Abstract

This paper addresses the form-meaning relation of multimodal communicative actions by means of a grammar that combines verbal input with hand gestures. Unlike speech, gesture signals are interpretable only through their semantic relation to the synchronous speech content. This relation serves to resolve the incomplete meaning that is revealed by gestural form alone. We demonstrate that by using standard linguistic methods, speech and gesture can be integrated in a constrained way into a single derivation tree which maps to a uniform meaning representation.

1 Introduction

Meaning in everyday communication is conveyed by a complex mixture of signals that includes the situated and dynamic context of language production and language perception. In face-to-face interaction, people rely on *utterance visible actions* (Kendon, 2004) to exchange information. For instance, in a multi-party conversation, a pronoun is often resolved by a pointing gesture towards the intended addressee; in living-space descriptions, people often create a virtual map so as to point to a designated location; when narrating stories people use hand movements to depict events or to provide visual characteristics of an object.

This project is concerned with embodied actions—also known as 'gesticulation', 'co-verbal gestures' or 'co-speech gestures'—that use the hand as a semantically intended medium for communication. The specific property of hand gestures is their *synchrony* with the co-occurring speech: a single thought is expressed synchronously in speech and in gesture, and is perceived as an integrated multimodal ensemble (McNeill, 2005). The synchronous nature of the multimodal signal is observed with the semantic relation between speech and gesture being one of redundancy (that is, the gestural signal "repeats" visually the spoken words) or a relation of complementarity (that is, the gesture adds propositional content to the final utterance). Whereas redundancy is not favoured in speech only, speech-gesture redundancy does not violate coherence (Lascarides and Stone, 2009), and it can facilitate learning and enhance expressiveness (Buisine and Martin, 2007).

In this project, we approach synchrony in multimodality by elevating formal language models to a description of multimodal input. In particular, we use well-established methods for composing a semantic representation of a signal from a representation of its form so as to provide a form-meaning mapping for multimodal communicative actions, consisting of spoken phrases and co-speech gestures. This will be achieved by developing a constraint-based multimodal grammar that takes verbal signals and hand gestures as input. The grammar captures generalisations about the well-formedness of the multimodal signal. Within the multimodal grammar one can elegantly capture the linguistic and visuo-spatial linkages at a conceptual level that trigger the synchronous production of speech and gesture: for instance, representing the interaction between a spoken signal and its synchronous

gesture is a matter of constraining the choices of speech-gesture integration in the grammar.

Our focus of study are spontaneous and improvised co-speech gestures that communicate meaning:[1] *depicting (representative)* gestures depict, model the object of reference or enact a specific behaviour. The depiction can be literal (also known as iconic gestures), e.g., making a round shape with hands when talking about a cake, or metaphoric, e.g, moving the hand from the left to the right periphery to refer to the past and the future. *Pointing (deictic)* gestures can identify concrete coordinates in Euclidean space (Lascarides and Stone, 2009), point to an abstract object in the virtual space (McNeill, 2005), or even nominate as prominent a word or a phrase (Kendon, 2004). *Performative (pragmatic)* signals perform a speech act, e.g., the hand moves away across the body with palm facing down to express negation. Finally, in *interactive* gestures, the hand is used as an interaction regulator as when extending an open hand towards the addressee to offer them the floor (Bavelas et al., 1995). Other spontaneous communicative actions include *beats*. These are formless flicks of the hand that beats time along with the rhythm of the speech, and they often serve pragmatic functions such as commenting on one's own utterance or giving prominence to aspects of the speech (Cassell, 2000).

The gesture categories do not form a typology of distinct classes; rather, they are spread among mutually inclusive dimensions, and so a single gesture can exhibit traces of one or more dimensions (McNeill, 2005). Utterance (1) taken from a corpus collected and annotated by Loehr (2004) exemplifies such multidimensional gesture: the horizontal hand movement with palms facing down literally depicts some salient feature of the synchronous speech content, namely objects positioned at the bottom, and at the same time this gesture is a recurrent metaphor of a completion of a process.[2]

(1) the BOTTOM worked FINE
Hands are rested on the knees and elevate to the body centre with palms facing downwards. Right and left hand perform a horizontal movement to the right and left periphery, respectively.

2 Main Challenges

We shall now address the major challenges arising from the ambiguous form of gesture. Considered out of specific context, the form of a hand signal is massively ambiguous, potentially mapping to open-ended meanings. For instance, a rotating hand movement performed by the whole hand can resemble the circular motion of an object such as a mixer or a wheel; it can also be a visual representation of

[1] The classification that follows is largely based on Kendon (2004).

[2] Throughout this work, small caps are used to indicate the pitch accented words and underlining is used to indicate the verbal segment temporally aligned with gesture; the gesture's transcription is given in italics after the verbal string.

the object being rotated by the hand; or each iteration can indicate distinct steps in an iterative process. Of course, many other propositions can be characterised by this hand movement. This is very roughly analogous to lexical sense ambiguity in language, where polysemous words can map to open-ended meanings if ones takes generative properties such as metaphor and nonce uses into account (Pustejovsky, 1995).

Further ambiguities concern the gestural category—representative or deictic—which affects the syntactic context. This ambiguity is useful as it allows us to differentiate between spatial and non-spatial content: deictic gestures provide spatial reference in the virtual situation and should thus receive spatial values, whereas representative gestures require qualitative values (Lascarides and Stone, 2009). A rough linguistic analogy is, for instance, the distinct categories of "duck"—a noun or a verb—leading to the syntactically ambiguous sentence "I saw her duck". The way this syntactic ambiguity is resolved depends on the context of use and resolving this ambiguity in form is logically co-dependent with resolving its interpretation in context: "I saw her duck, geese and chickens" would yield a syntactic and corresponding meaning representation distinct from that of "I saw her duck and hide in the hay".

Neither the form of the gesture nor the form of speech uniquely determine the linguistic phrase synchronous with gesture.[3] Following Lascarides and Stone (2009), we assume that computing the rhetorical connections between a gesture and its synchronous phrase, and resolving the meaning of the gesture to a specific value are logically co-dependent. With this in mind, consider the real example in (2) (Loehr, 2004).

(2) If I was TO REALLY TEACH someone how to be a professional musician ...
Hands are in the central space in front of speaker's body; palms face horizontally upwards. Along with "really", both hands perform a quick downward movement; possibly a conduit metaphor

Here the gesture stroke was performed while uttering the pre-head modifier, while annotators interpreted the gesture meaning as one where the open hands express the conduit metaphor (Lakoff and Johnson, 1980). The fact that annotators interpreted it in this way suggests that quantitative criteria alone—such as the timing of speech relative to gesture—are not sufficient to define adequate constraints on synchrony. This example also illustrates that in syntax, the gesture stroke interacts with the head daughter of the speech phrase, and in semantics, the content of the gesture is semantically related to that of the whole clause, in which way, the agent, patient and the idea transferred between them via teaching all serve to resolve the values of the participants in the conduit metaphor that is expressed by the gesture. However, this conduit interpretation is not available if the gesture temporally overlaps with only the subject daughter itself. Intuitively in this case, the

[3] In this paper, the term 'gesture' designates the expressive part of the whole movement, the kinetic peak of the excursion that carries the gesture's meaning—the so called *gesture stroke*.

gesture would simply denote the individual denoted by the subject, perhaps also placing him in a particular place that carries meaning. Finally, the gesture can receive a pragmatic interpretation that is paraphrasable as the parenthetical expression "I am informing you", which is possible by attaching the gesture to the S node. Despite the ambiguities in context, the result does not violate coherence—coherent multimodal actions tolerate certain unresolved ambiguities in interpretation, just as purely linguistic ones do.

Nonetheless, speech-gesture synchrony is not a free-for-all and our challenge is to identify the factors that make a multimodal act ill-formed. There is evidence in the literature that temporal alignment affects perception of speech and gesture integration, and the parameter that plays a role in perceiving a multimodal action as well-formed is prosody (Giorgolo and Verstraten, 2008).

To illustrate the effects of prosody on speech-gesture synchrony, consider the constructed example (3). Here it seems anomalous to perform the gesture on the unaccented "called" even though the gesture is intended as a depiction of something related to the act of calling. This ill-formedness would not arise if the gesture was placed along the whole utterance or a part of it which includes the prosodically prominent element "mother".

(3) * Your MOTHER called today.
The speaker puts his hand to the ear to imitate holding a receiver.

Ambiguity does not contradict our prediction that spontaneous gestures are a semantically intended communication source. In fact, they partially constrain the set of possible interpretations: this observation is valid not only for deictic and performative gestures whose recurrent form and orientation in the virtual space maps to a limited set of possible interpretations, but also for representative gestures whose imagistic resemblance with the object of reference is linked to an abstract meaning. By constructing a multimodal grammar we shall provide a methodology for the derivation of all possible interpretations in a specific context-of-use and for constraining speech-gesture ill-formedness.

We address the ambiguity of a disambiguated multimodal form by producing an underspecified logical formula which gives an abstract representation of what the signal means in any context of use. So, this abstract representation must support the full variety of specific interpretations of the gesture that occurs in different discourse contexts. How exactly it is going to resolve to a preferred value is a matter of discourse processing that is beyond the scope of our current goals. Multimodal ill-formedness is addressed by providing linguistic constraints of when speech and gesture can be synchronised. In this way, we address in a qualitative way the quantitative finding of Giorgolo and Verstraten (2008).

3 Form and Meaning of Gesture

Contrary to the decompositional analysis of lexical items or the semantic compositional approach to natural language, the meaning of a gesture cannot be determined decompositionally (McNeill, 2005).[4] A gesture obtains its meaning after conjoining the various gesture features—the shape of the hand, the orientation of the palm and fingers, the location of the hand and the direction of the movement—and linking them to the context of the accompanying speech. Recall that some ambiguity about the 'transfer' conduit (2) remains, and so formalising gesture content requires the framework to support ambiguity in coherent actions. The holistic aspect of gesture's form requires a description that is distinct from the tree descriptions of linguistic phrases. We benefit from previous unification-based models of gesture (Johnston, 1998), (Kopp et al., 2004) to formally regiment the contribution of each aspect of gesture in terms of Typed Feature Structures (TFSs). For instance, the form of the gesture in utterance (2) is representend in Figure 1. The representation is typed as *depicting_metaphoric* so as to distinguish the form features constrained by depiction from those constrained by spatial reference (Lascarides and Stone, 2009).

$$\begin{bmatrix} \textit{depicting\_metaphoric} & \\ \text{HAND-SHAPE:} & \text{open-flat} \\ \text{PALM-ORIENTATION:} & \text{upwards} \\ \text{FINGER-ORIENTATION:} & \text{forward} \\ \text{HAND-LOCATION:} & \text{centre-low} \\ \text{HAND-MOVEMENT:} & \text{straight-down} \end{bmatrix}$$

Figure 1: TFS Representation of Gesture Form

Following previous research on semantics of gesture (Lascarides and Stone, 2009), we use the framework of Robust Minimal Recursion Semantics (RMRS) (Copestake, 2007) to provide a form-meaning mapping of embodied actions. RMRS is fully flexible in the type of semantic underspecification it supports: one can easily leave the predicate's arity and the type of the arguments underspecified until resolved by the discourse context, for instance. This is useful, because each form feature value can resolve to a wide range of fully specific predications in context, and these possibilities are not of unique arity. For instance, the downward movement in (2) can be interpreted as offering knowledge that is held by the open hand. In this case, the logical form contributed by the movement should be a three-place predicate denoting an event $teach(e, x, y)$. On the other hand, the movement in the same gesture that is performed in the different (constructed) speech context (4) depicts the uniformity of the shape of the keel of boat, from the port to the starboard, which by the hand shape is curved. Thus here the movement resolves to the one-place predicate $uniform(x)$ where x denotes the shape of the keel.

[4]There are attempts of hierarchical organisation of gesture ((Fricke, 2008), inter alia) similar to the hierarchically organised syntactic constituents but these are at the level of the entire hand excursion from a rest position to its retraction to a rest, also knows as a *gesture unit*.

(4) The boat's keel is curved
same gesture as in (2)

Form-meaning mapping from a gesture stroke to its highly underspecified semantic representation consists in reading the gesture's predications directly off the feature structure as shown in Figure 2.

$$
\begin{array}{l}
l_0 : a_0 : [\mathcal{G}](h) \\
l_1 : a_1 : hand\_shape\_open\_flat(i_1), \\
l_2 : a_2 : palm\_orientation\_upwards(i_2), \\
l_3 : a_3 : finger\_orientation\_forward(i_3), \\
l_4 : a_4 : hand\_location\_centre\_low(i_4), \\
l_5 : a_5 : hand\_movement\_straight\_down(i_5) \\
h =_q l_i \;\; where \;\; 1 \leq i \leq 5
\end{array}
$$

Figure 2: RMRS Representation of Gesture Meaning

Each predication is associated with a (not necessarily unique) label ($l_0 \ldots l_5$), a unique anchor ($a_0 \ldots a_5$) and an index variable ($i_1 \ldots i_5$) that underspecifies its main argument. The label is used to determine the scopal position of its predicate in the logical form (so Figure 2 exhibits semantic scope ambiguities among the resolved predications). The anchor for each predication is used as a locus for adding arguments to the predication—for instance, $ARG(a, x)$ means that $hand\_shape\_open\_flat$ resolves in context to a predicate that takes (at least) two arguments and the second is x. The predication $hand\_shape\_open\_flat(i_1)$ underspecifies the referent i_1 depicted through the hand shape of the hand (i_1 can resolve to an individual variable x or to an event variable e). An RMRS predicate is resolved to a specific predicate (or a combination thereof) on the semantics/pragmatics interface. The range of possible specific predicates that a given predication can resolve to is limited by iconicity (Lascarides and Stone, 2009). Further, Lascarides and Stone (2009) motivate the introduction of an operator $[\mathcal{G}]$ that limits the scope of the predicates within the gesture modality. This captures constraints on co-reference between speech and gesture, and across different gestures.

The gesture's interpretation in context is logically co-dependent on how it is coherently related to its synchronous speech. Lascarides and Stone (2009) argue that there is an inventory of semantic relations between the gesture and the linguistic phrase: for instance, the gesture can *depict*, *elaborate*, *explain*, but not *contrast with* the information introduced by speech. In the grammar, we shall therefore introduce in semantics an underspecified semantic relation $vis\_rel(s, g)$ between the content denoted by a speech s daughter and the content denoted by a depicting gesture g daughter when they are combined via a grammar construction rule that reflects that s is the linguistic phrase that g is synchronous with. How this relation resolves is a matter of commonsense-reasoning. This is similar to the treatment of free adjuncts in language: the covert relationship between the content of the main clause and the proposition of the free adjunct must be determined in pragmatics.

4 Speech-Gesture Synchrony

4.1 What is Synchrony?

There is a very broad consensus within the gesture community that speech and co-speech communicative actions function in *synchrony* to convey an integrated message (McNeill, 2005), (Kendon, 2004). However, the conditions on synchrony are controversial: is synchrony defined solely in terms of temporal alignment (McNeill, 2005), (Engle, 2000) or are there other prevailing conditions (Oviatt et al., 1997)? Further confusion arises as to what the criteria are when considering the temporal extension of the gesture: is it the gesture stroke that is temporally aligned with the spoken signal, the gesture phrase from its beginning to its semantic peak, or the entire gesture excursion from a rest to a rest. We therefore start by working out our own definition as follows:

Definition 1 (Synchrony) *The choice of which linguistic phrase a gesture stroke is synchronous with is guided by: 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 an underspecified logical formula supporting (ii) and hence also (i).*

Whereas synchrony has already been defined in terms of (i) and (ii), the last factor is our contribution: we exploit standard methods for constructing form and meaning in formal grammars to constrain the choices of integrating speech and gesture into a single derivation tree, and thus to derive logical forms from syntax. An overall challenge is to constrain synchrony in a way that rules out ill-formed multimodal input, and nevertheless enables the derivation of highly underspecified logical formulae for well-formed input that will support pragmatic inference and resolve to preferred values in specific contexts. Note that this definition abandons simultaneity as a condition on synchrony. As attested in (2) and (3), this dovetails with the fact that our own perceptual system can make the judgement of which signals are synchronised and which are not.

The constraints on integrating speech and co-speech gesture into a single tree are guided by prosody (the literature offers enough evidence for the prosody-gesture interaction (Kendon, 1972), (McClave, 1991), (Loehr, 2004), (Giorgolo and Verstraten, 2008) inter alia), syntax (recall (2) and its subsequent discussion), and also the temporal performance of gesture relative to speech.

While there is a clear interaction between gesture and prosody, and between gesture and syntax-semantics of speech, we remain agnostic as to whether gesture, its dimension(s), content and composing phases interact with the distribution of information into theme and rheme. Cassell (2000) hypothesises that the type of *relation* between gesture and speech plays a central role in combining with either thematic or rhematic utterances. This information might be needed by a discourse processor but we are not convinced that information structure should constrain the

choices of attachment for linguistic phrases and gesture within the grammar. So, in the absence of convincing empirical evidence that speech-gesture synchrony is informed by the type of the tone and correspondingly, by the thematic and rhematic functions of an utterance, we shall limit ourselves to prosody, syntax-semantics and timing as central factors for combining speech with gesture within the grammar to produce a unified meaning representation.

4.2 Empirical Investigation

To spell out constraints on speech-gesture integration, we conducted empirical investigation on a 165-second collection of four recorded meetings annotated for gesture and intonation (Loehr, 2004). Our experiments were intended to shed light on the following questions: Does the temporal performance of gesture relative to speech constrain the choices of integrating gesture into the parse tree? Do gestures occur with a particular syntactic constituent, if any at all? Is the gesture stroke performed while uttering a prosodically prominent syllable?

Gesture and Syntax To check for the interaction between communicative gestures and syntax, we assigned syntactic labels to the gesture strokes. This analysis was preceded by a preprocessing step which involved insertion of sentence boundaries, replacement of shortened forms with the corresponding long ones (e.g., "I've" $>$ "I have"), and also replacement of the filled and unfilled pauses with dummy words to handle incomplete grammatical slots.

The syntactic annotation was strictly driven by the temporal performance of gesture relative to speech, and in particular, by the type of the overlap relation between gesture and speech. In general, we observed three (not necessarily exclusive) temporal relations of a gesture (G) overlapping the relevant spoken word(s) (S): (1) inclusion where $S\ during\ G$; (2) precedence where $start(G) \prec start(S)$ and/or $end(G) \prec end(S)$, i.e., the stroke starts or ends at some midpoint of the spoken word, and (3) sequence where $start(G) \succ start(S)$ and/or $end(G) \succ end(S)$, i.e., the stroke starts or ends at some midpoint of the spoken word. In case of inclusion, we have assigned the corresponding part-of-speech or syntactic labels of the included word(s). In case of precedence/sequence, there is generally a choice as to whether to include those midpoint words: provided that these word(s) were part of a syntactic constituent, they were included in the labelling, and otherwise they were ignored. Of course, if the inclusion (exclusion) of the midpoint words lead to distinct syntactic labels, all of the possibilities have been captured. And if the words overlapping the gesture did not form a syntactic constituent, this has been labelled as a "Non-constituent". Moreover, whenever the gesture starts at the midpoint of $word_1$ and finishes at midpoint of $word_2$, the gesture has been annotated in terms of the label of $word_1$, $word_2$ and their common syntactic label (if available). The results of the syntactic categories assigned to gesture strokes (G) are summarised in Table 1. Since every gesture potentially maps to more than one syntactic category, the total number of labels exceeds 100%.

Syntactic Category of G	Percent	Syntactic Category of G	Percent
S	6.38%	RB	7.45%
VP	10.64%	TO	2.13%
V *(present and past verb forms, base forms, modal verbs, present and past participles)*	27.66%	JJ *(positive and comparative adjectives)*	5.32%
NP	20.21%	DT	13.83%
NN *(singular and plural)*	9.57%	UH	1.06%
PRP *(personal and possessive)*	20.21%	C *(coordinating or subordinating conjunction)*	6.38%
IN	5.31%	Pause	8.51%
PP	1.06%	Non-constituent	6.38 %

Table 1: Gesture-Syntax Correlation

Discussion On the sole basis of the temporal performance of gesture relative to speech, the mapping of a gesture to a syntactic phrase is one-to-many without any restrictions on the syntactic category. Further, when a gesture overlaps a verbal head (a single verb form, a verb phrase, or an entire sentence), the ambiguous form of the hand signal often does not fully constrain the attachment of gesture to a particular tree node. This attachment ambiguity is observed with gestures spanning a verb only, a verb phrase, or an entire sentence, thereby allowing for more mappings beyond the strict temporal performance. To illustrate this, consider utterance (5) where the gesture stroke overlaps an entire sentence.

(5) So he MIXes MUD . . .
Speaker's left hand is rested on the knee in ASL-B, palm extended and facing up as if holding something. Right hand performs consecutively four rotation movements over the left palm.

Here there is ambiguity as to whether the contextually specific interpretation of the circular hand movement addresses the content of the verb arguments "mud" and "he". Specifically, there is not sufficient information coming from form whether this gesture is a literal depiction of a mixing action, or the hand signal elaborates on the speech by showing the manner of executing the mixing action over the object, or even that the hand signal enacts the event of mixing mud from the speaker's viewpoint, and the hand is thus an extension of the actor's body performing the mixing. Note that these ambiguities would also arise if the gesture was performed while uttering "mixes" only or even "he mixes".

To address these multiple possibilities, in the grammar we shall define rules where the synchronous phrase can be derived by attaching gesture to the verb head daughter and extending it over the arguments to the head, thereby allowing for a gesture to attach to the head only, and also to a (syntactic and/or prosodic) constituent. In this way, we shall address two important issues: firstly, synchrony cannot be defined solely in terms of temporal alignment, i.e., the incomplete meaning of gesture as derived from form does not constrain the synchronous phrase;

secondly, the inclusion of the arguments is grounded in the *synthetic* nature of gesture versus the *analytic* nature of the spoken words, for instance, the information about an event, the object of the event and the agent can be provided by a singular gesture performance and several linearly ordered lexical items (McNeill, 2005). A single utterance can thus receive more than one correct parse analysis where each one contributes a distinct relation between the speech daughter and the gesture daughter.

We predict that the same principle of exploring synchronicity beyond the strict temporal alignment can be applied to gestures overlapping a word sequence that does not form a syntactic constituent, and also to gestures overlapping a prepositional, adjectival or a noun head. Utterance (6) (McNeill, 2005) demonstrates that gestures can be extended over the preposition head arguments.

(6) and he goes up THROUGH the drainpipe
Right hand is extended forward, palm facing up, fingers are bent in an upward direction. The hand shape resembles a cup.

The stroke temporally overlapping with the preposition denotes some salient feature of upward direction and "interiority" (McNeill, 2005). One possible synchronous phrase is the gesture signal combined with the co-temporal verb particle and preposition (McNeill, 2005). From this perspective, the gesture *complements* the denotation of the temporally aligned elements by narrowing down to a specific content. Our prediction for the non-unique gesture attachment possibilities would also favour an attachment to a larger phrase containing the object, "the drainpipe". We anticipate that both synchronous analyses are legitimate and should be obtainable by the grammar so as to provide the necessary underspecified relations resolvable by contextual knowledge.

Similarly, we predict that in case of gestures overlapping non-head daughters such as determiners or modifiers, the synchronous phrase is obtained by linking the gesture to the non-head daughter, but also to a larger phrase resulting from the unification of the non-head daughter with its head. In this way, the information coming from the head can also serve to resolve the contextually specific interpretation (recall (2)).

As for gestures overlapping nouns and noun phrases, we predict that the type of relation between gesture and speech could possibly determine the preferred attachment. In example (1), for instance, the interpretation where the hand movement represents literally the bottom cupboards can be obtained by attaching the gesture to the overlapping noun phrase. At the same time, the gesture can resolve to the metaphor of completing some process only by an S attachment. We therefore intend to explore the type of relation $R(s, g)$ between the s speech daughter and the g daughter so as to provide all plausible contextually specific interpretations.

Since there is not enough evidence about the semantic interaction between a gesture and the rest of the syntactic labels, interjections, and conjunctions, we shall leave them for future research. Finally, gestures happening along an unfilled or a filled pause are not envisaged by the grammar performance.

Gesture and Prosody In his doctoral dissertation, Loehr (2004) sought evidence for simultaneity in the performance of the pitch accent and the gesture apex, i.e., the most prominent part of gesture which unlike the stroke and the post-stroke hold does not span some interval. Conversely, we need prosody inasmuch as it is a possible constraint on gesture form, particularly on the contentful part of gesture (see example (3)).

To test for correlations between the pitch-accented element and the gesture stroke, we checked automatically the number of strokes temporally overlapping a pitch accent. The statistical analysis was performed after removing the gestures overlapping non-communicative hand movements[5] and (filled or unfilled) speech pauses. The results are summarised in Table 2.

Temporal Overlap	**Percent**
Gesture stroke and pitch accent	78.41%
Gesture stroke and pitch accent $< 250msec$	97.73%

Table 2: Gesture-Prosody Correlation

Discussion The statistical analysis showed that 78.41% of the gesture strokes were overlapped by a pitch accent. Then we relaxed the overlap by plus/minus 250 msec which is the average duration of a word in the corpus. Under this condition, the gesture stroke-pitch overlap raised to 97.73% (there were two events performed with a positive or negative delay of 250–320 msec). Essentially, none of the words performed within these extra miliseconds crossed a constituency boundary: for instance the pitch was on the pre-head modifier or on the complement of the argument temporally aligned with the gesture stroke. Within the grammar, we shall therefore provide rules for attaching gesture to a phrase larger than the single prosodically prominent lexical item temporally aligned with the gesture stroke. This also motivates our prediction that gestures can be synchronised with a constituent larger than the element temporally aligning the gesture stroke. In this way, we address by means of qualitative criteria the findings of Giorgolo and Verstraten (2008) and the descriptive studies detailing the synthetic nature of gesture (McNeill, 2005). A possible way to think of this extension beyond the temporal alignment is that syntactically, gestures are roughly analogous to lexical items and semantically, they are analogous to utterances. By attaching gesture 'higher' than the temporally co-occurring item, we allow for establishing a speech-gesture relation after having exploited the semantics of a larger spoken phrase and the semantics of the gesture.

The empirical study also demonstrated that while prosody can make a multimodal utterance ill-formed, in syntax there is generally several choices for attaching gesture to a speech constituent. It is thus essential to find the right balance between prosodic well-formedness and the possible syntactic attachments.

[5] In the gesture community, non-communicative hand gestures are usually referred to as *adaptors*. These are practically grounded, meaningless bodily movements such as nervous ticks or movements satisfying bodily needs such as rubbing the eyes or scratching one's nose.

5 An HPSG-based Account

We choose the framework of HPSG to spell out the theoretical principles of the multimodal grammar. This extends Johnston's (1998) HPSG analysis of gesture to cover a wider variety of gestures and to regiment their *domain-independent* constraints on form and meaning. Our motivation to use HPSG stems from its mechanisms to induce structural prosody in parallel with the derivation of syntactic structures (Klein, 2000). In so doing, we show that isomorphism between prosodic and surface syntactic structures is not necessary for encoding well-formedness constraints. Moreover, the semantic component in HPSG is based on Minimal Recursion Semantics (MRS) (Copestake et al., 2005) which is entirely compatible with RMRS, the framework we need for representing the highly underspecified content of gesture given its form (see § 3). Finally, the grammar can be easily augmented with tone/information structure constraints (Haji-Abdolhosseini, 2003) once we establish whether there is evidence for a direct interaction between on one hand, the tonal type and hence the information type, and on the other hand, the gesture performance.

As detailed in § 1, gestures are multidimensional. We regiment this formally in a multiple inheritance type hierarchy (Pollard and Sag, 1994), as shown in Figure 3. In this way, a gesture consisting of, say, deictic and depicting dimensions can inherit information from the type *concrete* and the type *literal*.

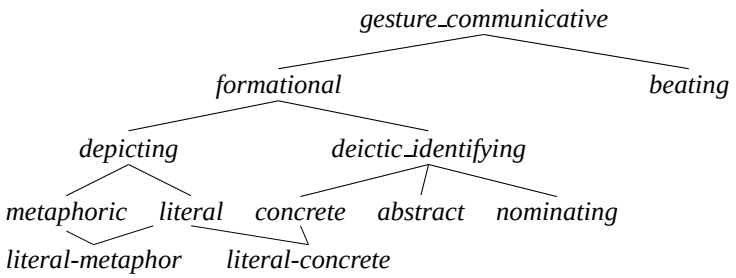

Figure 3: A Fragment of the Gesture Type Hierarchy

The type hierarchy of gestures is based on whether the form of the hand signal contributes some aspect of its meaning or not. In the former case, we distinguish *formational* actions, and in the latter, we talk about *beating*. The formational type subsumes *literal* depiction to account for form features which literally depict the object of reference, and *metaphoric* depiction where the form features are used as a metaphoric representation of the object of reference. Descriptive studies on deixis suggest that the form of the hand is dependent on its context and intended meaning. For instance, if the speaker designates an individual, the pointing is typically performed with an extended index finger, and if the speaker points to a class of objects, to an object exemplar, the pointing hand is typically open up (Kendon, 2004). This motivates us to represent deictic gestures as a subtype of *formational*. The deictic subtypes account for the distinct relations between the pointing signal

and the referent: the hand can identify a *concrete* referent at the spatio-temporal coordinates; it can point to an *abstract* representation of the referent; it can also *nominate* certain words or phrases as more prominent. Formless beat-like movements are typed as *beating*. This type hierarchy is intended as an illustration of gestural organisation rather than an exhaustive hierarchy of the possible gestural dimensions.

The mapping from hand movement to types on this hierarchy is one to many, thereby providing a representation of ambiguity about whether a communicative gesture is deictic, depicting, or a mixture thereof, and the ambiguity is resolvable only through its relation to speech. For this reason while investigating depicting and deictic gestures, we will analyse them in terms of this multidimensional perspective.

Synchronising linguistic and gestural input in the derivation tree involves unifying a feature structure typed as *gesture_communicative* (or any of its subtypes) and a feature structure typed as *spoken_sign* (or any of its subtypes). Upon unification, the multimodal signal is of type *depict(ing)_sign* which subsumes *depict_word*, *depict_phrase* and *depict_mtr(τ)*. The multimodal type hierarchy can be further extended with subtypes highlighting the type of the gesture signal.

While ambiguity in the type of gesture is regimented by mapping a gesture signal to more than one type in Figure 3, ambiguity in multimodal synchrony is reflected in the grammar by distinct rules constraining the permissible attachments. In this paper, we shall provide rules for integrating speech and representational co-speech gesture. The theoretical framework will be illustrated in terms of utterance (5) from Loehr's (2004) corpus.

5.1 Integration of Depicting Gesture and Prosodic Word

Our theoretical analysis begins with the straightforward case of attaching gesture to a single word.

Definition 2 (Situated Prosodic Word Constraint) *Gesture can attach to any syntactic head in the spoken utterance if 1. there is an overlap between the temporal performance of the gesture stroke and the head; 2. the head is a prosodically prominent word.*

The representation of Definition 2 in a constraint-based framework is illustrated in Figure 4. We shall now describe each aspect of this feature structure in turn.

This constraint accounts for a sign of type *depict_word* derived via unification of a single prosodic word of type *spoken_word* and a gesture of type *depicting*. As illustrated by example (3), the well-formedness constraints are guided by the relative temporal performance of both modalities: there must be a temporal overlap between the performance of the gesture phrase and the prosodic word. Otherwise, the multimodal signal is ill-formed. The temporal overlap entails the relations of inclusion, precedence and/or sequence, as detailed in § 4.2.

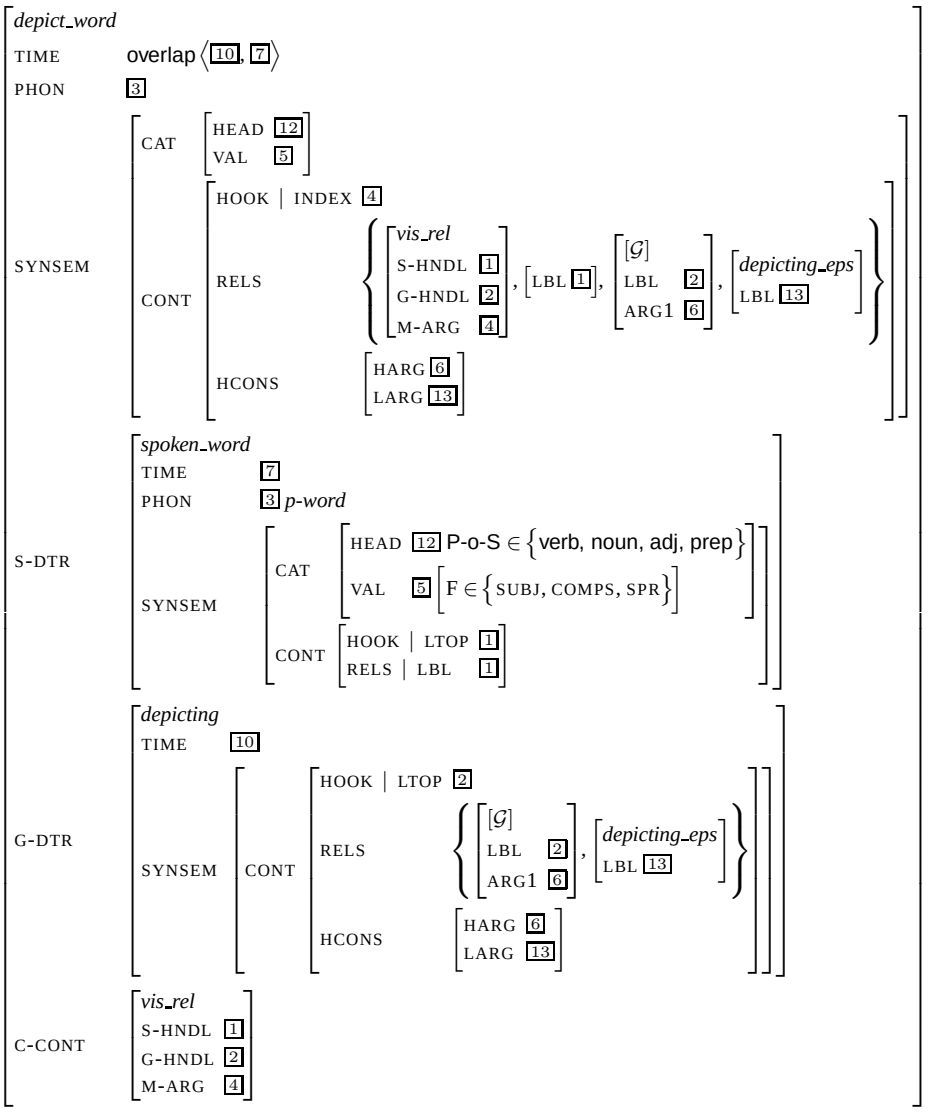

Figure 4: Situated Prosodic Word Constraint

For the gesture daughter, we record its temporal performance and its semantic contribution. The semantic components are encoded as follows: the local top is obtained via co-indexation with the label of the main predicate which is the operator $[\mathcal{G}]$. For the sake of readability, we gloss the set of elementary predications contributed by a depicting gesture as *depicting_eps*. These include every aspect of gesture meaning such as $l_1 : a_1 : hand\_shape\_open\_flat(i_1), l_2 : a_2 : palm\_orientation\_upwards(i_2)$, etc. It is vitally important to constrain these predications so that they appear within the scope of the $[\mathcal{G}]$ modality (see Lascarides and Stone (2009) for motivation): this is expressed by equating ARG1 of the operator with the label of the elementary predications within the HCONS condition.

For the speech daughter, it is equally important to record its timing, syntax-semantics information and also its prosody. The synchronicity between a depicting gesture and a lexical item necessitates the latter to be prosodically marked: we allow for the combination of a prosodically prominent word of type *p-word* and a gesture but we restrict the combination of an unstressed word "leaner" (Zwicky, 1982) of type *lnr* and a gesture. The head is not constrained to any particular category. In so doing, the gesture can be related to a verb ("MIXES mud"), a noun ("KING of Scotland"), a preposition ("THROUGH the drainpipe") or an adjective ("CLOSE to the station") as long as it is prosodically prominent. The VAL feature of the head indicates its potential to combine with other arguments. The underspecified semantic component of the speech daughter is defined in the familiar fashion in terms of its hook and relations features. The rule schema remains as unspecific as possible with respect to its EPs.

This rule contributes its own underspecified vis_rel (visualising relation) between the topmost label of the speech-daughter and the topmost label of the gesture daughter. This is specified by identifying S-HNDL of the relation with the local top label of the speech content (l_1) and G-HNDL of the relation with the local top label of the gesture content (l_2). Any relations contributed by the rule itself are specified within the C-CONT feature. The resolution of this relation is a matter of discourse which is not envisaged by this project. Based on Lascarides and Stone (2009), vis_rel is used to refer to the set of possible rhetorical relations between gesture and speech (e.g., *Narration, Depiction* or *Overlay*, but not *Contrast*).

We finally introduce an M-ARG (multimodal argument) attribute which serves as a pointer to the integrated multimodal signal and so it can be taken as an argument by any external predicate. This analysis is analogous to the treatment of conjunction in ERG where a $conjunction\_relation$ introduces an index which serves as a pointer to the conjoined entity.

The derivation of the mother node follows the algebra of Copestake, Lascarides and Flickinger (2001). It is strictly compositional: we unify the TIME, PHON and SYNSEM values of the daughters. The head feature is percolated up to the mother node and also the PHON value of the unified multimodal signal is identified with the PHON value of the speech daughter. The semantic representation involves appending the RELS and HCONS lists of S-DTR to the RELS and HCONS lists of G-DTR.

Applied to utterance (5), this constraint enables the gesture to attach to the verb "mixes": the verb is prosodically marked and the extension of its temporal performance overlaps the extension of the temporal performance of speech. In this case, vis_rel can resolve in context to a literal depiction of some mixing event. Alternatively, the gesture can also be combined with the NP "mud" which is prosodically prominent, it is a head of itself and its temporal performance overlaps the temporal performance of the gesture stroke. In this case, the verb "mix" would take two arguments: ARG1 will be identified with ARG0 of "he", and ARG2 will be identified with M-ARG of the depicting word "mud" + depicting gesture. Note that the derivation is still constrained: nothing licenses attaching the gesture to "he". Likewise, this constraint prohibits the gesture in (3) to attach to "called" or to "mother": the

former is not prosodically marked and the latter does not temporally overlap with the gestural performance.

In the next section, we shall focus on attaching gesture to a phrase larger than a single prosodic word.

5.2 Integration of Depicting Gesture and Spoken Phrase

Definition 3 (Situated Head-Argument Constraint) *Gesture can attach to the head daughter in the spoken utterance upon fully or partially saturating the head with the (externally and/or internally) selected arguments if: 1. the phrase is a prosodic constituent, 2. there is an overlap between the temporal performance of the constituent and the gesture stroke.*

We use partial of full saturation to remain neutral about the number of satisfied arguments. This is driven by the ambiguous form of the hand signal which corresponds to multiple attachment solutions. The formal rendition of this constraint is shown in Figure 5. The temporal condition, the semantic contribution of the rule, the semantics of gesture, and also the derivation of the mother node is consistent with the Situated Prosodic Word Constraint. We therefore forego any details about them.

Following the empirical analysis in § 4.2, this rule formalises synchrony beyond the strict temporal alignment of the signals. In so doing, the semantics of the head is provided with its "minimal specification" (Pustejovsky, 1995) which is necessary for resolving the incomplete meaning of gesture to one or more contextually-specific interpretations.

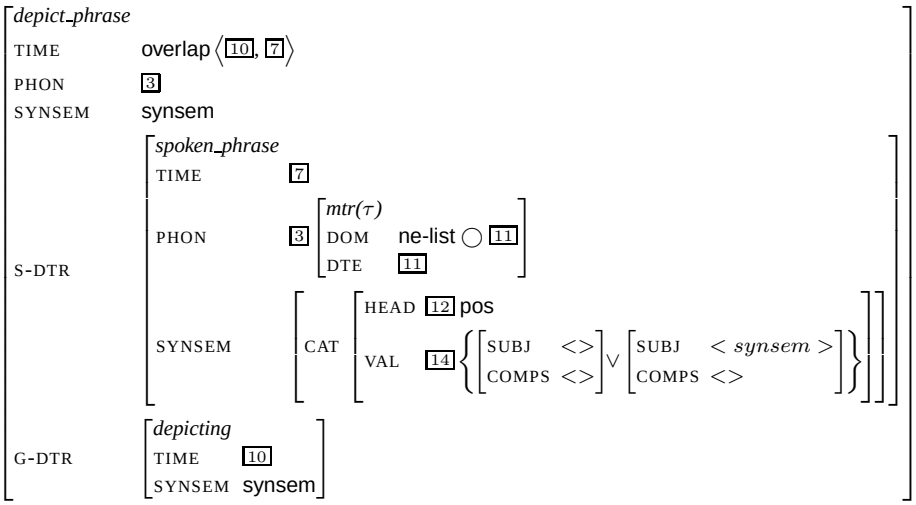

Figure 5: Situated Head-Argument Constraint

Prosody constrains the combination of both modalities: the PHON value of the speech daughter is restricted to type *mtr(τ)*—i.e., a metrical tree of any depth

(Klein, 2000). The domain union relation (◯) serves to interpolate the prosodically prominent element, the so called Designated Terminal Element (DTE), into the non-empty list of domain objects. In case of broad focus, the DTE element is in rightmost position. We make use of the disjunction operation in the SYNSEM | CAT | VAL list to remain as neutral as possible about the number of saturated arguments when the synchronisation of the gesture can take place. This constraint allows one to attach a gesture to a headed phrase whose complement requirements have been fullfilled or to a headed phrase whose both subject and complement requirements have been fullfilled.

It is important to underline the distinct status of vis_rel in the Situated Prosodic Word Constraint and in the Situated Head-Argument Constraint: whereas the former remains as vague as possible about the speech-gesture relation, the combination of the head with its arguments in the latter contributes to its minimal specification and hence the choices of resolving this relation are more constrained.

This constraint allows the G-DTR in (5) to attach to the VP "mixes mud" or to the S "he mixes mud": the temporal condition is complied; the prosodic word temporally overlapping gesture is an unsaturated syntactic head that needs to be saturated with the selected arguments: them being either "mud" only or both "mud" and "he". The inclusion of arguments into the synchronous phrase ultimately affects the gesture interpretation in context, as discussed in § 4.2.

The prosodic structure induced in parallel with the syntactic tree does not disrupt the traditional notion of syntactic constituency. Nevertheless, the syntactic structure is not necessarily isomorphic to the prosodic structure. Definition 3 constrains synchrony to a phrase where the head and the other elements are in a head-argument relation. From the perspective of an HPSG-based analysis, this involves specifying a rule so that a gesture phrase can be accommodated into a prosodic constituent that is distinct from the syntactic constituent. We therefore extend our analysis, and provide a further constraint, called Situated Prosodic Phrase Constraint (Figure 6), where the attachment is informed only by prosody, ignoring any SYNSEM values. Our motivation for this relaxation stems from the tight alignment between the speech rhythm and gesture performance: we have already observed that prosody can make embodied actions ill-formed. This constraint intergrates a gesture of type *depicting* to a metrical tree *mtr(τ)* of any depth. Similarly as before, synchrony requires temporal overlap between the gestural and the spoken modalities. The rest of the features remain the same.

The synchronisation is constrained: we unify the feature structure of both modalities making sure that the mother node inherits the semantic contribution of G-DTR. Since we have no access yet to the SYNSEM value of speech, we can only record the semantic component of gesture and add an underspecified relation vis_rel between both modalities. This relation outscopes the local top of the gesture content and the local top of the linguistic content whatever its SYNSEM is going to be.

Applying the situated prosodic phrase constraint to our working example in (5) enables the combination of the gesture and the phrase "he mixes": both modali-

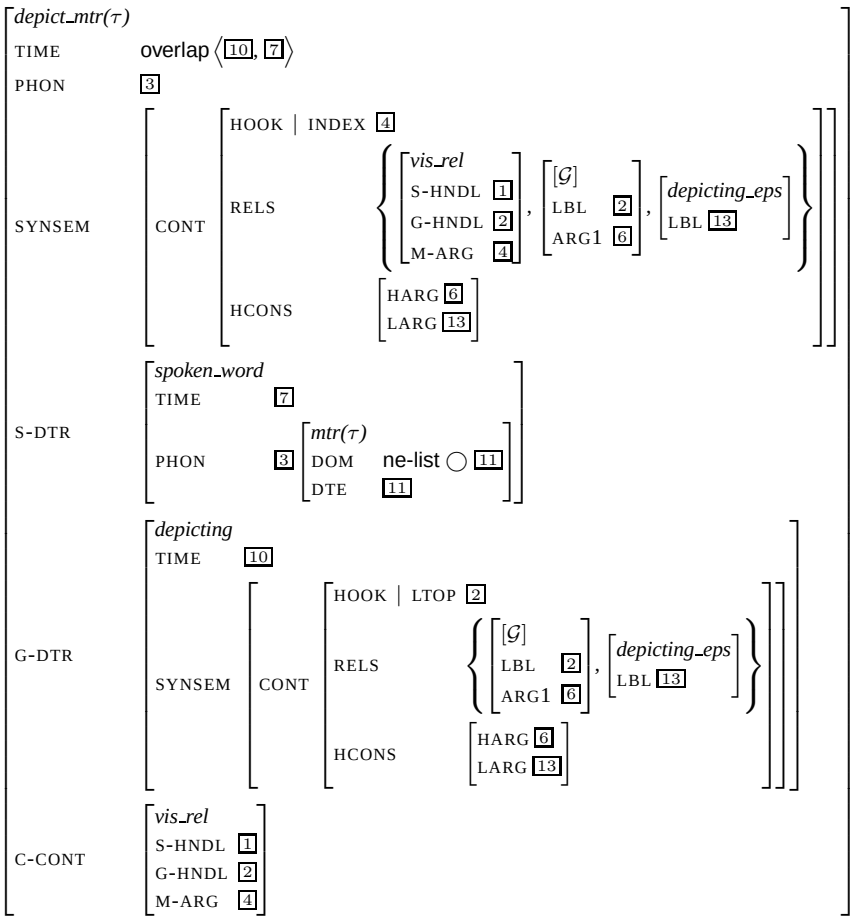

Figure 6: Situated Prosodic Phrase Constraint

ties overlap in time, and also the prosodic phrase is a metrical tree whose DTE is the prosodic word "mixes". Informally speaking, this synchronisation of modalities contributes some underspecified relation between the content of gesture and the content of speech. Whereas the gesture content is known (due to the compositional analysis), the speech content is going to be further specified once accessing the SYNSEM of the syntactic phrase. Upon that, the semantics of the depicting phrase will be able to incorporate the relevant elementary predications coming from the speech daughter: in this case, they will be roughly equivalent to: $l_1 : pron(x_4)$; $l_2 : pronoun\_q(x_4)$ $l_2 : RESTR(h_6)$ $l_2 : BODY(h_7)$; $l_3 : mix(e_1)$ $l_3 : ARG1(x_4)$ $l_3 : ARG2(x_9)$ and $h_6 =_q l_1$.

This rule is needed because it balances between syntactic constituency and prosodic constituency. Nonetheless, its specification would not be necessary in other formalisms that have isomorphic prosodic, syntactic and semantic structures (Steedman, 2000).

6 Conclusions

In this paper, we demonstrated that current methods for semantic composition can be extended to multimodal language so as to produce an integrated meaning representation based on the form of the spoken signal, the form of the co-speech gesture, and their relative timing. We also saw that the ambiguous gesture form provides one-to-many form-meaning mappings without violating coherence in the final interpretation.

The integration of speech and gesture into a single derivation tree is informed by linguistic criteria (prosody and syntax) and non-linguistic criteria (temporal relation between speech and gesture), and it produces a highly underspecified logical form that will be resolved to preferred values in specific context. Our generic rules—the Situated Prosodic Word Constraint and the Situated Head-Argument Constraint—provided the methodology for producing an integrated tree where on one hand, syntax permits multiple attachments which subsequently produce underspecified relations, and on the other, prosody constrains the well-formedness of the embodied act. Moreover, the Situated Prosodic Phrase Constraint illustrates that gestures can be elegantly integrated into a prosodic constituent, and so this rule demonstrates that isomorphism between prosodic and syntactic structure is not necessary for the derivation of the multimodal signal.

In future, we intend to extend those rules with analysis of deictic gestures where sequentiality of the performance of spoken and the gestural signal is common. We also hope to implement the theoretical findings into a computational multimodal grammar for English (Bender et al., 2002).

References

Bavelas, Janet B., Chovil, Nicole, Coates, Linda and Roe, Lori. 1995. Gestures specialized for dialogue. In *Personality and Social Psychology Bulletin*, volume 21, pages 394–405.

Bender, Emily M., Flickinger, Dan and Oepen, Stephan. 2002. The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars. In John Carroll, Nelleke Oostdijk and Richard Sutcliffe (eds.), *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan.

Buisine, Stéphanie and Martin, Jean-Claude. 2007. The effects of speech-gesture cooperation in animated agents' behavior in multimedia presentations. *Interact. Comput.* 19(4), 484–493.

Cassell, Justine. 2000. Nudge nudge wink wink: elements of face-to-face conversation for embodied conversational agents. *Embodied conversational agents* pages 1–27.

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

Copestake, Ann, Flickinger, Dan, Sag, Ivan and Pollard, Carl. 2005. Minimal Recursion Semantics: An introduction. *Journal of Research on Language and Computation* 3(2–3), 281–332.

Copestake, Ann, Lascarides, Alex and Flickinger, Dan. 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)*, pages 132–139, Toulouse.

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

Fricke, Ellen. 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.).

Giorgolo, Gianluca and Verstraten, Frans. 2008. Perception of speech-and-gesture integration. In *Proceedings of the International Conference on Auditory-Visual Speech Processing 2008*, pages 31–36.

Haji-Abdolhosseini, Mohammad. 2003. A Constraint-Based Approach to Information Structure and Prosody Correspondence. In Stefan Müller (ed.), *Proceedings of the HPSG-2003 Conference, Michigan State University, East Lansing*, pages 143–162, Publications, `http://cslipublications.stanford.edu/HPSG/4/`.

Johnston, Michael. 1998. Multimodal Language Processing. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*.

Kendon, Adam. 1972. Some relationships between body motion and speech. In A. Seigman and B. Pope (eds.), *Studies in Dyadic Communication*, pages 177–216, Elmsford, New York: Pergamon Press.

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

Klein, Ewan. 2000. Prosodic Constituency in HPSG. In *Grammatical Interfaces in HPSG, Studies in Constraint-Based Lexicalism*, pages 169–200, CSLI Publications.

Kopp, Stefan, Tepper, Paul and Cassell, Justine. 2004. Towards integrated microplanning of language and iconic gesture for multimodal output. In *ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces*, pages 97–104, State College, PA, USA, New York, NY, USA: ACM.

Lakoff, George and Johnson, Mark. 1980. *Metaphors We Live By*. Chicago and London: The University of Chicago Press.

Lascarides, Alex and Stone, Matthew. 2009. A Formal Semantic Analysis of Gesture. *Journal of Semantics* .

Loehr, Daniel. 2004. *Gesture and Intonation*. Washington DC: Georgetown University, doctoral Dissertation.

McClave, Evelyn. 1991. *Intonation and Gesture*. Washington DC: Georgetown University, doctoral Dissertation.

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

Oviatt, Sharon L., DeAngeli, Antonella and Kuhn, Karen. 1997. Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction. *CHI* pages 415–422.

Pollard, Carl and Sag, Ivan A. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press and Stanford: CSLI Publications.

Pustejovsky, James. 1995. *The Generative Lexicon*. MIT Press, Cambridge.

Steedman, Mark. 2000. *The Syntactic Process*. The MIT Press.

Zwicky, Arnold. 1982. Stranded *to* and phonological phrasing in English. *Linguistics* 20(1/2), 3–57.