBL to designate that their values are labels of spoken and gestural predications, respectively. As illustrated in Section 2.1, *vis_rel* is

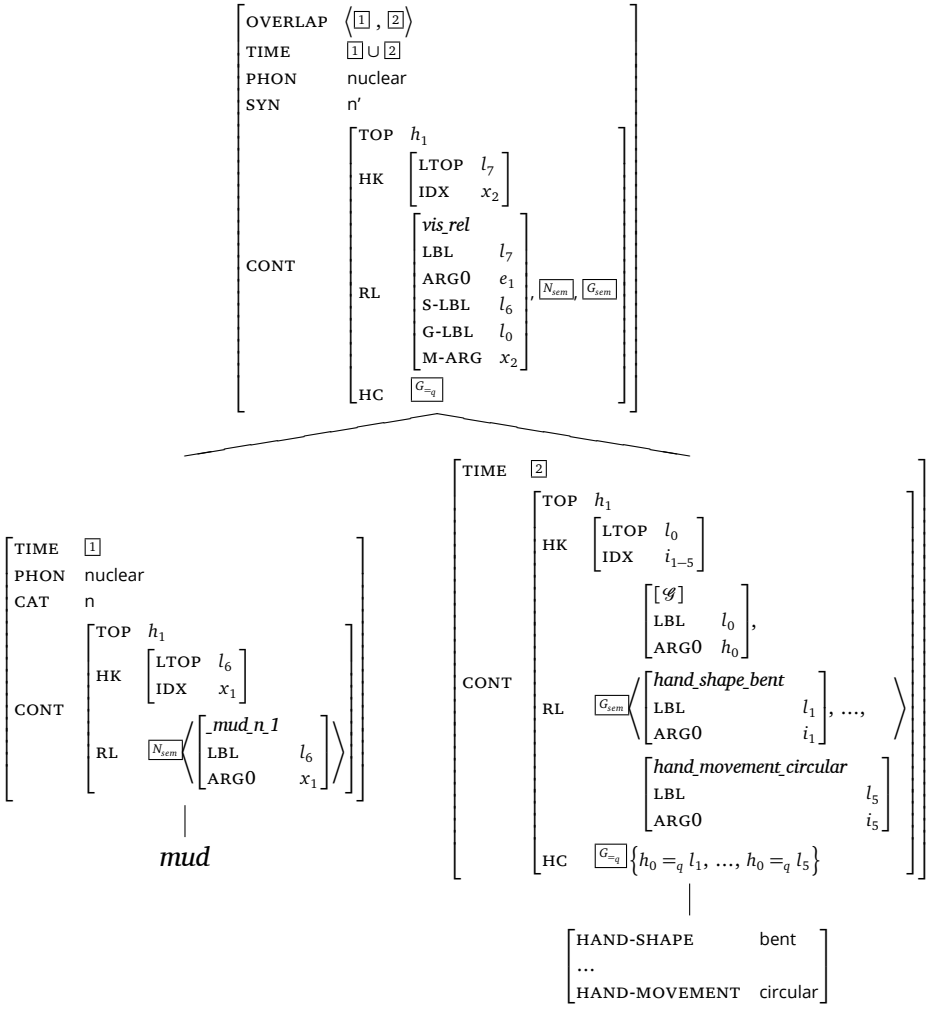


Figure 6: Derivation Tree for Depicting Gesture and the N “mud”

resolvable at the semantics/pragmatics interface to a specific value—e.g., Depiction, Elaboration—that is dependent on resolving the gestural denotation. Here, the attachment to “mud” would support an interpretation where the gesture designates some substance and the

fact that it was going round, which in turn would resolve *vis_rel* to Elaboration, as featured in the LF in (5). The truth conditional contribution of the gesture will thus ultimately be roughly analogous to an appositive or a non-restrictive relative clause modifying the noun. Note that given constraints on reference on the semantics/pragmatics interface, this attachment blocks the gesture referring to anything that is bridging related to "mixes" or "he".

The CONT of the mother is obtained by equating the TOP of the mother to the TOP of the daughters. The relations (abbreviated as RL) of the situated phrase are equal to the append of the predications of the gesture daughter $\boxed{\text{G}_{sem}}$ and the speech daughter $\boxed{\text{N}_{sem}}$, and also *vis_rel*. Further, *vis_rel* introduces a multimodal argument M-ARG which serves as a semantic index of the integrated speech-gesture signal (the hook's index is therefore equated to the index of M-ARG — $x_2$), and so it can be taken as an argument by any external predicate. Here, for instance, the verb "mix" would take two arguments: ARG1—corresponding to the subject—would be identified with ARG0 of "he", and ARG2—corresponding to the object—would be identified with M-ARG of the situated word, consisting of "mud" and the gesture.

### 6.1.2 Situated Prosodic Word Constraint and Deictic Gesture

We illustrate the syntactic derivation and the semantic composition for deixis and a spoken word using utterance (14). The derivation tree is shown in Figure 7. The Situated Prosodic Word Constraint licenses an attachment of the deictic gesture to the verb "enter": it is marked by a pre-nuclear accent, and it temporally overlaps the gesture.

The semantic composition proceeds in the same way as with depicting gestures. Since the gesture semantics features a quantifier *deictic_q*), the local top of gesture is distinct from the label of the quantifier. The semantic index is the underspecified variable $i_1$ bound by *sp_ref*. In composition, the deixis semantic predicates (as shown in 19) append to the semantic predicate $\boxed{\text{V}_{sem}}$ of the speech daughter—$l_4 : a_9 : \_enter\_v\_1(e_5)\ \ l_4 : a_9 : ARG1(u_1)\ \ l_4 : a_9 : ARG2(u_2)$. In so doing, the underspecified semantic index $i_1$ of the deixis unifies with the semantic index $e_5$ of the speech, and so the underspecified gesture variable $i_1$ of $sp\_ref(i_1)$ resolves to an event ($e_7$).

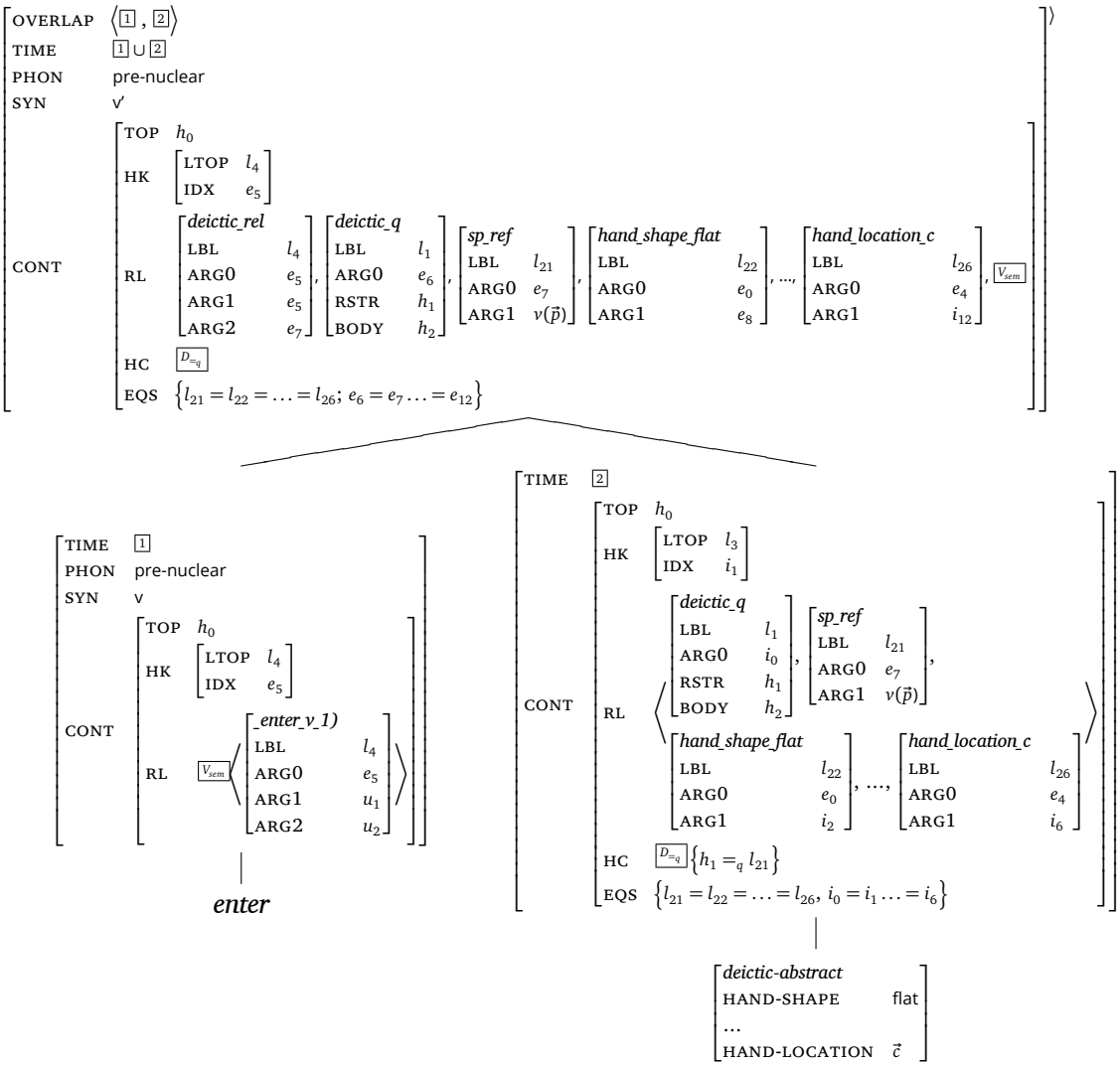


Figure 7: Derivation Tree for Deictic Gesture and the V "enter"

Like depicting gestures, deictic gestures are connected in semantics to their aligned speech via an (underspecified) relation. The construction rule therefore introduces the underspecified relation $deictic\_rel(e_5, e_7)$ between the semantic index $e_5$ of the speech predication and the semantic index $e_7$ of the deictic gesture. Pragmatics must then resolve this relation to a specific value: one possible resolution would be VirtualCounterpart—i.e., the deictic gesture denotes a virtual counterpart of the coordinates of entering the apartment door. Similarly to the treatment of non-scopal modification in language, this relation shares the same label as the speech head daughter since it further restricts the referent introduced by the gesture. Informally, the gesture here functions as an appositive in language and a rough linguistic paraphrase is "the entering event, the event at the coordinates pointed at".

### 6.2 *Speech Phrase and Gesture Alignment*

One of our central claims is that ambiguities as to which speech phrase a co-speech gesture aligns with are best modelled as attachment ambiguities within the grammar. As we demonstrated in Section 2.2, the relative timing of speech and gesture is not the only constraint on using such construction rules; also, temporal constraints should be weaker than *simultaneity*, contrary to McNeill (1992). Rather, we argued that the gesture should temporally overlap with its aligned speech (if it didn't, then by definition it wouldn't be co-speech gesture!) and furthermore temporally overlap with an *accented element* in the (aligned) speech unit. Thus a single utterance such as (1) or (14) can licence different speech-gesture alignments, each of them supporting a distinct range of plausible pragmatic interpretations in accordance with constraints on reference (see Section 2.1). Likewise, it is perfectly acceptable for the gesture in (1) to be performed only while uttering the accented word "mud", and still interpret the gesture in all the ways proposed in Section 2.2. In this section we provide the formal methodology of how to arrive at these interpretations.

As proposed in Section 2.2, we introduce construction rules that allow a gesture to align with an *entire constituent*—that is, a head combined with its arguments—in contrast to Rule 6.1 that aligns gesture with a (temporally overlapping, accented) word. From a descriptive perspective, the inclusion of more context into the speech aligned with

gesture is grounded in the "synthetic" nature of gesture versus the "analytic" nature of the spoken words (McNeill 2005). For instance, in example (1) the information about the direction of the mixing event (i.e., clockwise, downwards), the manner of performing the mixing action (i.e., using the entire hand) is denoted by a single visual performance and by several linearly ordered lexical items ("mixes", "mud"). For the purposes of a multimodal grammar it is essential to distinguish between temporal synchrony and alignment: whereas the former is a quantitative measurement of when the two modalities happen, the latter is a qualitative, linguistic notion pertaining to the syntax tree of speech and gesture and the meaning representation it corresponds to. By setting apart these two notions, we also ensure that the physical termination of the gesture does not enable attachment to a midpoint of a speech constituent.

With all this in mind, we now define the construction rule that allows a gesture to attach to a constituent larger than a single prosodic word:

**Construction Rule 6.2. Situated Spoken Phrase Constraint** *A depicting or deictic gesture can attach to any of the higher projections in the derivation tree of the nuclear/pre-nuclear accent element, which also form a syntactic and/or prosodic constituent* xp*, no matter what the syntactic label is if there is an overlap between the temporal performance of the gesture stroke and* xp.

The attachment of the gesture to any projection in the tree would allow for saturating the head with its selected arguments before the attachment takes place. This means that the attachments are licensed at each saturation step. In this way, we account for the fact that gesture can co-refer to any or all of these arguments in the fully resolved pragmatic interpretation. Note also that Rule 6.2 used 'syntactic and/or prosodic constituent' to refer to any phrase of a hierarchical organisation: prosodic or syntactic. Assuming an analysis where there is no isomorphism between syntax and prosody, this flexibility is necessary whenever there are mismatches between prosodic structure and syntactic structure.[8]

[8] In prior work on HPSG-based analysis of prosody (Klein 2000), prosodic structures are analysed in parallel with syntactic structures.

Since the attachments of depicting gesture to a speech phrase are analogous to the attachments of deixis to the speech phrase, we illustrate the possible attachments using the depicting gesture in utterance (1). Recall from Section 2.2 that the resolved LFs for this speech-gesture action featured coherence relations between: (i) the NP's denotation and the 'rotating' gesture, and (ii) between the VP's (or S's) denotation and the 'rotating' gesture. We discussed (i) in the previous section and we therefore forego any further details about it. Given the construction rule in 6.2, interpretation (ii) is supported as follows: attach the gesture to VP "mixes mud" (or to the S "he mixes mud"). In both cases, the gesture stroke temporally overlaps the nuclear prominent "mud", and so the gesture can attach to its VP projection or S projection. Both of these attachments force the gesture to qualify "mixes" (for the second argument to the underspecified coherence relation that's introduced by the construction rule must outscope $mix(e, y, x)$). They underspecify, however, the relative scope of the coherence relation with respect to the predication $mud(x)$ and $pron(y)$. If these resolve to being within the scope of the coherence relation, then the resolved interpretation of the gesture can co-refer to *he* and to the mud; if not, it can't.

Further to this, we claimed that utterance (10) was ill-formed since the gesture was performed along a non-accented item in an all-rheme utterance. Having introduced the construction rules 6.1 and 6.2, we are now in a position to account for the utterance's ill-formedness: the form of (10) doesn't meet the constraints for either of our construction rules. On the other hand, if the gesture was performed in a way that temporally overlaps the prosodic word "mother", then the rules we've proposed license attachments to the N "mother", the NP "your mother" and even to the S "your mother called".

### 6.3 *Spoken Word and Gesture Alignment: Temporal and Prosodic Relaxation*

The two construction rules we've proposed allow a co-speech gesture to align with a prosodic word or with a constituent that contains prosodic element(s) that overlap the temporal performance of the gesture. These constructions, however, are not sufficient as they do not reflect an important finding from our data. We used examples (12) and (13) to illustrate that when the referent of the deictic gesture is visually

salient, the deictic gestures does *not* necessarily overlap a prosodically prominent word and/or temporally overlap the semantically related word. The following rule takes this into account.

**Construction Rule 6.3. Deictic Prosodic Word with Defeasible Constraint.** *The constraints on temporal overlap in 6.1 and 6.2 are defeasible, i.e., a deictic gesture attaches to a word that is not prosodically prominent and/or whose temporal performance is adjacent to that of the deictic stroke if: (a.) the mapping* v *from gestured space* $\vec{p}$ *to space in denotation* v($\vec{p}$) *resolves to equality; and (b.) the temporal performance of the gesture overlaps (some portion of) the spoken utterance containing the word.*

This temporal/prosodic relaxation rule integrates a defeasible constraint with the view of producing LFs that in context resolve to the intended meaning. As attested by (13),[9] the relaxation of this contraint depends on the salience of co-present individuals and it is thus necessary only in utterances where the gesture denotation is physically present in the visible space, i.e., there is an equality between the physical space that the hand points at and the gesture referent. This rule accounts for the fact that certain characteristics of the context (i.e., salience of the individual pointed at) are required for the rule to apply. Otherwise, the interpretation could be infelicitous. Similar issues occur with deictic expressions and other referential expressions which require a salient individual in context for the utterance to be felicitous (see Lücking *et al.* 2006a).

Note also that this rule constrains the alignment to temporal overlap between (some portion of) the utterance and the gesture. This means that the grammar does not handle gestures performed either before or after the temporal performance of the utterance since anything beyond the clausal level is a matter of relating discourse units. For instance, while the temporal overlap between the gesture and the speech signal in (13) takes care of aligning the gesture and the semantically related element—i.e., "she" in (13)—the gesture in (12) does not overlap any portion of the utterance containing "mouse" and hence the grammar rule cannot attach the gesture to the noun "mouse". Similarly to relating purely linguistic discourse segments, relating the ges-

[9] Many more examples can be found in the AMI corpus.

ture in (12) with the noun "mouse" is a matter of discourse processing that lies outwith the scope of the (syntactic) grammar.

With this constraint in mind, let us examine the possible derivations of utterance (13). The Situated Prosodic Word 6.1 would license attachments to the temporally overlapping prosodically prominent "said". Although syntactically well-formed, this attachment would not produce the contextually preferred (and the most intuitive) interpretation: namely, an identity between the gesture referent and the speech referent. An alternative attachment is provided by Construction Rule 6.3: the deictic gesture may attach to "she" thereby providing an interpretation where the gesture denotation is identical to the denotation of the pronoun "she".

## 7 IMPLEMENTATION AND EVALUATION

The main challenge for the grammar implementation stems from the non-linear input of speech-and-gesture actions. Existing grammar engineering platforms for unification-based grammars typically only parse linearly ordered strings, and so they do not handle multimodal signals whose input comes from separate channels connected through temporal relations. Also, these parsing platforms do not support quantitative comparison operations over the time stamps of the input tokens. This is essential for our grammar since temporal overlap constraints choices of attachment.

To solve this, we pre-processed the XML-based Feature Structure (FS) input so that overlapping TIME values were 'translated' into identical start and end edges of the speech token and the gesture token as follows:

```
<edge source="v0" target="v1">
        <fs type="speech_token">
<edge source="v0" target="v1">
        <fs type="gesture_token">
```

This pre-processing step is sufficient since the only temporal relation required by the grammar is *overlap*, an abstraction over more fined-grained relations between speech (S) and gesture (G) such as (*precedence(start(S), start(G))* ∧ *identity (end(S), end(G))*).

The linking of gesture to its temporally overlapping speech seg-

ment happens prior to parsing via chart-mapping rules (Adolphs *et al.* 2008) which involve re-writing chart items into FSs. The `gesture-unary-rule` (Figure 8) rewrites an input (I) speech token in the context (C) of a gesture token into a combined speech + gesture token where the +GEST and +PROS values of the speech and gesture tokens are copied onto the output (O).

```
gesture-unary-rule := cm_rule &
 [+CONTEXT <gesture_token & [+GEST #gest]>,
  +INPUT <speech_token & [+PROS #pros]>,
  +OUTPUT <speech+gesture_token &
        [+GEST #gest, +PROS #pros]>,
  +POSITION "O1@I1, I1@C1" ].
```

Figure 8: Definition of `gesture-unary-rule`

The +PROS attribute contains prosodic information and the +GEST attribute is a feature-structure representation. The +POSITION constraint restricts the position of the I, O and C items to an overlap (@), i.e., the edge markers of the gesture token should be identical to those of the speech token, and also identical to the speech + gesture token. This chart-mapping rule recognises the gesture token overlapping the speech token and it records this by "augmenting" the speech token with the gesture feature-values.

Gestures overlapping more than one speech token were handled by further chart-mapping rules that distributed the gestural information onto multiple speech tokens within the temporal span of the gesture. So a gesture overlapping, say, three speech tokens, would get split into three gesture tokens. Then, the `gesture-unary-rule` was applied so as to instantiate a speech + gesture token for each speech token temporally overlapping the gesture. The result of this chart-mapping operation is multiple gesture-marked speech tokens whose span is identical to the span of the gesture.

A separate rule was also required for concrete deixis to account for the permitted precedence and sequence relations between the speech token and the concrete deictic gesture token. This rule (which we omit for the sake of space) remains neutral about the positional (and hence temporal) relation between the gesture token and the speech token,

thus allowing a gesture token of type *deictic-concrete* to attach to each speech token from the input chart.

In the grammar, we extended the ERG word and phrase rules with prosodic and gestural information where the +PROS and +GEST features of the input token are identified with the PROS and GEST of the word and/or lexical phrase in the grammar. We then added a gesture lexical rule (Figure 9) which projects a gesture daughter to a complex gesture-marked entity for which both the PROS and GEST features are appropriate.

```
gesture_lexrule := phrase_or_lexrule &
 [ ORTH [ PROS #pros,
          GEST no-gesture],
   ARGS <[ ORTH [ GEST gesture-form,
                  PROS p-word & #pros ]]>].
```

Figure 9: Definition of `gesture_lexrule`

In line with Definition 6.1, this rule constrains PROS to a prosodically prominent word of type *p-word* thereby preventing a gesture from plugging into a prosodically unmarked word. The *gesture-form* value is a supertype over the distinct gesture types—depicting and deictic. The GEST feature of the mother is of type *no-gesture* to block any further recursive instantiation of this rule. The `gesture_lexrule` is inherited by a lexical rule specific to depicting gestures, and by a lexical rule specific to deictic gestures. In this way, we can encode the semantic contribution of depicting gestures which is different from the semantic contribution of deixis. For the sake of space, Figure 10 presents only the `depicting_lexrule`. The semantic information contributed by the rule is encoded within C-CONT.

The rule introduces an underspecified *vis_rel* between the main label `#dltop` of the spoken sign (via the HCONS constraints) and the main label `#glbl` of the gesture semantics (via the HCONS constraints). Note that these two arguments are in a *geq* (greater or equal) constraint. This means that *vis_rel* can operate over any projection of the speech word; e.g., attaching the gesture to “mud” in (1) means that the relation is not restricted to the EPs contributed by “mud” but it

```
depicting_lexrule := gesture_lexrule &
[ARGS <[ SYNSEM.LOCAL.CONT.HOOK.LTOP #dltop,
         ORTH [ GEST depicting] >,
 C-CONT [ RELS <![ PRED vis_rel,
                   S-ARG #arg1,
                   G-ARG #arg2 ],
                 [ PRED G_mod,
                   LBL #glbl,
                   ARG1 #harg ],
                 [ LBL #larg1 ],...!>,
          HCONS <!geq&[ HARG #arg1,
                        LARG #dltop ],
                  qeq&[ HARG #arg2,
                        LARG #glbl ],
                  qeq&[ HARG #harg,
                        LARG #larg1 ],
                  ...!>]].
```

Figure 10: Definition of `depicting_lexrule`

| **Gesture Grammar Coverage Profile** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Aggregate** | **total items** ♯ | **positive items** ♯ | **word string** $\phi$ | **lexical items** $\phi$ | **distinct analyses** $\phi$ | **total results** ♯ | **overall coverage** % |
| $90 \leq$ *i-length* $< 95$ | 126 | 91 | 93.00 | 26.41 | 1.89 | 91 | 100.0 |
| $70 \leq$ *i-length* $< 75$ | 78 | 53 | 71.00 | 12.00 | 1.00 | 53 | 100.0 |
| $60 \leq$ *i-length* $< 65$ | 249 | 179 | 60.00 | 9.42 | 1.00 | 179 | 100.0 |
| $45 \leq$ *i-length* $< 50$ | 18 | 14 | 49.00 | 7.00 | 1.00 | 14 | 100.0 |
| **Total** | **471** | **337** | **70.18** | **14.31** | **1.24** | **337** | **100.0** |

Table 1: Coverage Profile of Test Items generated by [incr tsdb()]

can be also be over the EPs of a higher projection. Here, the implemented analysis differs from the theoretical one in that we formalise in semantics the gesture attachment ambiguities as per Situated Spoken Phrase Constraint: that is, *vis_rel* can operate over any projection of the gesture-marked sign.

The gesture's semantics is a bag of EPs, all of which are outscoped

by the gestural modality [$\mathscr{G}$]. The rule therefore introduces in RELS a label (here `#larg1`) for an EP which is in *qeq* constraints with [$\mathscr{G}$]. The instantiation of the particular EPs comes from the gestural lexical entry. In the real implementation, the number of these labels corresponds to the number of features.

The evaluation was performed in the tradition of testing wide-coverage grammars, by means of a manually crafted test suite (Oepen *et al.* 1997). We created a test suite covering different gesture types, prosody and the following linguistic phenomena: intransitivity, transitivity, complex NPs, modification, negation and coordination. The test set contained 471 speech-gesture items (71.5% well-formed and 28.5% ill-formed) covering the full range of prosodic (prosodic markedness and unmarkedness) and gesture (the span of depicting/deictic gesture and its temporal relation to the prosodically marked elements) permutations. The gestural vocabulary was limited since a larger gesture lexicon has no effects on the performance. To test the grammar, we used the [incr tsdb()] competence and performance tool (Oepen 2001) which enables batch processing of test items and which creates a coverage profile of the test set (see Table 1). The values are as follows: the left column separates the items per aggregation criterion (the length of test items);[10] the next column shows the number of test items per aggregate; then we have the number of grammatical items; average length of test item; average number of lexical items; average number of distinct analyses and total coverage.

We manually verified the coverage. While the grammar successfully parses all well-formed examples, the inclusion of a separate chart-mapping rule for concrete deixis results in overgeneration. We believe that the alternative method of enforcing strict precedence or strict sequence is too restrictive with respect to the possible interpretations supported by the distinct attachment configurations.

Finally, we also verified that the newly introduced rules did not change the coverage or increase the ambiguity of the existing broad-coverage grammar. We therefore ran both the ERG grammar and the gesture grammar on the ERG testsuite. The results shown in Table 2 were generated by both the ERG grammar and by the grammar

[10] Note the length here does not correspond to the actual length of tokens in each test item, since the tool also counts the XML tags.

| 'en-test-en-grm' Coverage Profile | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Aggregate** | **total items** ♯ | **positive items** ♯ | **word string** $\phi$ | **lexical items** $\phi$ | **distinct analyses** $\phi$ | **total results** ♯ | **overall coverage** % |
| $55 \leq$ *i-length* $< 60$ | 3 | 3 | 55.00 | 108.00 | 2.00 | 3 | 100.0 |
| $45 \leq$ *i-length* $< 50$ | 7 | 7 | 49.00 | 69.00 | 16.86 | 7 | 100.0 |
| $40 \leq$ *i-length* $< 45$ | 17 | 17 | 43.00 | 69.50 | 4.94 | 16 | 94.1 |
| $35 \leq$ *i-length* $< 40$ | 32 | 32 | 37.00 | 41.87 | 2.84 | 32 | 100.0 |
| $30 \leq$ *i-length* $< 35$ | 30 | 30 | 31.00 | 32.57 | 2.37 | 30 | 100.0 |
| $25 \leq$ *i-length* $< 30$ | 13 | 13 | 25.00 | 42.00 | 1.67 | 12 | 92.3 |
| $15 \leq$ *i-length* $< 20$ | 13 | 13 | 19.00 | 15.58 | 1.83 | 12 | 92.3 |
| **Total** | **115** | **115** | **34.13** | **43.99** | **3.63** | **112** | **97.4** |

(generated by [incr tsdb()] at 8-jul-2005 (04:42 h))

Table 2: [incr tsdb()] Coverage Profile of ERG Test Items Parsed by ERG and Gesture Grammar

equipped with the gesture rules. In other words, the gesture rules had no effects on the existing rules.

## 8 CONCLUSIONS

The work presented here advances a new theory in which the form-meaning mapping of speech-gesture actions was analysed using well-established methods from linguistics such as constraint-based syntactic derivation and semantic composition. In particular, we captured the mapping of form of speech-gesture actions to their meanings within a constraint-based grammar: the construction rules were inspired by examining real data and were further implemented within a wide-coverage grammar for English. The highly ambiguous gesture form was captured using underspecified semantics, which allowed us to account for the range of specific interpretations that a given gesture can take in its context of use. The ambiguities notwithstanding, we demonstrated that the speech-gesture attachments are constrained by the form of the speech signal, thus showing that the difference in ambiguity between linguistic input and gesture input is more a matter of degree than a difference in kind.

## ACKNOWLEDGEMENTS

The authors are very grateful to Elżbieta Hajnicz and the anonymous reviewers, Daniel Loehr, Matthew Stone, Mark Steedman, Emily Bender, Bob Ladd, Michael Johnston, Jonathan Kilgour, Ulrich Schäfer, Stephan Oepen, and also EPSRC for funding this work, as well as ERC (grant number 269427).

## REFERENCES

Dorit ABUSCH (2014), Temporal Succession and Aspectual Type in Visual Narrative, in Luka CRNIČ and Uli SAUERLAND, editors, *The Art and Craft of Semantics: A Festschrift for Irene Heim*, volume 1, pp. 9–29, MIT Working Papers in Linguistics, Cambride, MA.

Peter ADOLPHS, Stephan OEPEN, Ulrich CALLMEIER, Berthold CRYSMANN, Daniel FLICKINGER, and Bernd KIEFER (2008), Some Fine Points of Hybrid Natural Language Parsing, in *Proceedings of the Sixth International Language Resources and Evaluation*, ELRA.

Stergos AFANTENOS, Eric KOW, Nicholas ASHER, and Jeremy PERRET (2015), Discourse parsing for multi-party chat dialogues, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, pp. 928–937, Lisbon.

Hiyan ALSHAWI (1992), *The Core Language Engine*, Cambridge: MIT Press.

Nicholas ASHER and Alex LASCARIDES (1998), Bridging, *Journal of Semantics*, 15(1):83–113.

Nicholas ASHER and Alex LASCARIDES (2003), *Logics of Conversation*, Cambridge University Press.

Janet Beavin BAVELAS and Nicole CHOVIL (2006), Hand gestures and facial displays as part of language use in face-to-face dialogue, in V. MANUSOV and M. PATTERSON, editors, *Handbook of Nonverbal Communication*, pp. 97–115, Thousand Oaks, CA: Sage.

Johan BOS (2004), Computational Semantics in Discourse: Underspecification, Resolution, and Inference, *J. of Logic, Lang. and Inf.*, 13(2):139–157, ISSN 0925-8531, doi:10.1023/B:JLLI.0000024731.26883.86, `http://dx.doi.org/10.1023/B:JLLI.0000024731.26883.86`.

Jean CARLETTA (2006), Announcing the AMI Meeting Corpus, *The ELRA Newsletter*, 11(1):3–5.

Jean CARLETTA (2007), Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus, *Language Resources and Evaluation*, 41(2):181–190.

Justine CASSELL, David MCNEILL, and K.E. MCCULLOUGH (1999), Speech-Gesture Mismatches: Evidence for One Underlying Representation of Linguistic and Non-Linguistic Information, *Pragmatics and Cognition*, 7(1):1–33.

Ann COPESTAKE (2007), Semantic composition with (robust) minimal recursion semantics, in *DeepLP '07: Proceedings of the Workshop on Deep Linguistic Processing*, pp. 73–80, Association for Computational Linguistics, Morristown, NJ, USA.

Ann COPESTAKE and Ted BRISCOE (1995), Semi-Productive Polysemy and Sense Extension, *Journal of Semantics*, 12:15–67.

Ann COPESTAKE, Dan FLICKINGER, Ivan SAG, and Carl POLLARD (2005), Minimal Recursion Semantics: An introduction, *Journal of Research on Language and Computation*, 3(2–3):281–332.

Ann COPESTAKE, Alex LASCARIDES, and Dan FLICKINGER (2001), An Algebra for Semantic Construction in Constraint-based Grammars, in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL/EACL 2001)*, pp. 132–139, Toulouse.

Markus EGG, Alexander KOLLER, and Joachim NIEHREN (2001), The Constraint Language for Lambda Structures, *Journal of Logic, Language and Information*, 10:457–485, ISSN 0925-8531, doi:10.1023/A:1017964622902, `http://portal.acm.org/citation.cfm?id=595849.596040`.

Randi ENGLE (2000), *Toward a Theory of Multimodal Communication: Combining Speech, Gestures, Diagrams and Demonstrations in Structural Explanations*, Stanford University, phD thesis.

Dan FLICKINGER (2000), On Building a More Efficient Grammar by Exploiting Types, *Natural Language Engineering*, 6 (1) (Special Issue on Efficient Processing with HPSG):15 – 28.

Ellen FRICKE (2008), *Grundlagen einer multimodalen Grammatik des Deutschen: Syntaktische Strukturen und Funktionen. Habilitationsschrift*, Europa-Universität Viadrina Frankfurt (Oder), manuskript. 313 S., (Erscheint 2010 im Verlag de Gruyter.).

Gianluca GIORGOLO (2012), Integration of Gesture and Verbal Language: A Formal Semantics Approach, in Eleni EFTHIMIOU, Georgios KOUROUPETROGLOU, and Stavroula-Evita FOTINEA, editors, *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*, volume 7206 of *Lecture Notes in Computer Science*, pp. 216–227, Springer Berlin Heidelberg, ISBN 978-3-642-34181-6, doi:10.1007/978-3-642-34182-3_20, `http://dx.doi.org/10.1007/978-3-642-34182-3_20`.

Gianluca GIORGOLO and Ash ASUDEH (2011), Multimodal Communication in LFG: Gestures and the Correspondence Architecture , in Miriam BUTT and Tracy Holloway KING, editors, *The Proceedings of the LFG 2011 Conference*,

pp. 257–277, Hong Kong, http://cslipublications.stanford.edu/LFG/16/abstracts/lfg11abs-giorgoloasudeh2.html.

Gianluca GIORGOLO and Frans VERSTRATEN (2008), Perception of speech-and-gesture integration, in *Proceedings of the International Conference on Auditory-Visual Speech Processing 2008*, pp. 31–36.

Erving GOFFMAN (1963), *Behavior in Public Places: Notes on the Social Organization of Gatherings*, The Free Press.

Alex GRZANKOWSKI (2015), Pictures Have Propositional Content, *Review of Philosophy and Psychology*, 6(1):151–163, ISSN 1878-5158, doi:10.1007/s13164-014-0217-0, http://dx.doi.org/10.1007/s13164-014-0217-0.

Florian HAHN and Hannes RIESER (2010), Explaining Speech Gesture Alignment in MM Dialogue Using Gesture Typology, in Paweł ŁUPKOWSKI and Matthew PURVER, editors, *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 99–109, Polish Society for Cognitive Science, Poznań.

Jerry R HOBBS (1985), On the Coherence and Structure of Discourse, Technical report csli-85-37, Stanford University, Center for the Study of Language and Information.

Michael JOHNSTON (1998a), Multimodal Language Processing, in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia.

Michael JOHNSTON (1998b), Unification-based Multimodal Parsing, in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL 1998, pp. 624–630, Association for Computational Linguistics, Stroudsburg, PA, USA, doi:http://dx.doi.org/10.3115/980845.980949, http://dx.doi.org/10.3115/980845.980949.

Michael JOHNSTON, Philip R. COHEN, David MCGEE, Sharon L. OVIATT, James A. PITTMAN, and Ira SMITH (1997), Unification-Based Multimodal Integration, in Philip R. COHEN and Wolfgang WAHLSTER, editors, *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 281–288, Association for Computational Linguistics, Somerset, New Jersey.

David KAPLAN (1989), Demonstratives, in J. ALMOG, J. PERRY, and H. WETTSTEIN, editors, *Themes from Kaplan*, Oxford.

Andrew KEHLER (2002), *Coherence, Reference, and the Theory of Grammar*, CSLI Publications.

Ruth KEMPSON, Wilfried MEYER-VIOL, and Dov M GABBAY (2000), *Dynamic syntax: The flow of language understanding*, Wiley-Blackwell.

Adam KENDON (1972), Some relationships between body motion and speech, in A. SEIGMAN and B. POPE, editors, *Studies in Dyadic Communication*, pp. 177–216, Pergamon Press, Elmsford, New York.

Adam KENDON (2004), *Gesture. Visible Action as Utterance*, Cambridge University Press, Cambridge.

Ewan KLEIN (2000), A constraint-based approach to English prosodic constituents, in *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp. 217–224, Association for Computational Linguistics, Morristown, NJ, USA, doi:http://dx.doi.org/10.3115/1075218.1075246.

Alexander KOLLER, Michaela REGNERI, and Stefan THATER (2008), Regular tree grammars as a formalism for scope underspecification, in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, Columbus, Ohio.

Stefan KOPP, Paul TEPPER, and Justine CASSELL (2004), Towards integrated microplanning of language and iconic gesture for multimodal output, in *ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces*, pp. 97–104, State College, PA, USA, ACM, New York, NY, USA, ISBN 1-58113-995-0, doi:http://doi.acm.org/10.1145/1027933.1027952.

Stefan KOPP, Paul A. TEPPER, Kimberley FERRIMAN, Kristina STRIEGNITZ, and Justine CASSELL (2007), *Trading Spaces: How Humans and Humanoids Use Speech and Gesture to Give Directions*, pp. 133–160, John Wiley & Sons, Ltd, ISBN 9780470512470, doi:10.1002/9780470512470.ch8, `http://dx.doi.org/10.1002/9780470512470.ch8`.

Peter KÜHNLEIN, Manja NIMKE, and Jens STEGMANN (2002), Towards an HPSG-based Formalism for the Integration of Speech and Co-Verbal Pointing, in *Proceedings of Gesture – The Living Medium*, Austin, Texas.

Alex LASCARIDES and Matthew STONE (2006), Formal Semantics for Iconic Gesture, in *Proceedings of Brandial'06, the 10th International Workshop on the Semantics and Pragmatics of Dialogue (SemDial10)*, pp. 125–132, Universitätsverlag Potsdam, Potsdam, Germany.

Alex LASCARIDES and Matthew STONE (2009a), Discourse Coherence and Gesture Interpretation, *Gesture*, 9(2):147–180.

Alex LASCARIDES and Matthew STONE (2009b), A Formal Semantic Analysis of Gesture, *Journal of Semantics*, 26(4):393–449.

Stephen C. LEVINSON (1983), *Pragmatics*, Cambridge University Press, Cambrdige.

Daniel LOEHR (2004), *Gesture and Intonation*, Georgetown University, Washington DC, doctoral Dissertation.

Andy LÜCKING, Hannes RIESER, and Marc STAUDACHER (2006a), Multi-modal Integration for Gesture and Speech, in David SCHLANGEN and Raquel FERNÁNDEZ, editors, *brandial'06 – Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 106–113, Universitätsverlag Potsdam, Potsdam.

Andy LÜCKING, Hannes RIESER, and Marc STAUDACHER (2006b), SDRT and Multi-modal Situated Communication, in David SCHLANGEN and Raquel FERNÁNDEZ, editors, *brandial'06 – Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 72–79, Universitätsverlag Potsdam, Potsdam.

David MCNEILL (1992), *Hand and Mind. What Gestures Reveal about Thought*, University of Chicago Press, Chicago.

David MCNEILL (2005), *Gesture and Thought*, University of Chicago Press, Chicago.

Richard MONTAGUE (1988), The Proper Treatment of Quantification in Ordinary English, in Jack KULAS, JamesH. FETZER, and TerryL. RANKIN, editors, *Philosophy, Language, and Artificial Intelligence*, volume 2 of *Studies in Cognitive Systems*, pp. 141–162, Springer Netherlands, ISBN 978-94-010-7726-2, doi:10.1007/978-94-009-2727-8_7, http://dx.doi.org/10.1007/978-94-009-2727-8_7.

Jana Bressem MÜLLER, Cornelia and Silva H. LADEWIG (2013), Towards a grammar of gesture – a form based view, *Body–Language–Communication: An International Handbook on Multimodality in Human Interaction. (Handbooks of Linguistics and Communication Science 38.1.)*, pp. 707–733.

Stephan OEPEN (2001), [incr tsdb()] — Competence and Performance Laboratory. User Manual, Technical report, Computational Linguistics, Saarland University, Saarbrücken, Germany.

Stephan OEPEN, Klaus NETTER, and Judith KLEIN (1997), TSNLP — Test Suites for Natural Language Processing, in John NERBONNE, editor, *Linguistic Databases*, pp. 13–36, CSLI Publications, Stanford, CA.

Patrizia PAGGIO and Costanza NAVARRETTA (2009), Integration and representation issues in the annotation of multimodal data, in Costanza NAVARRETTA, Patrizia PAGGIO, Jens ALLWOOD, Elisabeth ALSÉN, and Yasuhiro KATAGIRI, editors, *Proceedings of the NODALIDA 2009 workshop Multimodal Communication — from Human Behaviour to Computational Models*, volume 6, pp. 25–31, Northern European Association for Language Technology (NEALT).

Thies PFEIFFER, Florian HOFMANN, Florian HAHN, Hannes RIESER, and Insa RÖPKE (2013), Gesture Semantics Reconstruction Based on Motion Capturing and Complex Event Processing: a Circular Shape Example, in *Proceedings of the*

*SIGDIAL 2013 Conference*, pp. 270–279, Association for Computational Linguistics, http://aclweb.org/anthology/W13-4041.

Livia POLANYI (1985), A Theory of Discourse Structure and Discourse Coherence, in *Proceedings of the 21st Meeting of the Chicago Linguistics Society*, Chicago, Illinois: Linguistics Department, University of Chicago.

Uwe REYLE (1993), Dealing with Ambiguities by Underspecification: Construction, Representation and Deduction, *Journal of Semantics*, 10:123–179.

I. A. SAG and T. A. WASOW (1999), *Syntactic Theory: A Formal Introduction*, Center for the Study of Language and Information, Stanford, California, ISBN 1575861615 (hard cover), 1575861607 (paper).

Mark STEEDMAN (2000), *The Syntactic Process*, The MIT Press.

Francis & Mark Turner STEEN (2013), Multimodal Construction Grammar, *Language and the Creative Mind*, pp. 255–274.