
THE UNIVERSITY *of* EDINBURGH

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

*Katya Alahverdzhieva*


Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2013

# Abstract

This thesis concerns the form-meaning mapping of multimodal communicative actions consisting of speech signals and improvised co-speech gestures, produced spontaneously with the hand. The interaction between speech and speech-accompanying gestures has been standardly addressed from a cognitive perspective to establish the underlying cognitive mechanisms for the synchronous speech and gesture production, and also from a computational perspective to build computer systems that communicate through multiple modalities.

Based on the findings of this previous research, we advance a new theory in which the mapping from the form of the combined speech-and-gesture signal to its meaning is analysed in a *constraint-based multimodal grammar*. We propose several construction rules about multimodal well-formedness that we motivate empirically from an extensive and detailed corpus study. In particular, the construction rules use the prosody, syntax and semantics of speech, the form and meaning of the gesture signal, as well as the temporal performance of the speech relative to the temporal performance of the gesture to constrain the derivation of a single multimodal syntax tree which in turn determines a meaning representation via standard mechanisms for semantic composition. Gestural form often underspecifies its meaning, and so the output of our grammar is *underspecified logical formulae* that support the range of possible interpretations of the multimodal act in its final context-of-use, given the current models of the semantics/pragmatics interface.

It is standardly held in the gesture community that the co-expressivity of speech and gesture is determined on the basis of their temporal co-occurrence: that is, a gesture signal is semantically related to the speech signal that happened at the same time as the gesture. Whereas this is usually taken for granted, we propose a methodology of establishing in a systematic and domain-independent way which spoken element(s) gesture can be semantically related to, based on their form, so as to yield a meaning representation that supports the intended interpretation(s) in context. The 'semantic' alignment of speech and gesture is thus driven not from the temporal co-occurrence alone, but also from the *linguistic properties* of the speech signal gesture overlaps with. In so doing, we contribute a fine-grained system for articulating the form-meaning mapping of multimodal actions that uses standard methods from linguistics.

We show that just as language exhibits ambiguity in both form and meaning, so do multimodal actions: for instance, the integration of gesture is not restricted to a unique speech phrase but rather speech and gesture can be aligned in multiple multimodal

syntax trees thus yielding distinct meaning representations. These multiple mappings stem from the fact that the meaning as derived from gesture form is highly incomplete even in context. An overall challenge is thus to account for the range of possible interpretations of the multimodal action in context using standard methods from linguistics for syntactic derivation and semantic composition.

# Acknowledgements

Only with the help of many people was I able to complete this journey. First and foremost, I would like to thank my supervisor, Professor Alex Lascarides, who mentored me on this rather unconventional topic, guided me and supported me throughout the years. For her patience, brilliant ideas and immediate response. I feel lucky, grateful and privileged for having worked with Alex.

I am deeply grateful to Daniel Loehr who provided me with his collection of annotated video recordings. To Michael Kipp who kindly provided me with the labelling tool Anvil and offered me technical assistance whenever necessary. I would also like to thank the participants who did the gesture annotation in my experiment. I gratefully acknowledge Mark Steedman and Ewan Klein who gave me extremely helpful comments and ideas that I used in this thesis. Also big thanks to Jean Carletta and Jonathan Kilgour who helped me a lot with the NXT tool. I also benefited a lot from discussions over mail with Sasha Calhoun. In 2010, I participated in the first school in Gesture Studies, where I was lucky to meet and talk about my work with Adam Kendon, Mandana Seyfeddinipur, and the summer school participants.

I gratefully acknowledge Dan Flickinger who made the initially impossible task of multimodal grammar implementation possible. It is thanks to him that the implemented grammar came to life. Big thanks to Emily Bender for her help with the grammar engineering challenge. To the people from the DELPH-IN community—Ulrich Schäfer, Peter Adolphs, Berthold Crysmann—who gave me useful instructions how to handle the grammar engineering platforms. And Stephan Oepen who helped me with the grammar profiling system. I would also like to express my gratitude to Nicholas Asher who offered me great help with SDRT. I would also like to thank the anonymous reviewers of my submissions for the insightful comments. And my office mates who contributed the nice, friendly and quiet environment in our office: Yansong, Jeff, Neil, Sharon, Ioannis, Sean. All images should be attributed to Tudor Thomas, who did an excellent work in turning my screenshots in these lively drawings.

I am deeply grateful to my examiners, Professor Bob Ladd and Dr. Michael Johnston, for their insightful comments and criticism, and also for turning my viva into a very relaxed and enjoyable experience.

Many thanks are due to EPSRC who provided the funding for my PhD studies, and also to the JAMES project that funded two of my conferences.

And of course, to my mum who always encouraged me in my pursuits and never questioned my choice to live far from home, to my sister who showed me that PhDs

are accomplishable. And to Hervé Saint-Amand, for being next to me and for his unconditional love and support.

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Katya Alahverdzhieva*)

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Through the physical co-location of people known as *co-presence* [Goffman, 1963], people exchange information with each other by a range of meaningful and visibly accessible communication channels. In Figure 1,[1] for instance, the arrangements of the people's bodies in the shared space, the directions of their faces, and their hand movements convey that the participants are engaged in a conversation and that the person on the right-hand side is probably holding the floor, whereas the participant on the left-hand side is listening.

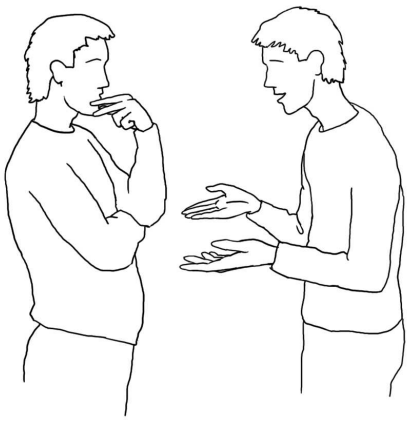


Figure 1: Conversation

In face-to-face interaction, people tend to have the same judgements as to which of the speech-accompanying behaviours are communicative and thus relevant for the topic of conversation, and which are not [Kendon, 2004]. To illustrate this, let's con-

[1] The photo is due to Tudor Thomas.

sider again Figure 1: imagine that the participant on the right was offering something (literally or figuratively) to his interlocutor, then the placement and the shape of his hands would be perceived as deliberately communicative and it would thus be part of the speaker's contribution to the discourse.

In utterance production and in utterance perception, people make use of 'visible bodily actions' [Kendon, 2004] as a source for providing and perceiving discourse-related information. Face-to-face dialogue is thus an *embodied* process of information exchange in that people deploy bodily behaviours in the communicative act; for instance, in verbal route directions, people frequently use their hands to navigate in the surrounding space; when narrating stories people rely on hand movements to depict events or to provide visual characteristics of an object. Face-to-face dialogue is also *situated* in the context in which the conversation takes place, making its meaning dependent on its relation to the world in the specific time and space of the communicative act; for instance, the identification of an addressee in a multi-party conversation may depend entirely on the orientation of the speaker's body in the context in which the conversation takes place.

## 1.1 What This Thesis is About

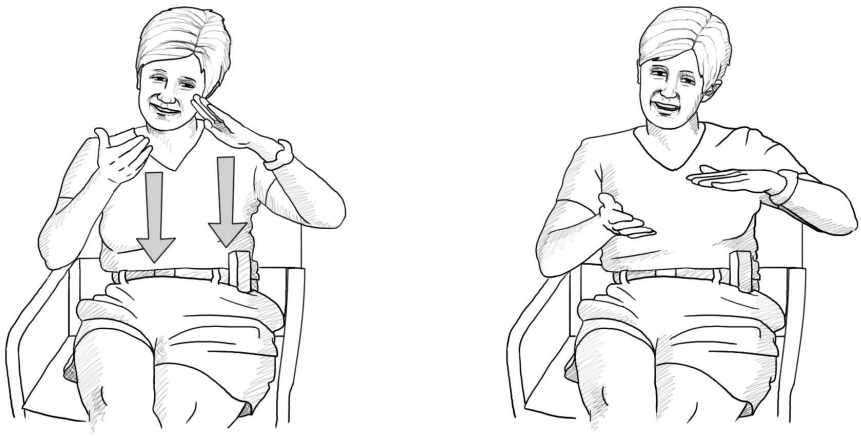
To illustrate the focus of this thesis, let's consider utterance (1.1)[2] [Loehr, 2004] extracted from a longer conversation where the speaker was describing how she would stack books to maintain a level position of her crooked cupboards.[3]

(1.1) ... and I have books that I [$X^*$stack] together [laughter] to hold it ...

Prior to "stack", hands are brought together with extended palms to the frontal space, adopting the shape of a flat horizontal container. Then along with "stack" they perform a brisk downward movement that resembles manipulating flat objects, as illustrated in Figure 2. Both the speech and the hand movement are relevant for the

[2]For the utterance transcription, we have adopted the following convention: the speech signal synchronous with the expressive phase of the gesture, the so called stroke, is underlined, and the signal synchronous with the hold phase after the stroke is underlined with a curved line. Here we have also included those words that start/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) and/or the tune: H(igh), L(ow) or X(underspecified). More details about the annotation are given in Section 4.1.

[3]For this example and for all subsequent examples that are cited as Loehr [2004], we are grateful to Daniel Loehr who kindly provided us with an annotated multimodal corpus. We used this corpus to study depicting gestures.



(a) and I have books that I stack... (b) ... together to hold it

Figure 2: Gesture Depicting Stacking Books, example (1.1) [Loehr, 2004]

conveyed meaning of stacking books, and both are produced and perceived as a coherent idea unit.

This thesis is concerned with signals like the one in (1.1) that use the hand as a semantically intended medium for communication in tandem with speech. In the literature, these hand signals are known as *co-speech gesture, co-verbal gesture* or *gesticulation,* and the whole speech-and-gesture ensemble is referred to as a *multimodal action* or an *embodied action.*

### 1.1.1 Main Properties of Co-speech Gesture

In contrast to conventionalised gestures where meaning directly follows from form—e.g., the thumbs up, the OK sign or the shush movement—co-speech gestures are spontaneous and improvised, and so they have no predefined and well-established form from which a unique meaning can be derived. It has been observed in the gesture community that the co-speech hand signal is massively ambiguous, potentially mapping to open-ended meanings (e.g., Lascarides and Stone [2009b]). A right-handed circular motion performed by the wrist can literally denote a circular movement such as the turning of a wheel, it could denote the object being rotated such as the wheel itself, it could metaphorically refer to an iterative process, and even each iteration can designate the distinct steps in an iterative process. It is only in context, in particular in the context

of speech, that the co-speech gesture contributes a specific meaning: the content and the corresponding meaning representation delivered by the rotating movement in the context of "He mixes mud" is distinct from that in the context of "It's a huge, long boring process". Whereas in the former case the hand can literally denote the mixing event or even the mud being mixed, in the latter context the hand can be a metaphorical denotation of a tedious and repetitive event. Likewise, some of the possible denotations of the gesture in (1.1) without the speech context include: a flat container, a flat object being contained, or even the size of an object measured from the top palm to the bottom palm. Of course, this is just an incomplete list of this gesture's possible denotations, given just its form.

The ambiguity notwithstanding, gesture and its form abstracted away from the context-of-use does convey some meaning no matter how incomplete this meaning might be. This suggests that the choice of which speech element(s) gesture can be linked to is *not* a free-for-all but there are certain constraints that govern this integration (cf. Giorgolo and Verstraten [2008], Harrison [2010]). More specifically, we have observed that the speech-and-gesture integration can be anomalous due to at least the following three factors:

1. First, the anomaly can be due to form—here we include the temporal performance of gesture in relation to the temporal performance of the prosodic prominence in speech (we discuss this in Section 1.3.2.1).

2. Second, it can be due to meaning—if we consider again the rotating movement example, the circular hand motion can never denote a rectangular concept, and hence it cannot be linked to a speech phrase of a rectangular denotation (we illustrate this in Section 1.3.2.2).

3. And finally, it can be due to the discourse context—for instance, identifying the denotation of a deictic gesture is a matter of salience in the communicative situation, and a failure to identify the gesture's referent would render the multimodal action anomalous (we illustrate this in Section 4.3).

This observation flags up an important claim in this thesis, namely that gestural form is *not* semantically vacuous, and that it conveys some meaning (no matter how underspecified) that can be resolved to a specific interpretation in the speech context. To validate this claim, we performed an experimental study where we asked two participants (one native speaker of English, one native speaker of German) to annotate

| | # overlaps | Total gestures | Percent |
|---|---|---|---|
| *Overlapping gesture meanings* | 41 | 54 | 75.926% |

Table 1: Inter-annotator Agreement on Meanings Assigned to Gestures

the hand movements in a video recording. The annotators were instructed to identify the communicative signals, to mark the gesture boundaries, to assign them a dimension from a list of possibilities (these included iconic, metaphoric, deictic gestures and beats),[4] and to provide them with a brief description of their meaning in context. The coding was performed with the sound turned on.[5] We shall postpone the discussion of the complete results until Chapter 4. Now it suffices to say that we established 75.926% agreement on the assigned gesture meanings as displayed in Table 1. Due to the nature of the work, the inter-annotator agreement was measured manually.

The results attest that human judgements about which movements are communicative gestures and what they mean in context are relatively robust, suggesting that recognising and interpreting gestures should be systematic. What we counted as disagreement with respect to the assigned meaning was, for instance, when one annotator interpreted the gesture in (1.1) as depicting "the action stacking books and boxes", and the other annotator described the same movement as depicting "the size and shape of the boxes referred to in the speech". Those findings are consistent with our previous claim that despite the ambiguity, gestural form imposes abstract constraints on what gestures mean in context. There is no contradiction between the two as we intend to use standard methods for semantic underspecification to produce a meaning representation that captures the incomplete meaning of gesture as mapped from its form, and that also supports the exact range of possible interpretations in context.

To better understand gesture form and meaning, we can draw an analogy with the role of intonation in conversation [Ladd, 2008]: similarly to intonation, gestures are to an extent idiosyncratic in that their form is designed on the spot and so it varies from individual to individual, or from one context to another. At the same time, there are prevalent features and properties of gesture that cut across speakers, and that humans perceive in a similar way. A possible example is the use of pointing when navigating on a virtual map: we say that gesture form is idiosyncratic in that some speakers might

[4] An in-depth introduction into gestures, their dimensionality and internal organisation is given in Chapter 2.

[5] For the annotation instructions, please refer to Appendix A.

use an extended index finger while others might use a vertically extended open palm. Essentially, both gesture forms would be produced in context to identify a certain landmark and both would be recognised as such.

A key feature of gestures is that the different ways they are understood depend on the content of the speech that is uttered when the gesture is performed.[6] Gestures function in tandem with the co-occurring speech within a single communication system: a "single thought" is expressed synchronously in two modes and is perceived as an integrated multimodal ensemble [McNeill, 1992; McNeill, 2005]. The integrated nature of the multimodal utterance is observed when the semantic relation between speech and gesture is one of redundancy (the gesture signal "repeats" visually the spoken words without contributing distinct content; for instance, uttering "I had one coffee this morning" while extending vertically an index finger along with "one coffee") or when the relation is one of complementarity (the gesture adds propositional content to the final utterance, or it qualifies the speech act being performed; for instance, the gesture in (1.1) contributes content to the whole multimodal action by displaying the horizontal orientation of the stacked books). Whereas redundancy violates Grice's Maxim of Quantity for co-operativity in conversation [Grice, 1975], speech-gesture redundancy does not violate coherence [Lascarides and Stone, 2009b], and it can facilitate learning and enhance expressiveness [Buisine and Martin, 2007]. Note that even when speech and gesture convey the same content, they may not be redundant for parsing, with the gesture serving to disambiguate the speech and/or vice versa.[7]

### 1.1.2 Thesis Aims

With this in mind, there does not yet exist a consistently formalised model of multimodal actions that is predictive about speech-gesture anomalies of the type discussed in the previous section, that combines speech and gesture in a constraint-based way so as to account for the anomalies coming from form, that interfaces their underspecified meaning derived from the ambiguous form with any existing pragmatic theory, and that is domain-independent. The original contribution of this thesis is to fill this gap: we are going to demonstrate that this can be achieved by re-using the research methodology

[6] This is the reason why the annotators were instructed to annotate the gestures with the sound switched on. This is in contrast with other annotation practices that initially code gestures with the sound switched off (for instance, Mandana Seyfeddinipur (personal communication), Bressem [2008], Harrison [2010]).

[7] Experimental studies have shown that speakers rely on gestures for disambiguating lexical ambiguities, see for instance Holler and Beattie [2007].

applied to the development of wide-coverage grammars. The overall aim of this thesis is thus to articulate the form-meaning mapping of multimodal communicative actions consisting of speech and co-speech gesture. We view form as realised on the following levels: first, the form of the spoken signal refers to its syntactic and prosodic properties; second, form designates the gestural category—depicting or pointing—which ultimately has effects on which linguistic element the gesture could align with and hence how the multimodal action would be interpreted in context (see Section 2.2.1.2). In the previous section, we claimed that the form of the gesture is not semantically vacuous and that it contributes some meaning within and outwith context. So the third level of form concerns the physical shape of the hands and their movement while executing the gesture. Further in Section 1.1.1 we propose a refinement of this definition, and in Section 5.2.1 we argue for a particular way of formally rendering the various aspects of gesture form. We also stated that the anomaly of the speech-and-gesture integration is a matter of form where form pertains to the temporal performance of the speech signal relative to the temporal performance of the hand signal. So the relative timing of the speech and gesture signal is yet another aspect of form that we shall account for. Mapping form to meaning includes all these realisations of multimodal form (including the refinement that we propose in Section 2.2.1.3 and its formal rendition in Section 5.2). We intend to use standard methods from linguistic theory such as constraint-based syntactic derivation and semantic composition to map multimodal form to meaning, yielding an underspecified logical form that will support pragmatic inference in any domain.

Further, this thesis contributes to the existing formal models of multimodal integration (detailed in Section 3.5) in that the speech-and-gesture integration takes place not at the level of the spoken utterance but at a sub-utterance level, i.e., our model involves combining gesture with smaller *syntactic constituents*. For instance, while Johnston and Bangalore [2000] used multimodal terminal symbols—that is, a terminal symbol is a triple composed of representations for the spoken input stream, the gesture input stream and their combined meaning—we achieve speech-and-gesture integration by syntactic adjunction in the grammar. In so doing, we contribute a formal model of multimodal integration which is sensitive to how form influences multimodal alignment, without breaking the constituent structure of the input speech elements (cf. Paggio and Navarretta [2009]).

Based on the empirical finding that the interaction between speech and gesture is on the level of *form* and assuming that any information about form is a matter of a

grammar (as detailed in Section 1.3), we analyse the form-meaning mapping in terms of a constraint-based grammar for multimodal language that takes verbal signals and hand gestures as input and that defines rules for the alignment of speech and gesture in a *single derivation tree*. The grammar captures generalisations about the well-formedness of the multimodal action, about multimodal ill-formedness, and also about multimodal signals that, although grammatically well-formed, can never produce the intended meaning in the communicative act. The multimodal grammar intends to account for the (underspecified) meaning representations that are derivable from form and that are compatible with the plausible interpretations in any context in which that act may be performed. The outcome of this study shall be a deeper understanding of the link between syntax and semantics of language, and also a grammar framework for the domain-independent development of wide-coverage multimodal grammars.

### 1.1.3 Steps to Achieve Our Aims

The different steps of how we intend to achieve the aims presented in Section 1.1.2 can be summarised as follows:

1. To extract generalisations from multimodal corpora about the syntactic and semantic well-formedness of multimodal signals composed of speech and gesture.

2. Using the extracted generalisations, to provide a precise grammar theory of multimodal signals that models the form of the speech, the form of the gesture and the form of their combination, producing multimodal logical forms as mapped from the multimodal form using standard methods of composition from linguistics. The theory should be able to scale up to any grammar formalism.

3. To formalise the grammar theory into a constraint-based grammar formalism.

4. To implement the theoretical constraints by extending a wide-coverage computational grammar for English with rules for speech and gesture integration.

5. To evaluate the grammar coverage using a manually crafted test suite, in analogy to the tradition of evaluation of phenomenon-based linguistic grammars.

## 1.2 What This Thesis in Not About

To avoid any potential confusion, we would like to mention a few directions of research that are outside the scope of this thesis.

This thesis is concerned with the syntax and semantics of multimodal actions composed of speech and co-speech gesture. Although we take care to produce meaning representations that are compatible with the assumptions made by the current formal models of the semantics/pragmatics interface (discussed in Section 1.3.1), any formal modelling that involves inference mechanisms for the final interpretation in context and/or discourse coherence is outwith the scope of this project.

Our language of study is English, and also the gestures of study are produced by English speakers. Our corpus investigation, as well as our grammar construction is carried out for English language. Any cross-linguistic analysis remains extraneous to our aims.

The grammar implementation is performed by using existing grammar engineering platforms for implementing large-scale language grammars. We do not intend to modify the existing platforms and to adapt them for multimodal input. (As we shall see in Chapter 7, the current machinery presents a number of challenges for the multimodal grammar implementation.)

This thesis deals with gestures in a discrete formal representation, namely typed feature structures. The recognition of the visual signal and the mapping from the visual input to a symbolic representation is outwith the scope of our work.

## 1.3 Why a Multimodal Grammar?

We claim that speech and gesture combine *within the grammar*, producing a single syntactic tree that maps to a unified meaning representation. Our motivation for a grammar-based approach stems from our underlying assumptions about the particular model of the semantics/pragmatics interface which has access to the compositional semantics of the elementary units but *not* to their form. However, a wide range of studies on multimodal communication demonstrate that the interaction between speech and gesture is, importantly, on the level of form—that is, the interpretation of a gesture is constrained by the linguistic structure (cf. Harrison [2010], Giorgolo and Asudeh [2011]), including its prosody. Linguistic form is also shown to have effects on human judgements concerning multimodal grammaticality [Giorgolo and Verstraten, 2008].

Further to this, grammars capture linguistic generalisations on form and meaning that correspond to human judgements of linguistic competence and the rules that govern the language production. Cognitive models of language production suggest that the multimodal utterance is a product of the constant interaction on a conceptual level between linguistic and visuo-spatial information [Alibali, Kita, and Young, 2000; Kita and Özyürek, 2003; Hostetter, Alibali, and Kita, 2007]. We address these models from a computational perspective by exploring the plausible interactions between linguistic structures and gestures, thereby enriching our view of grammars with co-verbal objects.

This section is structured as follows: in Section 1.3.1 we introduce our main assumptions about the semantics/pragmatics interface. Then in Section 1.3.2 we provide some constructed and empirically extracted multimodal utterances that provide empirical evidence for capturing the speech and gesture interaction in the grammar. We conclude with Section 1.3.3 where we search for a motivation from a cognitive perspective by discussing some studies about the mental processes involved in multimodal production.

### 1.3.1 Assumptions About the Model of Pragmatic Theory

We assume that the pragmatic processing of multimodal actions is analogous to the pragmatic processing of discourse units, so as to make our formal model of multimodal signals be supported by well-established pragmatic theories for unimodal communication. In particular, 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 the coherence-based theory is that discourse is hierarchically organised into a structure of elementary discourse units bound by coherence relations—for instance, Elaboration, Explanation, Contrast, Contiguity and Cause-Effect. The complete list of available relations is always finite, and it is specific to the pragmatic theory. We shall not discuss them in detail, as the values of these relations are not a matter of the current research. The coherence relations are inferred on the basis of: (i) the semantic content of the discourse units; and (ii) information about the context in which the utterance takes place, including real world knowledge and the mental state of the participants. Essentially, the construction of a pragmatically preferred logical form and the inference of a coherence relation has access only to the compositional semantics of the elementary units and *not* to their form. We assume that

the form-meaning mapping happens in a separate module, which outputs an abstract and partial meaning representation. This abstract representation is augmented with contextual knowledge and complex commonsense reasoning—which is captured via a logic of defeasible reasoning (e.g., Hobbs, Stickel, and Martin [1993], Asher and Lascarides [2003]) or probabilistic inference (e.g., Marcu and Echihabi [2002])—which then serves to resolve that abstract and partial meaning representation to a fully specific interpretation.

Traditionally, the elementary units of discourse are clauses. Following Lascarides and Stone [2009a], we assume that gestures are also elementary units, and so the interpretation of a multimodal action involves inference about the relation between speech units and gesture units, and also between gesture units and other gesture units. By making the gesture an elementary discourse unit, we treat it as a proposition (and not just an attribute of the linguistic action), which allows for exploring its rich contribution to coherent conversation (see Kendon [2004] and Lascarides and Stone [2009b] for detailed motivation for gestures expressing propositions).

We further assume that the choice of what speech content the gesture relates to determines the events and objects that can be used as antecedents for resolving the semantic values for objects and relations contributed by the gesture. In line with theories of dynamic semantics and discourse interpretation [Hobbs, 1985; Kehler, 2002; Asher and Lascarides, 2003], we assume that there are constraints on the availability of the discourse parts that can serve as antecedents for resolving the semantic values of a discourse unit. For linguistic discourse, antecedents for anaphora can be within the same discourse unit or in the coherently related discourse unit. Following Lascarides and Stone [2009b], we carry over these constraints into embodied actions—in other words, any semantic element that acts as an antecedent for resolving the underspecified meaning of a gesture to a specific value must be a part of a discourse unit to which the gesture is connected with a coherence relation. In addition, Lascarides and Stone [2009b] observe additional constraints on antecedents for resolving gesture interpretation; constraints that we assume here, namely: the antecedent for resolving gesture can be introduced by a gesture or a linguistic discourse unit, but antecedents for resolving linguistic anaphora must be linguistic. This captures the fact that it seems rather unnatural for referents introduced only in gesture to serve as antecedents for anaphoric expressions in speech such as pronouns. Lascarides and Stone [2009b] exemplify this using Kendon's [2004] Christmas cake narrative. Utterances (1.2) and (1.3), Figure 3[8]

[8] In this instance, we have no information about the gesture phases and the pitch accents.

(a) and it was [pause 1.02 sec] this sort of [pause 0.4 sec] size…

(b) …and he'd cut it off in bits

Figure 3: Fragment of Kendon's [2004] Christmas Cake Narrative, examples (1.2) and (1.3)

are fragments from a longer narrative where the speaker described how his father, a grocer, would sell slices from a large cake at Christmas time.

(1.2) …and it was [pause 1.02 sec] this sort of [pause 0.4 sec] size
*Along with the speech, the speaker uses his index fingers to outline a large square object in the frontal space.*

(1.3) and he'd cut it off in bits
*Using the same frontal space, the speaker performs a cutting action with his right hand, with an open palm and vertically placed.*

Lascarides and Stone [2009b] argue that the hypothetical situation of continuing the narrative with "and it would get frosting all over it" where the spoken "it" would refer to one of the cut-off cake slices introduced by the gesture in (1.3) would be quite unnatural. In Section 5.2.2.4, we argue for a particular way how to represent this formally in the grammar. Also, in Section 5.2.2.4, we provide an example demonstrating that this constraint varies across gesture types.

A full pragmatic model of multimodal communication must therefore identify which speech signals gesture aligns with. However, form—that is, the prosody and syntax of the linguistic unit and the relative timings—imposes constraints on the possible speech-and-gesture alignment configurations. This in turn raises the notion of multimodal grammaticality. In the next section, we tackle this observation by introducing empirical data.

## 1.3.2 Empirical Evidence

### 1.3.2.1 Speech-Gesture Alignment and Prosody

We begin with the constructed example in (1.4), which reflects the intuitions of native speakers about multimodal grammaticality.

(1.4) * 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" in a sentence of a single intonation phrase even though the gesture is intended as a depiction of something related to the act of calling. This anomaly would not arise if the gesture happened along the whole utterance (or a part of it) which, importantly, includes the prosodically prominent element "mother": for instance, "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 — they identify a linguistic expression as distinct from the current context. Since an 'out-of-the-blue' or an all-rheme utterance advances the discussion by contributing contrastive effects in the hearer's model (not necessarily novel information), we expect that a gesture aligns with the contrastive component(s) signalled by prosodic prominence (in (1.4), the contrast takes scope over "mother" and any of its higher projections). This explains ill-formedness only with the additional assumption that the gesture must overlap with the pitch accent of the corresponding speech phrase. This intuitive judgement is in line with the empirical findings of Giorgolo and Verstraten [2008] who isolated prosody as the parameter that influences the perception of multimodal well-formedness vs. multimodal ill-formedness.

Considering that form (here, prosody) plays a role with respect to what part of the speech signal a gesture could possibly align with, we can extend the notion of speech-and-gesture alignment to cover grammaticality. We define the notion of grammaticality in terms of the placement of gesture relative to speech: ungrammatical (and hence misaligned) multimodal actions comprise those actions where the timing of gesture relative to the timing of speech does not conform to the rules by which speech and gesture are combined in everyday language use, and which native speakers perceive as ill-formed or misaligned. To account for these ungrammaticalities, we shall advance a grammar theory that is predictive about multimodal grammaticality. The ungrammaticality of purely linguistic units such as "dog the barked" is traditionally captured by systematic

Figure 4: Gesture Depicting “greasy”, example (1.5) [Kendon, 2004]

grammar rules for constituency order, agreement, etc. Based on the native speakers’ intuitions, our conjecture for multimodal grammaticality is that a gesture cannot be combined with an unaccented item in an all-rheme utterance. We leave the results of whether this conjecture is empirically validated and whether it can be applied to deictic gestures to Chapter 4. Similarly to purely linguistic grammars, we believe that the notion of multimodal grammaticality is not a binary classification of fully acceptable vs. unacceptable structures but there are speech-and-gesture combinations that would receive gradient judgements of their grammaticality.

#### 1.3.2.2 Speech-Gesture Alignment and Syntax/Semantics

To illustrate how syntax influences the decisions of which speech phrase gesture aligns to with respect to the derived meaning, consider utterance (1.5) where moving the gesture to a different speech element would result in a different gesture meaning or even in incoherence. This example is taken from Kendon [2004, p. 129] where the speaker discussed how an old sausage and pastry factory was taken over by new owners. When transferred to the new owners, the factory was rather filthy which the speaker described as “they made everything greasy”. Along with “greasy…”, the speaker’s hands spread out to the left and right periphery as illustrated in Figure 4 so as to designate some spatial extent, some closed area being made greasy [Kendon, 2004].

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

We view the semantic effects of speech-and-gesture alignment as arising from the (underspecified) semantic content of gesture which is directly bound to the co-temporal "greasy" or any of its higher projections. The interpretation would not change much if the gesture onset was moved a few milliseconds earlier so that it happened along "make everything greasy" or if it was held further so as to span "make everything greasy in the whole room". However, this gesture would not produce the intended meaning if it was performed earlier so that it coincided exactly with the time the subject noun phrase "they" was uttered, with no temporal overlap with any part of the verb phrase. Given the particular model of the semantics/pragmatics interface (Section 1.3.1), the gesture temporally overlapping only "they" could not convey anything related to greasiness. This means that it cannot be coherently related to the denotation of the verb phrase, and hence it cannot align with it either (given the rough definition of alignment from Section 1.1.2). Alternatively, the gesture temporally overlapping "made everything greasy" can convey the extent of greasiness, and so it can be coherently related to the content of this VP, including "greasy", and hence in syntax it can align with it.

Based on that, we further extend the notion of alignment to include the syntax and semantics of the multimodal action. Therefore, within the grammar we shall model speech-and-gesture alignment not only in terms of timing but also in terms of purely linguistic factors such as the syntax of the utterance — here, the temporal co-occurrence with an argument or any of its syntactic projections affects what the gesture can mean. In Chapter 5, we propose grammar rules which reflect the semantic effects of the alignment in syntax.

Note also that Kendon [2004] used this example as an illustration of how speakers re-use and revise gestural form similarly to linguistic phrases: the interlocutor did not hear the word "greasy" and asked the speaker to repeat the word. Then the speaker repeated not only the linguistic material but the entire speech-gesture ensemble. This suggests not only of the composite nature of the multimodal signal, but also that speakers employ gesture in the same way they do speech. Even though we do not model the anaphoric relations of multimodal actions in discourse, this example confirms our starting claim for applying standard linguistic methods to multimodal actions, and thus making the formal model of gesture as uniform as possible to that of purely linguistic units.

Another example that demonstrates the interaction between the gestural performance and the syntax of the linguistic phrase for the multimodal meaning is the naturally occurring example (1.6), Figure 5, taken from the corpus of Loehr [2004].

Figure 5: Gesture Representing the Conduit Metaphor of Teaching, example (1.6) [Loehr, 2004]

(1.6) If I was to [$_{PN}$rea]lly [$_{N}$teach] someone how to be a professional musician
*The right hand is holding a small object and the left hand is open flat, relaxed with palms facing up; both hands move to the frontal space to possibly denote a conduit metaphor.*

The author interpreted the gesture as a "conduit for a hypothetical situation", i.e., the hypothetical situation of teaching. Note that the expressive part of the gesture (following the original annotation of the author) was performed along the adverbial pre-head modifier "really" but did not temporally overlap the head word "teach". However, the conduit interpretation is accessible only after linking the gesture with the head of the phrase—the verb "teach". This interpretation would arise no matter whether the gesture was synchronous with the pre-head modifier only, with the verb head, or even with both the pre-head modifier and the verb head. But intuitively the interpretation of the gesture would be different or even incoherent if it was synchronous only with the subject noun phrase "I". The fact that the annotator interpreted the gesture in this way suggests that gestures interact with the syntax of the temporally co-occurring speech phrase — in this case, the head of the phrase.

While we believe that the interaction between gesture and syntax is affected by constituency, we do not expect that there is any restriction on the syntactic label of the linguistic phrase that gesture could possibly align with. In (1.5), for instance, the gesture could align with the adjective only “greasy”, with the adjectival phrase “greasy in the whole room place”, with the verb phrase “made everything greasy in the whole place” or even with the whole sentence “they made everything in the whole room place”. We even argue that we can arrive at the gestural interpretation of greasiness even if the gesture happened along with the verb “made” only. In Chapter 4, we perform an empirical investigation to verify whether the syntactic category constrains gesture performance. Also, in Section 5.3, we propose grammar construction rules that will demonstrate how to access the content of “greasy” if the gesture was performed along with “made” only.

#### 1.3.2.3 Negation in Speech and Negation in Gesture

Based on an in-depth study of audio-visual recordings of everyday interactions, Harrison [2010] found that gestures of negation undergo the same grammatical organisation as language. The gestures of negation include those hand movements where the hand moves across the body with palm facing down to express denial or rejection. This gesture does not contribute propositional content to the utterance, but it rather designates the negative speech act performed in speech [Harrison, 2010]. The author studied the temporal unfolding of negation gestures in relation to the temporal unfolding of negation in speech. The linguistic structures of concern were the ‘node’ and ‘scope’ of negation defined as follows:

> The node is the location of a negative form, and the scope is the stretch of language to which the negation applies. [Harrison, 2010].

It was found that the node of negation was synchronised with the expressive part of the gesture, the so-called stroke, and the scope of negation was synchronised with the hold after the gesture stroke. We shall illustrate this with example (1.7), taken from Harrison [2010]. Here the speaker was talking how he ended his relationship with his girlfriend to date other girls from his school, but his plan failed since the other girls were friends with his ex-girlfriend.

(1.7) I was like [pause] I was like [$_N$ no]-go territory
*Right hand moves across the body with palm facing down.*

(a) I was like …I was like… (b) …no-go… (c) …territory

Figure 6: Illustration of a Gesture of Negation, example (1.7) [Harrison, 2010]

Along the pause in speech, the speaker prepared the gesture by bringing his right arm across the body with a horizontally open palm (Figure 6(a)). Then along with the scope-bearing element in language, the negation “no”, which is also the prosodically prominent element, the speaker moved his arm rapidly towards the right periphery on the horizontal axis (Figure 6(b)), and finally the hand was held still while uttering the verbal material outscoped by “no” (Figure 6(c)). Note also that the hand did not reach a resting position before the speaker had uttered the verbal material outscoped by the negation—“territory”.

#### 1.3.2.4 Summary

The empirical data demonstrate that the speech phrases a gesture can align with cannot be adequately defined only in terms of the temporal performance of one mode relative to the performance of the other. The fact that the constituent structure of the utterance—prosodic and/or syntactic—influences the decisions on speech-gesture integration, and hence the corresponding interpretation of the multimodal action motivates constraining the alignment in the grammar. This is possible via construction rules that articulate the speech phrase(s) that gesture can align with: in other words, the alignment of speech and gesture in the grammar is a matter of a constraint-based syntactic attachment in a single derivation tree. Likewise, if there is a choice as to which phrase gesture could align to, we model this via structural—that is, attachment—ambiguity. The advantages of modelling multimodal actions via a grammar is not only that we can account for the constraints of form on the alignment, but also that we can use *standard*

methods for semantic composition to map multimodal form to multimodal meaning. Given the assumptions about the semantics/pragmatics interface (recall Section 1.3.1), the construction rules for speech and gesture alignment introduce an (underspecified) relation $R(s,g)$ between the content $s$ of the speech signal and the content $g$ of the gesture signal, which demonstrates the $s$ and $g$ are coherently related. Pragmatic processing would then assign a more specific value to this relation. This approach is analogous to the analysis of free adjuncts in language where the attachment in syntax determines what the free adjunct modifies, and semantics introduces an underspecified relation between its content and the content of the phrase it modifies so as to handle the distinct ways the free adjunct can be related with its matrix clause. For instance, the relation between the free adjunct and the matrix clause in "Opening the drawer, he found a revolver" is Narration, whereas in "Tired and exhausted, none of the students passed" the matrix clause and the free adjunct are connected through Explanation.

Carried over multimodal actions, we can illustrate this using example (1.8), Figure 7. This utterance, taken from Kendon [2004, pp. 113–116], is a small fragment of a narrative where the speaker described how different types of cheese used to be prepared and stored in his father's owned grocery shop.

(1.8) He used to go down there and throw …[$_{X^*}$grou]nd rice over it.
*The speaker moves his right hand forward; fingers are flexed inward in contact with the palm; the tip of the thumb is resting on the first joint of the index finger. The hand is moved twice by extending the wrist. The gesture resembles scattering a handful of dust/powder over some surface.*

Here we can infer an Elaboration relation between the content of speech and the content of gesture, that is the gesture elaborates on the specific way of executing the throwing motion.[9] This can be roughly paraphrased as "Sieving it through his fingers, he used to throw ground rice". In comparison, imagine that the gesture in (1.3) was accompanied by "He sold it to the customers". In this case, the relation between the gesture content and the speech content would be Narration—that is, the event of selling the cake to customers (conveyed in speech) is a continuation of the event of slicing the cake in bits (conveyed by the gesture). The linguistic paraphrase in this instance would be "By slicing the cake in bits, he would sell it to his customers". So not only are we assuming a unified pragmatic model of speech and gesture interpretation, but our aim

[9]This analysis is possible with the additional assumptions that there is an alternative set comprising the various ways in which a throwing action can be performed.

Figure 7: Gesture Depicting Throwing Ground Rice, example (1.8) [Kendon, 2004]

goes further as we apply a unified method of semantic composition to purely linguistic elementary units (that is, clauses) and multimodal elementary units (that is, a clause and a co-speech gesture).

To recapitulate, the advantages of capturing the speech-gesture interaction through a multimodal grammar are two-fold: first, we guarantee a smooth transition between syntax, semantics and pragmatics; and second we capture underlying evidence from the multimodal data about the interaction between the form of the speech signal (its prosody and syntax) and the form of the gesture signal (its temporal performance relative to the one of speech).

### 1.3.3 Cognition: an Inseparable System

Our formal model is built upon the inseparable system of speech and gesture at a cognitive level which is surface realised *in parallel* by spoken material and by visual material. There is evidence in the literature that spontaneous co-speech gestures are part of the conceptual planning of the utterance, and they are formed at a level prior to the decisions made about the surface realisation of the linguistic utterance [Alibali, Kita, and Young, 2000; Hostetter, Alibali, and Kita, 2007]. Importantly, the macro- and micro-levels of planning an utterance include planning what modalities to employ, what information to package into gesture vs. into speech through a constant interaction

between the two planners: the one responsible for the linguistic messages, and the one for the visuo-spatial representations.

At the level of production, the literature offers sufficient evidence about this on-line interaction. For instance, Kita et al. [2007] established through production tests that the gestural encoding is constrained by the packaging of information in syntactic structures. De Ruiter, Bangerter, and Dings [2012] attested empirically that the frequency of the gesture modality can be reliably predicted from the frequency of the speech modality: the material that is hard to verbalise and could provoke decrease in speech production is not compensated for by increased gesture performance. This is also closely related to the finding that gesture is suspended along with the speech disfluency and then it is resumed once the speech is resumed [Seyfeddinipur and Kita, 2001; Seyfeddinipur, 2006]. The inseparable nature of speech and gesture modalities was given a theoretical account in the Growth Point theory of McNeill [2005]. The growth point is a minimal unit that combines the modes of imagistic thinking and linguistic thinking. The growth is a dialectic process that initiates the unpacking of imagistic categories into gesture and of linguistic categories into spoken words. Within the dialectic process the two semiotic modes of imagistic thinking and spoken thinking feed each other, rather than one mode feeding the other.

By drawing on a unified formal model of speech-gesture alignment, we can understand the conceptualisation of linguistic and imagistic messages, as well as their interaction on the level of *production*, in that we treat gesture signals and speech signals as objects suitable for syntactic (and corresponding semantic) manipulation. In particular, via a multimodal grammar, we can elegantly capture the linkages at a conceptual level that trigger the synchronous production of speech and gesture: for instance, defining how gestural form and content is informed by the linguistic capabilities is a matter of constraining the choices of alignment in the grammar; defining the interaction between the spoken and the gesture signal is a matter of establishing a coherence relation between them.

## 1.4 Speech-Gesture Alignment

Given our assumptions about the model of the semantics/pragmatics interface outlined in Section 1.3.1 and the empirical finding that speech and gesture interaction is dependent on form (including syntax, prosody and timing of speech relative to the timing of gesture), we can now refine the concept of speech-gesture alignment. We first provide

some background information as to how the speech and gesture interaction is generally described in the literature.

It has become commonplace in the gesture community to designate the interaction between speech and co-speech gesture as **synchrony** and **co-expressivity** [McNeill, 1992; McNeill, 2005; Kendon, 2004]. For instance, McNeill [1992] proposed to view synchrony as a co-temporal performance which reflects the speech and gesture interaction on three different levels:

i. phonological: the expressive part of the gesture, the so-called stroke, happens at the same time as the prominent syllable in speech. It may precede it, but never follow it;

ii. semantic: if speech and gesture happen at the same time, they convey the same idea unit;

iii. pragmatic: if speech and gesture happen at the same time, they serve the same pragmatic function.

On a similar account, the meta-communicative signal hypothesis of Engle [2000] predicted that multimodal timing was the factor determining the composite and the communicative nature of the spoken and of the gestural signal.

By contrast, we find evidence for the temporal speech-gesture **"asynchrony"**: Morrel-Samuels and Krauss [1992] demonstrated empirically that the onset of gesture preceded or coincided with the onset of the most closely associated lexical item. This difference in the gesture initialisation in relation to the one of speech was also accounted for in the Sketch model of speech and gesture production [De Ruiter, 1998].

Also, Clark [1996, p. 178] provided empirical evidence for the gesture **sequentiality** as in utterance (1.9):

(1.9) I got out of the car, and I just ...
*Demonstration of turning around and bumping the head on an invisible telephone pole.*

By the same token, Oviatt, DeAngeli, and Kuhn [1997] attested only 25% overlap of users' commands and the associated pointing gesture in 'point-and-click' devices, and around 50% sequential integration of speech and pen-based input, with the pen input preceding the one of speech.

These studies demonstrate that there is not yet a conclusive methodology of how to establish the integrated nature of two signals: for instance, if synchrony is indeed a guiding factor, how do we explain multimodal actions of asynchronous speech and gesture that are still perceived as semantically well-formed? We assume that these discrepancies in the literature are due to two underlying reasons, the first one being the lack of well-defined criteria of what is considered the temporal extension of gesture: is it the gesture stroke that is temporally synchronous with the spoken signal, the gesture phases comprising the material from the beginning of gesture to its semantic peak, or the entire gesture excursion from a rest to a rest? Whereas some analyses precisely state that gesture is identified with the stroke [McNeill, 2005], others forgo gesture phase partitions [Morrel-Samuels and Krauss, 1992; Engle, 2000]. The second reason is related to the factors that influence the decision of which speech signal gesture should be linked to. Whereas McNeill [2005] identifies synchrony with co-expressivity, we claim that the linking of gesture to speech is a matter of form (here we include prosodic and/or syntactic constituency, and also relative timing), meaning and pragmatic coherence. We use the concept *speech-gesture alignment* to designate this more complex integration pattern.

With all this in mind, we are now in a position to spell out our own definitions of synchrony, co-expressivity and alignment as follows:

**Definition 1.4.1. Speech-Gesture Synchrony.** *The synchrony between speech and gesture is based on the temporal performance of one signal relative to the temporal performance of the other, that is, two signals are synchronous if they happen within the same time frame.*

**Definition 1.4.2. Speech-Gesture Co-expressivity**. *If the speech signal and the gesture signal convey complementary or redundant information, they are co-expressive.*

**Definition 1.4.3. Speech-Gesture Alignment.** *The choice of which linguistic 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 an underspecified logical formula supporting (i) and hence also (ii). The*

*derivation of the single multimodal syntactic structure—constrained by the prosody of the speech signal gesture temporally overlaps—is achieved within the grammar.*

A few things should be noted about these definitions: Definition 1.4.1 is solely based on timing and it does not consider the meanings of the speech signal and the gesture signal. In contrast, Definition 1.4.2 is driven from the signals' meanings and it stays neutral about the temporal relations of speech and gesture. Finally, Definition 1.4.3 is a more complex notion that encompasses both form—that is, the timing of speech relative to the timing of gesture, and also the prosodic properties of the speech signal—and meaning—that is, what are the meaning representations mapped from the common multimodal syntactic structure, how are speech and gesture semantically related and what are the preferred interpretations in context? More specifically, recall from Section 1.3.1 that we used the concept of alignment to designate that speech and gesture are connected through a coherence relation which is inferred on the basis of the gestural semantics. Also, driven from the empirical evidence that gesture interacts with the form of the linguistic signal (its prosody and syntax) and that this interaction has semantic effects (recall example (1.5) and the subsequent discussion), we extended the notion of alignment to encompass form—that is, multimodal syntax—and meaning—that is, multimodal semantics. In other words, we argue for multimodal alignment within the syntactic grammar on the basis that the temporal performance of gesture interacts with the constituent structure of the spoken utterance, which in turn has effects on the interpretation(s) of the multimodal action, given our assumptions about the model of pragmatic theory (Section 1.3.1). This means that the integration within the grammar is guided not by timing per se, but rather by constituency. Our approach can be compared with prior work on speech and gesture integration where integration is not achieved within a single derivation tree and where the information about the relative timings of the input modalities is captured outside the grammar [Giuliani and Knoll, 2007].

With this in mind, we shall propose grammar construction rules for attaching gesture to a speech phrase in a single multimodal syntax tree that, using standard methods for semantic composition, would map to a (partial) meaning representation. The construction rules for this attachment introduce a semantic relation between the content of speech and the content of gesture, which captures the fact that speech and gesture are coherently connected (see Lascarides and Stone [2009b]). The re-construction of

this meaning representation to a pragmatically preferred interpretation and the processing of this semantic relation happens externally to the grammar at the semantics/pragmatics interface. Essentially, the definition of alignment covers the syntactic attachment of speech and gesture in a single syntax tree (under some constraints, e.g., what are the prosodic properties of the speech signal, when did the gesture happen relative to the speech signal), the meaning representation(s) mapped from this syntax tree and also the interpretation of the multimodal action in context.

This notion is also in line with Bergmann and Kopp [2007] for whom speech and gesture alignment involves the following aspects:

> ...the meaning that the verbal and non-verbal behaviors convey, the form they take up in doing so, the manner in which they are performed, their relative temporal arrangement, as well as their coordinated organization in a phrasal structure of utterance [Bergmann and Kopp, 2007].

Whereas alignment 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. In this way, our model goes beyond previous work where speech and gesture co-expressivity is not derived from syntax [McNeill, 1992]. An overall challenge is to constrain the alignment configurations in a qualitative 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. Our programmatic plan is thus similar to that of the development of large-scale grammars since we intend to provide analyses for all well-formed signals in a domain-independent fashion. Note also that Definition 1.4.3 is not entirely based on quantitative factors such as simultaneity. In this way, we guarantee that speech-gesture alignment is obtained by exploring the linguistic properties of the multimodal action. This definition also dovetails with the fact that our own perceptual system can make the judgement of which signals are/can be aligned and which are not/cannot be (recall utterances (1.4), (1.5) and (1.6) and the related discussions).

## 1.5 This Thesis in Context

This thesis benefits from the previous findings concerning the integrated nature of gesture and speech, as discussed in the gesture literature. First, we capture the claim from

the descriptive studies that speech and gesture are an integrated "ensemble" conveying a "single thought" [Kendon, 2004; McNeill, 2005] by producing *a single derivation tree* for the multimodal signal which yields a formal meaning representation that is defined in terms of the meaning of the speech, the meaning of the gesture, and their mode of combination. Second, models of speech and gesture production from psycholinguistic studies have suggested that there is a constant information exchange at a pre-verbal level between the module responsible for generating gestures and the module responsible for generating linguistic messages [Kita and Özyürek, 2003]. We represent this interaction by constraining the choices of alignment in the grammar. Third, there is evidence in the literature that gestures of negation undergo the same grammatical organisation as language [Harrison, 2010], and so we are going to model speech and gesture by using standard linguistic methods for analysing syntax and for composing logical forms from those syntactic analyses. We also benefit from previous formal, but domain-specific, approaches to gesture, such as pen and voice devices [Johnston, 1998a; Johnston, 1998b] or embodied conversational agents [Cassell et al., 1998]. Our distinct contribution to this body of formal work is to formalise gesture's representation and its integration into the syntactic tree while abstracting away from the specific domain of application.

The previous approaches to co-speech gesture notwithstanding, a consistent domain-independent model where gesture is formalised in parallel with speech for the purposes of interpretation does not yet exist. This thesis contributes a methodology for the precise domain-independent derivation of speech and co-speech gesture into a single syntax tree that maps to an (underspecified) multimodal meaning representation, thereby supporting the range of plausible interpretations in context.

## 1.6 Thesis Overview

This thesis is structured as follows: Chapter 2 provides extensive empirical material that forms the basis for describing the main challenges for the formal models arising from gestural ambiguity. Chapter 3 proceeds with an overview of the related literature, including the descriptive, cognitive and formal models of speech-gesture interaction. In this chapter, we show how our work fits in the broader context of related studies and how it contributes to certain under-researched areas. In Chapter 4, we present our empirical studies, which shed light on the speech-gesture interaction at the level of form (prosody and syntax) and meaning. We use these empirically extracted generalisa-

tions to spell out a set of theory-independent grammar construction rules in Chapter 5. Chapter 6 describes the formalisation of the grammar rules in the HPSG framework. Chapter 7 discusses the implementation of the theoretical rules within the LKB/PET parsing platform and its evaluation within the [incr tsdb()] grammar profiling system.

## 1.7 Published Work

The results of this thesis have been disseminated in several publications. These include:

- Alahverdzhieva and Lascarides [2010] discusses the formal analysis of depicting co-speech gesture and speech in the HPSG grammar framework (here presented in Chapter 6).
- Alahverdzhieva and Lascarides [2011b] discusses how deictic gesture can be integrated with speech in formal grammars (here presented in Chapter 5).
- Alahverdzhieva and Lascarides [2011a] provides a formal analysis of deictic gesture and speech in the HPSG grammar framework (here presented in Chapter 6).
- Alahverdzhieva and Lascarides [2011c] discusses the semantic composition of speech and co-speech gestures for the purposes of multimodal grammars (here presented in Chapter 5).
- Alahverdzhieva, Flickinger, and Lascarides [2012] provides an overview of the implementation of the multimodal grammar within the LKB/PET grammar engineering platform (here presented in Chapter 7).

# Chapter 2

# Data

*Evvi mai cosa più visibile, più comune e più semplice del gestire dell'uomo?*
*Eppure quanto poco si riconosce di esso!*[1]
[De Jorio, 1832]

In the previous chapter we argued for analysing the speech-gesture alignment within the grammar on the grounds that the gesture performance is constrained by the form of the linguistic phrase. Taking also into account our assumptions about the pragmatic theory which accesses only the compositional semantics of the linguistic unit and/or gesture unit, we analyse any information about form in the grammar. We also stated that despite the constrained interaction between speech and gesture, the ambiguity of gesture opens up multiple alignment configurations. We formalise these as attachment ambiguities in the syntax tree, which in turn yield distinct meaning representations.[2] Our main challenge is thus to provide a formal model of speech-gesture alignment without under-determining or over-determining the possible analyses and their representations of meaning.

The aim of this chapter is to first introduce the focus of our work—co-speech gestures—and then to show, through examples of multimodal actions, why we view gestures as both ambiguous and constrained. In Section 2.1, we provide a taxonomy of hand movements, gesture dimensions and their structural organisation. In Section 2.2, we introduce empirical data that illustrate the range of ambiguity on one hand, and the constrained speech-gesture alignment on the other.

[1] Was there anything more visible, more common and more simple than the gesturing of men? And yet we know so little of it.

[2] Although the logical forms are distinct, they may be sometimes truth-conditionally equivalent. This happens in purely linguistic grammars too [Copestake and Flickinger, 2000].

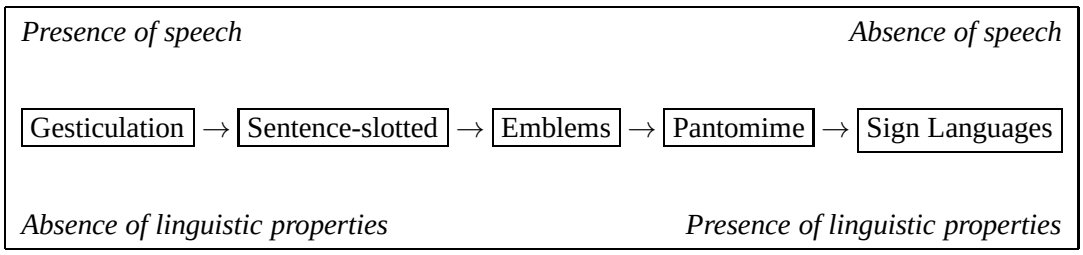



Figure 8: Kendon's Continuum [Kendon, 1988; McNeill, 1992]

## 2.1 Gesture Background

### 2.1.1 Gesture Dimensions

Our study focuses on improvised hand gestures performed spontaneously with speech. To clarify what gestures are in, and out, of our study, Figure 8 presents a taxonomy of all hand movements as originally proposed by Kendon [1988] and McNeill [1992]. In the gesture community, this taxonomy is known as "Kendon's Continuum", named in honour of Adam Kendon by McNeill [1992]. The hand movements are ordered on the basis of the degree of speech presence that is necessary to understand the hand signal, and the degree of linguistic properties that the gesture exhibits. The leftmost side is taken up by speech-accompanying hand movements, known as *co-speech gestures, co-verbal gestures* or *gesticulation*. One of their key features is that they can be understood only in the context of the co-occurring speech, as detailed in Section 1.1.1. Further, *sentence-slotted* hand movements play a syntactic (and a corresponding semantic) function by filling a grammatical slot in a syntactically incomplete utterance of "mixed syntax" [Slama-Cazacu, 1976]. Recent studies carried over German data indicate that the integration of a gesture into the syntactic structure of an utterance follows the underlying grammatical organisation [Ladewig, 2010]. To illustrate this, consider again utterance (1.9)[Clark, 1996], repeated in (2.1), where the verb position is filled by a depicting head gesture.

(2.1) I got out of the car, and I just ...
*Demonstration of turning around and bumping the head on an invisible telephone pole.*

*Emblems* or 'narrow gloss' gestures [Kendon, 2004]—such as thumbs up, thumbs down or the shush movement—have a socially and/or culturally established form associated with a specific meaning. In *pantomimic* gestures, the speaker takes on the role

of a protagonist or participant that is referred to in speech, and enacts the event that the gesture's meaning refers to; e.g., a toddler acting out a tantrum originally performed by another child [Doherty-Sneddon, 2003, p. 65]. Finally, the right-most side is occupied by *sign languages*—such as American Sign Language—which form an autonomous language system.

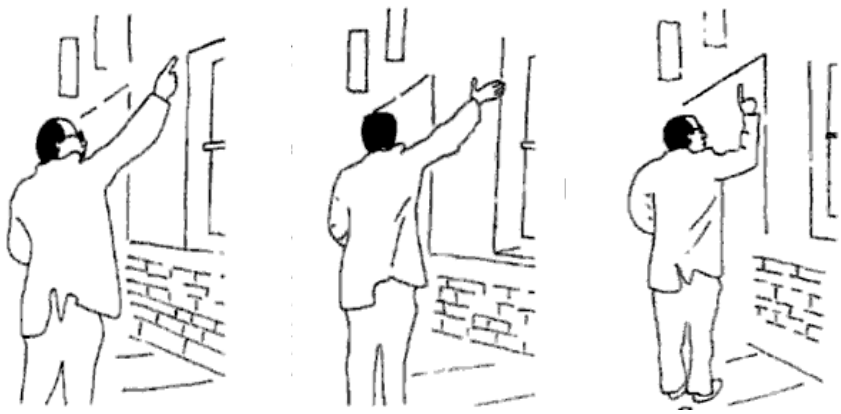
McNeill [2005] observed that moving from the left-hand side to the right-hand side of this spectrum, the coherent performance of the gesture becomes less reliant on speech having to be present, and thus the linguistic properties of the gesture increase. Whereas gesticulations can only be understood by linking them to the speech context, sign languages have the power of a stand-alone system for communication.

This thesis concerns the form-meaning mapping of *co-speech gestures*. We analyse their ambiguous form as one that yields an abstract and partial representation of meaning, supporting an open-ended number of specific interpretations in their context of use. The aligned speech is a vital source for resolving this abstract meaning to a specific interpretation. This is in contrast with narrow gloss gestures whose form is linked to a socially and/or culturally conventionalised meaning. As we shall see below, there also exist borderline movements whose recurrent patterns place them somewhere between improvised gestures and those of conventionalised form.

Based on their function, Kendon [2004] differentiates the following broad classes of co-speech gestures:

1. **Descriptive (Depicting/Representative)**. The hand depicts, models the object of reference or enacts a specific behaviour. The depiction can be *literal* (also known as iconic)—for instance, tracing a path with hands in the frontal space while talking about (the size of) a cake as in Figure 3(a), page 12. Alternatively, it can be *metaphoric*—for instance, moving the hand from the left to the right periphery to refer to the past and the future. The function of the depicting gestures is understood as visually characterising the referent in terms of qualitative features which contribute content to the proposition denoted by the whole multimodal action, with the gesture content being either redundant or complementary with respect to the speech content. These gestures are "a part of the referential content of their respective utterances" [Kendon, 2004, p. 158].

2. **Pointing (Deictic)**. The hand signal contributes to the propositional content of the utterance by highlighting a region in space so as to identify the referent's location in Euclidean space. The pointing can be *concrete* as when pointing to



(a) You'll notice . . . (b) . . . that's also part of . . . (c) [pause 0.3 sec] Laxton Junior school

Figure 9: Nomination Deictic Gesture, example (2.2) [Kendon, 2004]

an object or individual present in the communicative situation; for instance, the speaker points with an extended index finger or an extended arm to a physically present individual while uttering "I saw this man" with the gesture happening along with "this man".

The pointing can also be *abstract* [McNeill, 2005]: here the referent is a virtually created object in the gestured space just in front of the speaker, which is physically absent from the communicative situation; for instance, a speaker says "And there was this guy" and he uses his index finger to refer to an individual when, in fact, the individual denoted by "this guy" is not physically present.

Kendon [2004] also introduces *nomination* pointing which possesses the metanarrative function of giving prominence to a word or phrase by simultaneously extending an index finger, and so the pointing has no propositional effects on the utterance. Utterance (2.2), Figure 9, borrowed from Kendon [2004, p. 142], is a fragment from a guided tour of Northant. An important landmark of the town was an old public school, which the speaker presented by extending his index finger to emphasise the word "Laxton" (displayed in Figure 9(c)).

(2.2) You'll notice that's also part of [pause 0.3 sec] [$_N$ Laxton] Junior School.
*Along with "Laxton", the speaker raises his voice and extends his index finger to nominate the word prominent.*

3. **Performative (Pragmatic/Recurrent)**. Gestures with a recurrent form-meaning pairing can be organised in *gesture families* [Kendon, 2004; Müller, 2004; Bressem, 2007]. Unlike depicting gestures, pragmatic gestures do not add "referential content" to the multimodal action, but rather they qualify the speech act performed by speech. For instance, the 'ring' gesture with the tip of the index finger and thumb closed together and the other three fingers held loose in a crescent-like shape has the very abstract idea of precision, exactness of the speaker's arguments; the open hand with a vertical palm indicates a termination of a line of actions; palm facing down as in Figure 6, page 18 conveys negation, denial or interruption of a process; the open hand supine as in Figure 1, page 1 expresses an offer or readiness to accept something from the interlocutor.

Bavelas et al. [1995] also introduced the notion of **interactive gestures** to designate the use of the hand as a resource for regulating the interaction rather than for conveying topic-related information. For instance, a hand with open palm extends towards the interlocutor to offer them the floor, to ask for help, or to cite their contribution. Other spontaneous communicative actions include **beats**, also known as batons [McNeill, 1992]. These are formless flicks of the hand, beating the time along with the rhythm of the speech. They often serve pragmatic functions such as commenting on one's own utterance or giving prominence to aspects of the speech [Cassell, 2000]. Unlike nomination pointing, beats are formless and they often superimpose other gestures [McCullough, 2010].

Within the class of co-speech gestures, the need for speech accompaniment is also a matter of degree. Generally, descriptive gestures lack a specific interpretation out of context, and they can be understood only by reasoning about the semantic or pragmatic relation they bear to speech. In comparison, the form of the performative and interactive gestures is somewhere on the border between metaphoric gestures and emblems, and so a general meaning can be abstracted from them.

This thesis concerns the full spectrum of co-speech gestures with a minor deviation from the above-mentioned nomenclature: we treat the group of hand movements with recurrent form features that usually serve a pragmatic function ("performative" and "interactive") as metaphorically depicting gestures or as a combination of metaphorically depicting and pointing gestures depending on how the speech content and gesture content are semantically connected. If the gesture qualifies the speech act, then we treat those movements as metaphorically depicting. And if the gesture serves a meta-narrative function and has spatial properties at the same time, we treat that move-

Figure 10: Hand Gesture Pointing at Another Participant, example (2.3)

ment as metaphorically depicting-deictic. This treatment is consistent with previous observations in the literature that gestures are multidimensional performances where a single gesture can display properties of one or more gesture classes (see (2.4) below and the related discussion).

To illustrate our deviation from the gesture literature with respect to the nomenclature of pragmatic gestures, consider utterance (2.3), Figure 10. This utterance is extracted from a multi-party conversation where the participants discussed the design of a remote control [Carletta, 2006]. The speaker on the right side used a pointing gesture towards the interlocutor seated diagonally from her so as to acknowledge a statement previously made.

(2.3) 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.*

Whereas in the gesture literature this hand movement might be assigned the category of an interactive gesture, we treat it as both concrete deictic—it has spatial

properties—and metaphorically depicting—the open hand metaphorically manifests acceptance of a previous statement. Likewise, while the gesture community designates the gesture in (1.7), page 17 as pragmatic, we treat it as metaphorically depicting.

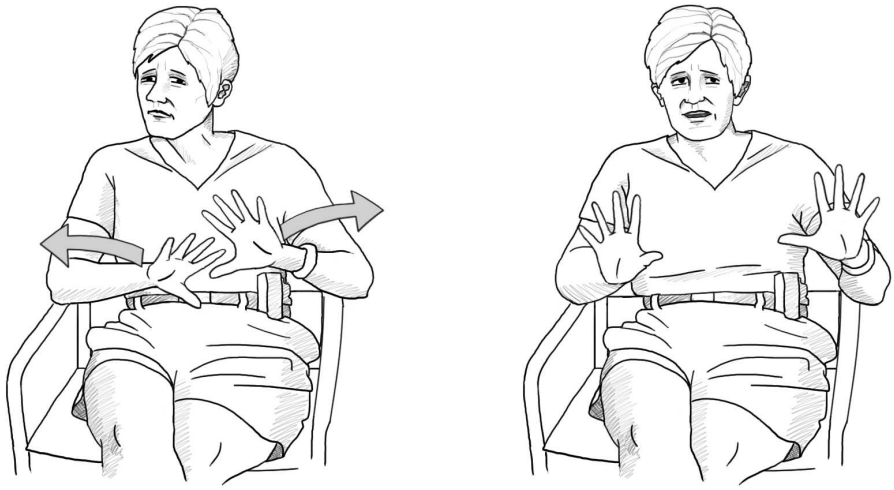
Even though our nomenclature does not distinguish pragmatic and interactive gestures as a stand-alone class, we shall account for the distinct ways these gestures designate the referent by an underspecified relation between the speech content and the gesture content which supports the distinct ways speech and gesture are connected, following the approach to modelling semantic connections among sentences in coherent text. The specific value of this relation accounts for the way the gesture refers to the salient features of the speech denotation. For instance, the gesture in (2.3) identifies the spatial coordinates of the speaker referred to by "she", and so we can infer an Identity relation between the speech content and the gesture content. This gesture also qualifies the positive attitude towards the statement being made rather than the content of the statement itself. Intuitively here, we can infer a Metatalk relation as discussed in the discourse literature [Polanyi, 1985]. Instead of limiting the gesture interpretation to a specific class (e.g., pragmatic), our model introduces a semantic relation between speech and gesture which supports the various ways they can be connected (e.g., via Identity, Metatalk).

Example (2.3) also demonstrates that co-speech gestures cannot be distributed among mutually exclusive types. Following McNeill [2005], we analyse gestures as multi-dimensional performances where the dimensions include *literal depiction, metaphoric depiction, deixis* and *emphasis by beat*. This means that a single gesture can display features of more than one of these dimensions. To illustrate this, let us consider utterance (2.4), Figure 11 extracted from a conversation where the speaker discussed how her cupboard doors were crooked and those at the bottom were level [Loehr, 2004].

(2.4) The $[\underline{_{PN}\text{bottom}}]$ worked $[_{N}\text{fine}]$ ...

*Both hands are rested on the knees. The speaker lifts them in the frontal space with palms almost facing forward, fingers extended and moves them rapidly to the left and right periphery.*

The hand movement across the x-axis literally depicts some salient feature of the synchronous speech content, namely objects positioned at the bottom. This gesture is also a recurrent metaphor of a completion of a process. The fact that two annotators out of four assigned a literally depicting and a metaphorically depicting dimension to this



(a) The bottom... (b) ...worked fine

Figure 11: Example of Gesture's Multidimensionality, example (2.4) [Loehr, 2004]

hand movement suggests that even in context gestures are at heart multidimensional, and that people tend to perceive this multidimensionality. Likewise, a vertically held right palm bent at a 90-degree angle while saying "Turn around the next corner" is depicting in that the bent form of the hand is a visual representation of the shape of a corner, and deictic in that the fingers are oriented in the direction of the turning, namely left. We shall formally capture the multidimensionality of communicative gestures in a typed hierarchy where a subtype can inherit information from more than one supertype (we discuss this at length in Section 6.3). In this way, the gesture in (2.4) would inherit information from the types *literal depiction* and *metaphoric depiction*. This also explains how we can arrive at metaphorically depicting–concrete deictic dimension for the gesture in (2.3). The multiple inheritance representation also allows us to account for the fact that beats often superimpose other dimensions [McCullough, 2010].[3]

### 2.1.2 Structural Organisation of Gesture

The gestural excursion from lifting the hands into the gesture space to retracting them to a rest is known as a *gesture unit* [Kendon, 2004]. A gesture unit can contain one

[3]Note that the gestural multidimensionality is not unanimously accepted: for instance, some authors claim that one dimension always dominates over the other dimension, making the gesture unidimensional (Mandana Seyfeddinipur (personal communication)).

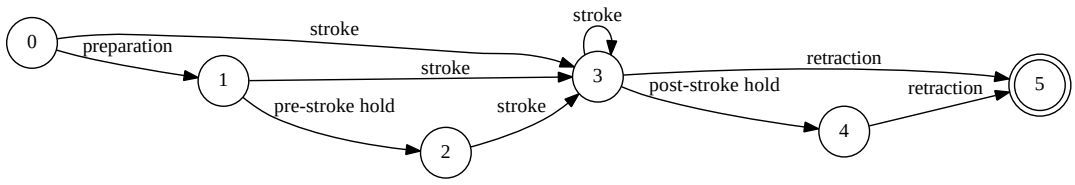



Figure 12: Finite-State Representation of Gesture Phases

or more *gesture phrases* which are made up by the following *gesture phases*[4] (see the finite-state representation of gesture phases in Figure 12):

- a non-obligatory *preparation*: hands are lifted from the rest position to the frontal space to perform the semantically intended motion
- a non-obligatory *pre-stroke hold*: hands are sustained in a position before reaching the kinetic peak; this phase serves as a trigger for the lexical item(s) produced during the expressive part of the gesture and it thus indicates the point where the speech-gesture synchrony is about to begin
- an obligatory *stroke* (one or more): the kinetic peak carrying the gestural meaning. It can be both static or dynamic: in the former case, the pointing forelimbs are stationary in the expressive position (e.g., the stroke of a deictic gesture is often static), and in the latter the gesture's meaning is derived from a movement of the pointing forelimbs.
- a non-obligatory *post-stroke hold*: the hands sustain their expressive position reached during the stroke
- an obligatory *recovery (retraction)*: hands return to a rest position.

For the purposes of mapping multimodal form to multimodal meaning, we consider only the content-bearing part of the gesture—the stroke. This does not prevent us from accounting for the way post-stroke holds can convey information about the semantic scope of elements in the logical forms introduced by speech (recall utterance (1.7), page 17 where the post-stroke hold conveys information about the scope of the negation in gesture and speech). Instead of aligning gesture stroke and post-stroke hold

[4]The technical terms *gesture phrase* and *gesture phase*, first introduced by Kendon [1972], have become standard in the gesture community.

with the temporally co-occurring speech elements, our grammar construction rules for speech and gesture alignment (detailed in Chapter 5) are based on the following two constraints:

1. the temporal performance of the stroke relative to the temporal performance of the speech, and

2. the syntactic and/or prosodic structure of the linguistic phrase (as attested by empirical investigation).

This enables us to consider speech elements whose temporal performance is outside the temporal performance of the gesture stroke, but still form a constituent structure with the temporally co-occurring elements. Our motivation for not considering the post-stroke hold is that its function is generally rather vague and we lack sufficient empirical evidence to support any concrete and formalisable hypothesis with respect to its contribution to the meaning of the multimodal action. While the post-stroke hold often indicates that the speaker elaborates on the idea expressed during the stroke, it can also accompany termination of the expressive phase—that is, the energy articulated by the hand is no longer perceived despite it being held in a post-stroke position.

## 2.2 Main Challenges

### 2.2.1 Range of Ambiguity

Gestural ambiguity and constrained speech-gesture alignment are the bread-and-butter of this thesis. The ambiguity pertains to the gestural form that can adopt multiple interpretations in context. As an illustration, recall our starting example in (1.1) (vertical hands depicting stacking books) and the numerous ways this gesture can be interpreted as discussed in Section 1.1.1. We also observed in Section 1.3.2 that speech and gesture cannot be aligned in unrestricted ways — there are form-based rules that constrain the choices of alignment. We formalise the permissible alignments in terms of syntactic attachments of the gesture to the linguistic phrase.

This section introduces empirical data so as to shed more light on the gestural ambiguity and the challenges associated with it. We shall see that the ambiguity concerns the form-meaning mapping, the distinct attachments of the gesture signal to the (aligned) linguistic phrase, and also the syntactic and semantic ambiguity between the

speech signal and the gesture signal. Whenever possible, we support our assumptions by real examples from the following multimodal corpora:

1. A 165-second collection of four recorded meetings annotated for gesture and intonation [Loehr, 2004]. The gesture annotation includes marked gesture phrases and gesture phases, and the intonation annotation, based on the ToBI guidelines, includes low or high pitch accent tones, intermediate phrases and intonation phrases. The corpus was labelled with the view of exploring the interaction between intonation and gesture.

2. A 5.53-min recording from the Talkbank Data[5] whose domain is living-space descriptions and navigation giving. We annotated it with gesture phrases, gesture phases, pitch accents (nuclear, non-nuclear and pre-nuclear) and prosodic phrases.

3. Observation IS1008c, speaker C from the AMI corpus [Carletta, 2006].[6] The corpus is a multi-party face-to-face conversation among four people discussing the design of a remote control. We equipped it with the following annotation layers: gesture phrases, gesture phases, pitch accents (nuclear, non-nuclear and pre-nuclear) and prosodic phrases.

We postpone the detailed discussion of the multimodal collections and the ways in which we extended their existing annotations to Chapter 4. Now we focus on some interesting observations concerning the data itself. Whenever data-driven multimodal actions were not available we made use of constructed examples. We also use constructed examples to illustrate the various ways in which multimodal communicative actions can be ill-formed.

#### 2.2.1.1 Underspecified Form-Meaning Mapping. Underspecified Semantic Relation between Aligned Speech and Gesture

The mutual "partnership" [Kendon, 2004] of speech and gesture in conveying a single meaning dovetails with the fact that the form of each signal reveals something about content. Gesture, along with speech, makes a semantic contribution to the final logical form of the multimodal communicative action. However, gesture form is often

[5]The video clip can be found here `http://www.talkbank.org/media/Gesture/Cassell/kimiko.mov`

[6]`http://corpus.amiproject.org/`

ambiguous even in the context of the utterance. For instance, recall from our experiment in Section 1.1 that whereas one annotator interpreted the gesture as the action of stacking books, another interpreted it as describing the shape and size of the boxes.

A standard approach for handling cases where even the disambiguated representation about form is insufficient to determine a complete interpretation or where form remains imprecise and vague with respect to the referent's 'granularity' is semantic underspecification—e.g., Alshawi [1992], Reyle [1993], Pinkal [1996] and Bunt [2007]. One of the central claims in this thesis is that the form-meaning mapping of gesture can be expressed by an *Underspecified Logical Form* (ULF) which captures the incomplete meaning derived from gesture form by yielding an abstract and partial representation of what it can mean in the context-of-use. The ULFs are resolvable to specific values by augmenting them with contextual information, including the content of the aligned speech signal. Essentially, the logical forms produced by gesture are only partial descriptions supporting the possible interpretations in context. A rough equivalent would be the semantic composition produced by shallow NLP techniques such as part-of-speech taggers which often underspecify information about syntactic constituency and semantic arity. Our programmatic approach consists in mapping gestural form to a partial and underspecified meaning representation that can be resolved at the pragmatics interface. By using a framework that is able to express underspecified semantics, the grammar will output an abstraction over the possible interpretations—in (1.1), for instance, a meaning predication will underspecify its main argument (and hence it could resolve to either an event or an individual) and its arity.

Following Lascarides and Stone [2009b], we assume that resolving the meaning of the gesture to a specific value and computing the rhetorical connections between a gesture and its corresponding speech phrase are logically co-dependent. In other words, the denotation of a depicting gesture and the denotation of the speech phrase are linked through an inventory of rhetorical relations which are inferable using the discourse context and commonsense reasoning [Lascarides and Stone, 2009b]. For instance, utterance (2.4) illustrates one possible relation: the gesture provides an Explanation of what is said in the co-occurring speech (the fact that the bottom cupboards were straight, as depicted through the gesture, explains why they worked OK, as expressed in speech). The set of rhetorical relations also include Depiction (the gesture provides a visual depiction of what is said in speech), Elaboration (the gesture provides more specific content relative to the accompanying speech signal), Narration (the gesture continues narrating a story as delivered in speech) and others. Lascarides

and Stone [2009b] argue that certain relations are not allowed — gesture can reveal the same content as speech, it can underspecify speech content, it can also contribute content to the speech content, but it cannot *contrast* it. There can, on the other hand, be a Contrast relation between two gestures and even between a gesture and a prior discourse segment that has some speech as part. This is our formal rendition of how speech and gesture combine together to convey a "single thought". Since the grammar aims to produce an underspecified relation between speech and gesture, but not to resolve it, we forgo any more details about the rhetorical relations between speech and gesture. An analysis of them can be found in Lascarides and Stone [2009a].

We now turn to utterance (1.8), repeated in (2.5), which demonstrates why inferring the rhetorical relation between speech and depicting gesture is dependent on resolving gesture's underspecified meaning.

(2.5) He used to go down there and throw . . . [$X^*$grou]nd rice over it.
*The speaker moves his right hand forward; fingers are flexed inward in contact with the palm; the tip of the thumb is resting on the first joint of the index finger. The hand is moved twice by extending the wrist. The gesture resembles scattering a handful of dust/powder over some surface.*

Kendon [2004] interpreted the gesture stroke temporally overlapping with the verb as denoting some salient feature of the act of throwing, namely throwing some small particles (like dust, rice) over an extended surface. Interpreting the gesture in this way suggests an inference where the hand movement of scattering small particles *elaborates* the throwing motion. In this case, we establish an Elaboration relation between speech and gesture. An alternative interpretation where the gesture movement *depicts* the manner of throwing rice supports a different relation—a relationship that is captured via Depiction. Of course, many more interpretations can also arise in context.

Deictic gestures also display ambiguity with respect to the way they can denote distinct features of the qualia structure [Pustejovsky, 1995] of the referent, and so the gesture relates through a range of relations with the various roles of polysemous words. An example from Clark [1996, p. 168] illustrates this: George points at a copy of Wallace Stegner's novel *Angle of Repose* and says:

1. "That book is mine."
2. "That man was a friend of mine."
3. "I find that period of American history fascinating."

While in all contexts the pointing signal denotes the physical object book, the semantic relation between the speech and gesture changes depending on the specific meaning of the speech NP: in 1., the speech denotes the physical artefact book, and so does the gesture. Given that we can infer an Identity relation between the speech denotation and the deixis denotation. In 2., there is a reference transfer from the book to the author, and so we can infer an AgentiveRelation between the deixis and the speech NP. In 3., the reference transfer is from the book to the author, and so we can infer a ContentRelation between speech and deixis. Likewise, it is anomalous to relate the speech and deixis via Identity in "That is a beautiful city" while pointing at a poster of Florence. In this instance, the denotation of "that" (i.e., Florence) is not identical to the denotation of the deixis (the physical object identified by the pointing gesture—the poster), but the two signals are rather related through Depiction (Alex Lascarides, personal communication).

In Section 1.3.2.4 we argued that the construction rules for speech-and-gesture alignment introduce a relation $R(s,g)$ between the speech $s$ content and gesture $g$ content, which designates that $s$ and $g$ are coherently related. To account for the distinct ways depicting vs. deictic gestures coherently relate with speech, we now refine this relation to an underspecified visualising relation $vis\_rel(s,g)$ between the content denoted by speech $s$ and the content denoted by the depicting gesture $g$, and an underspecified deictic relation $deictic\_rel(s,g)$ between the content $s$ of speech and the content $g$ of deixis. How this relation is going to resolve is a matter of discourse context and commonsense-reasoning: for instance, the possible resolutions of *vis_rel* include Depiction, Elaboration, Narration or Explanation (but not Contrast) and of $deictic\_rel(s,g)$—Identity, MetaphoricalCounterpart [Lascarides and Stone, 2009b]. This approach 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 (recall Section 1.3.2.4).

#### 2.2.1.2 Gestural "Syntactic" Ambiguity

We assume that the different gestural dimensions correspond to distinct "syntactic" categories—depicting vs. deictic—which map to distinct ULFs: depicting gestures adopt a *qualitative* category in that they are represented through qualitative, non-spatial features that produce (underspecified) predications. For instance, the form of the gesture in (2.5) can be represented through a range of features that characterise the shape of the hand, the orientation of its palms and fingers and the movement performed

Figure 13: Hand Gesture Pointing at a Landmark in the Virtual Space, example (2.6)

during the gesture, and that are perceived as resembling its reference in context: the motion of the loosely held hand resembles the action of scattering small particles over an extended surface. In contrast, deictic gestures are at heart *quantitative*: they provide spatial reference as determined by the spatial coordinates of the hand [Kranstedt et al., 2006; Lascarides and Stone, 2009b].

With this in mind, the "syntactic" ambiguity of gestures is a matter of assigning distinct gesture categories to the same gesture form. To illustrate this, consider the multimodal action in (2.6), Figure 13, extracted from the Talkbank corpus and the constructed utterance in (2.7). The form of the gesture outside of context is ambiguous as to whether it is a depicting gesture, and hence it should be analysed in terms of qualitative values, or it is deictic, and hence it should receive quantitative values.

(2.6) I [$_N$turn] [$_{PN}$left] on [$_N$Elm] Street...
*Speaker's right arm is extended in the central space with the palm slightly bent and fingers pointing left.*

(2.7) All edges of Gaudí's Casa Batlló are curved.
*Same gesture as in (2.6)*

Gesture's syntactic form can be disambiguated only in the context of speech: for instance, accompanied by the speech signal in (2.6), the gesture receives quantitative values, and in the context of (2.7) – qualitative ones. Disambiguating gesture "syntactic" category is essential for its interpretation: namely, a gesture of qualitative values is interpreted as providing non-spatial descriptive aspects of its denotation, and so its form bears a close resemblance to what it means in context. Conversely, a gesture of quantitative values is interpreted as identifying the spatial coordinates of some referent in the physical space. 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 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".

#### 2.2.1.3 The Form of Gesture. Underspecified Semantics

One of the main challenges addressed in this thesis concerns the *ambiguous gesture form* which potentially maps to open-ended meanings. In Chapter 1, we defined gesture form in terms of the physical shape and the movement performed during the gesture. We now refine the notion of form to encompass the following two axis:

1. **Gesture Form:** By form, we understand the non-arbitrary arrangement of the distinct aspects of gesture that our perceptual system recognises as pertaining to its visual characteristics: the shape of the hand (for instance, open flat, fist, bent finger(s), extended finger(s)), the orientation of the palm and fingers (for instance, forward, to the torso, to the left or right periphery), location of the hand (centre, low centre, high centre, etc.) and movement (straight forward, straight down, straight left, etc.). In line with previous work [Kopp, Tepper, and Cassell, 2004; Lascarides and Stone, 2006; Hahn and Rieser, 2010], in Chapter 5.2.1 we argue for formalising gesture form in terms of feature-value pairs. The specific values are based on the gesture type hierarchy, proposed in Section 6.3.

2. **Attachment in a Derivation Tree:** In Chapter 1, we argued for aligning speech and gesture in a single derivation tree on the grounds that this treatment in syntax would allow us to use standard methods for semantic composition to build the target meaning representation. We formally capture the speech-gesture alignment through an attachment in a single derivation tree that maps to an under-

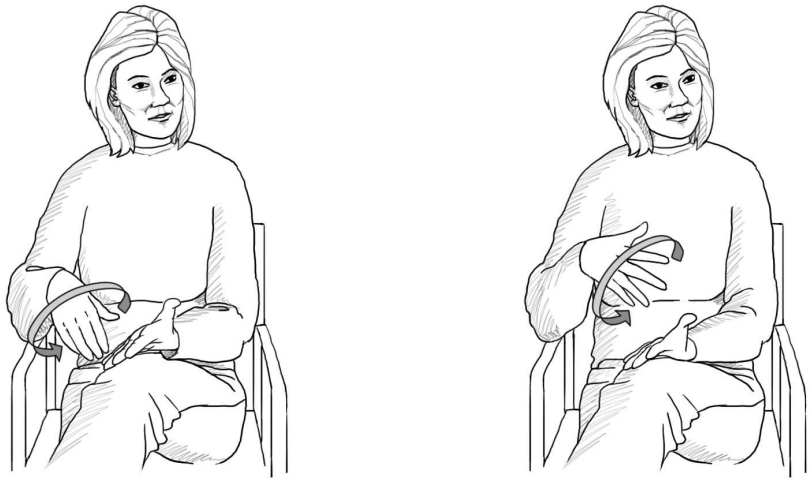


Figure 14: Gesture Depicting Mixing Mud, example (2.8) [Loehr, 2004]

specified semantics. Also, given our earlier claim that the alignment contributes the underspecified relation $vis\_rel(s,g)$ or $deictic\_rel(s,g)$ to the logical form, the semantic component of the construction rule that attaches speech to gesture will add this predication to the logical form of the combined speech and gesture constituent.

**Ambiguity in Gesture Form.** The ambiguity in form has as its consequence that the same gesture can map to different predications, and these predications are not necessarily of unique arity. This ambiguity persists within the same context and also across contexts. This also confirms our claim that the form-meaning mapping of gesture signals is, in fact, open-ended.

To illustrate how the same gesture maps to predications of different arities in distinct contexts, consider example (2.8), Figure 14 [Loehr, 2004].

(2.8) So $[_{H^*}$he mix]es $[_{X^*}$mud] ...
*Speaker's left hand is rested on the knee with palm open supine. The right hand is held loose with fingers facing downwards over the left hand. The speaker performs consecutively four rotation movements with her right hand over the left palm.*

The utterance is extracted from a longer narrative where the speaker described the

renovation of a house with a drywall. The circular hand movement can be interpreted as a visual depiction of the mixing event *e* performed by some agent *x* (from the character's viewpoint it could be the speaker herself) over the object *y*. In this case, gesture form would contribute the three-place predicate $mix(e,x,y)$. It is perfectly acceptable for the same gesture to appear in a completely different context such as the constructed utterance in (2.9) where the hand movement depicts a salient feature of the staircase, namely, the fact that it is spiral. Thus here the hand shape and movement resolves to the one-place predicate $spiral(x)$ where *x* denotes the stairs.

(2.9) She descended the spiral staircase
*Same gesture as in (2.8)*

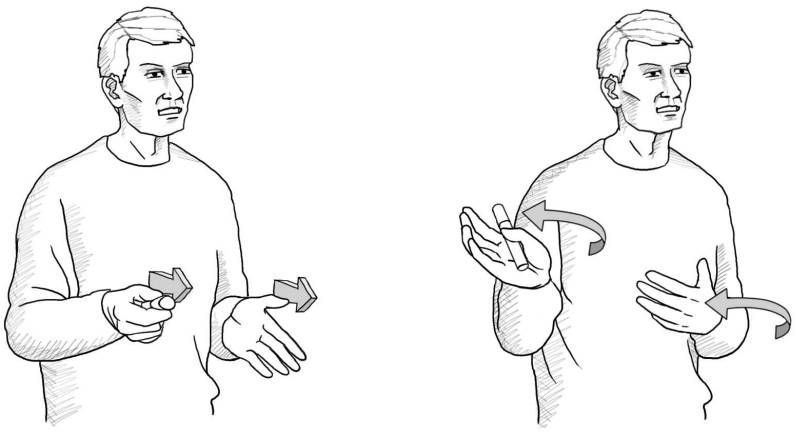
Further, we have observed that even within the same speech context, the disambiguated gesture form under-determines meaning, and so a single gesture can map to predications, which are not necessarily of the same arity. To illustrate this, consider utterance (2.10), Figure 15[7] taken from a university lecture on Cognitive Science where the speaker, a university professor, discusses the course literature.

(2.10) I can give you [$_N$other] books that would totally trash experimentalism.
*Both hands are in parallel with palms open vertical. They perform a short forward move to the frontal centre. The same hand shape is used in the second stroke, but here the hands move from the centre to the right periphery.*

A possible denotation of the parallel placement of the hands in Figure 15(a) is a container 'containing' experimentalism or a containee of books about experimentalism and so, both denotations would yield a one-place predicate $books(x)$. Moreover, the hand shape, combined with the forward direction of the movement can denote the conduit metaphor [Lakoff and Johnson, 1980] of the act of giving some books to the audience, in which case, the vertical palm form would contribute a 4-place predicate $give(e,x,y,z)$ where *x* refers to the agent of giving books *y* to someone *z*.

The conduit gesture in Figure 15(a) is anaphorically related to the subsequent movement of throwing away, Figure 15(b): both hands appear in the same container–containee position when shifting towards the right. This is also in line with the previous discussion of example (1.5) where the speaker responded to a clarification request

[7] The video clip is available at `http://www.talkbank.org/media/ClassBank/Lecture-unlinked/feb07/feb07-1.mov`: *00:03:41–00:03:43*



(a) I can give you other books... (b) ...that would totally trash experimentalism

Figure 15: Gesture Depicting (the Event of Giving) Books, example (2.10)

by repeating the entire multimodal action. These examples show that like purely linguistic discourse, gestures have the potential of dynamically changing the context in which subsequent multimodal actions are interpreted. Even though this falls outside the scope of multimodal grammar development, these examples confirm our starting assumptions that gestures follow the same principles as language and so the formal apparatus for analysing gesture can be reliably drawn from linguistic theory.

The form of the deictic signal as well displays imprecision with respect to the region pointed out by the hand. For instance, when pointing in the direction of a book with an extended index finger, does the deictic gesture identify the physical object book, the location of the book—e.g., the table—or the cover of the book? Often there is not an exact correspondence between the region identified by the pointing hand, the so called 'pointing cone' [Kranstedt et al., 2006] and the reference. Our formal model does not intend to solve this imprecision of reference since it has no effects on multimodal perception. Certain imprecisions in the interpretation of deixis remain unresolved even in context, just as certain imprecisions can be tolerated in purely linguistic utterances.

Based on Lascarides and Stone [2009b], we formalise the location of the pointing hand with the constant $\vec{c}$, that marks the physical location of the tip of the index finger. This combines with the hand shape, orientation and movement to determine the region $\vec{p}$ actually marked by the gesture—for instance, a stationary stroke with an extended

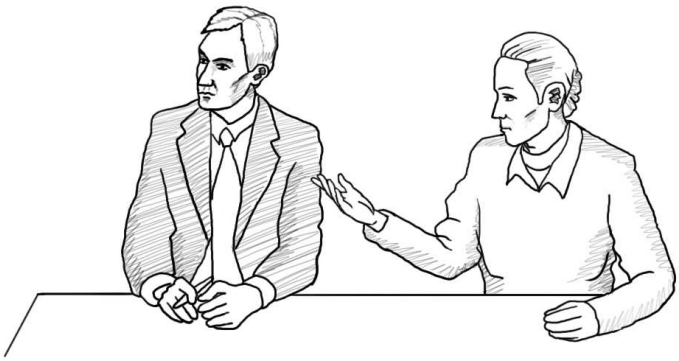


Figure 16: Hand Gesture Pointing at an Individual in the Communicative Event, example (2.11)

index finger will make $\vec{p}$ a line (or even a cone) that starts at $\vec{c}$ and continues in the direction of the index finger. Often gestures identify referents that are not salient in the communicative situation — this is the case with abstract deixis where the hands places individuals and/or events on a virtually created map in the frontal space. To account for this inequality between the gestured space and actual denotation, Lascarides and Stone [2009b] use the function $v$ to map the physical space $\vec{p}$ designated by the gesture to the space $v(\vec{p})$ it denotes. We illustrate this with example (2.11), Figure 16 extracted from the AMI corpus: here the speaker was discussing that the cost of living in the East was lower than in the West. Then along with "You..." she extended her right hand towards the other interlocutor, who presumably was from the tropics. In this instance, the referent introduced by the hand is at the exact coordinates in the visible space the gesture points at and therefore the function $v$ resolves to equality.

(2.11) ... [$_N$You] guys come from tropical [$_{NN}$countries]

*The speaker turns to the right towards the other participant pointing at him using her right hand with palm loosely open up.*

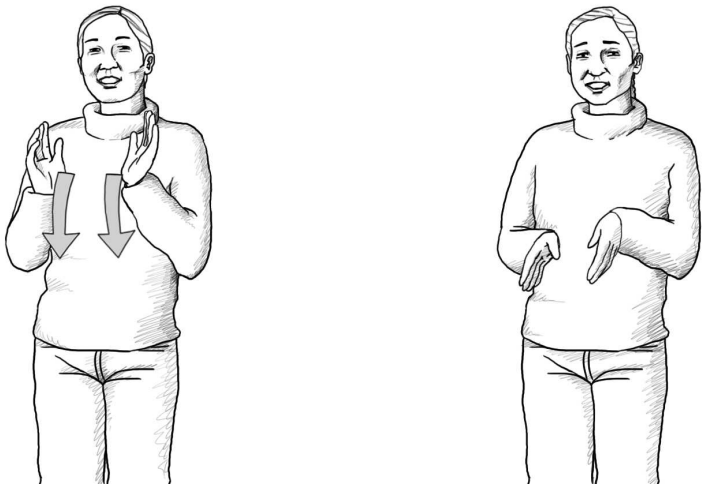


Figure 17: Hand Gesture Placing a Virtual Apartment in the Frontal Space, example (2.12)

By contrast, in (2.12), Figure 17, extracted from the Talkbank conversation where the speaker described the layout of her flat, *v* would *not* resolve to equality since the referent for "apartment" was not physically present at the coordinates identified by the pointing hand.

(2.12) 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.*

**Attachment Ambiguity.** The form of the gesture, the form of speech and their relative timing do not uniquely determine the speech phrase gesture aligns with. Similarly to "John saw the man with the telescope", gestures exhibit a non-unique choice of attachment to the speech phrase which reflects the fact that they can interpreted in various ways. We will illustrate this using example (1.6) [Loehr, 2004], repeated in (2.13).

(2.13) If I was to [$_{PN}$rea]lly [$_{N}$teach] someone how to be a professional musician

*Hands are open flat, relaxed with palms facing up; they move to the frontal space to possibly denote a conduit metaphor.*

Recall from Section 1.3.1 that we assume that the decision of what speech content gesture relates to determines the discourse parts that can serve as antecedents for resolving the semantic values of the gestural parts. This decision is driven by the speech-gesture alignment: for instance, an attachment of the gesture to the entire clause would support a semantic relationship between the content of the gesture and the content of the clause, in which way, the agent, patient and the idea transferred between them via teaching can all be accessed for resolving the values of the participants in the conduit metaphor [Lakoff and Johnson, 1980] that is expressed by the gesture. The alternative interpretation where the gesture emphasises "really" would be supported by an attachment to the pre-head modifier, and so "teach" would not feature in the logical form for resolving gesture. This attachment, however, would not support the conduit interpretation.

As we already observed in Chapter 1, interpreting the gesture in (2.13) as the conduit metaphor is possible only after linking the gesture with the speech head-daughter "teach" although the contentful part of the gesture, the stroke, is performed while uttering the adverbial pre-head modifier "really". We shall come back to this point in Chapter 4 where we argue that quantitative criteria—such as the timing of speech relative to gesture—are not sufficient on their own to define adequate constraints on speech-gesture alignment.

Attachment ambiguity is also observed in deictic dimensions. In utterance (2.12), for instance, there is no information coming from the form of the hand, nor from its relative timing that it should attach to "enter" only, and not to "enter my apartment", in which case the form of the hand would be related to the rectangular shape of, say, an entrance door to an apartment. Intuitively in this case, the gesture directs not only to the point of entering the apartment, but also to the entrance door which by the hand shape is rectangular. As discussed in Section 1.3.1, the constraints that theories of discourse coherence impose on which entities in the context are available for further reference would block the door from being a part of the meaning of the gesture if it were related only to the content of the word "enter". Such a discourse structure does not make the door available — the gesture is not coherently connected to a semantic representation that features $door(x)$.

To recapitulate, our formal model fully supports ambiguity and partial meaning in that we map form to an underspecified, highly factorised meaning representation whose main variable can resolve to either an event $e$ or to an individual $x$ in context. The underspecified predicate-argument structure characterises the main difference be-

tween how gesture form maps to meaning and verbal input maps to meaning.

### 2.2.2 Not a Free-for-All

Despite the syntactic and semantic ambiguities, in Chapter 1 we claimed that the different ways gesture can align with speech are not unrestricted. Instead, the alignment between speech and gesture is constrained by the form (that is, the prosody) of the linguistic signal. We use the term multimodal ill-formedness to refer to those speech and gesture ensembles that can never happen in everyday language use, based on an extensive study of real data. To illustrate the effects of prosody on speech-gesture alignment, consider again the constructed example (1.4), repeated in (2.14).

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

We stated that the performance of the gesture along with the unaccented "called" seems anomalous despite the semantic relation between the verb and the gesture. Intuitively in this case, the gesture should be performed while uttering the accented "mother" or any phrase containing "mother". Note also that even by attaching the gesture to the subject daughter "your mother", we can still arrive at an interpretation where the gesture depicting the calling event continues the denotation of the speech phrase, and so they can be related through Continuation.

This example illustrates that despite the multiple syntactic attachments and semantic interpretations licensed by gestural ambiguity, speech-and-gesture alignment is not unconstrained. We shall therefore equip our grammar with construction rules that not only constrain the choices of attachment, but that also account for the highly flexible and open-ended gestural interpretations. There is no contradiction between both as an overall challenge is to find the right balance between constraining the mapping from form to meaning on the one hand while determining the resolutions to the range of possible interpretations in discourse context on the other.

## 2.3 Summary and Next Steps

In this chapter we provided the necessary background information to understand the scope of this thesis and the main challenges we shall be dealing with. This thesis studies speech-accompanying gestures produced spontaneously with the hand. Their

main property is that their ambiguous form underspecifies their meaning, and so they can only be understood by aligning them with the speech phrase and by reasoning about the interpretations licensed by this alignment. This is unlike emblems whose form alone maps to a conventionalised meaning.

By drawing on both constructed and naturally occurring examples, we demonstrated the range of gestural ambiguities. We first demonstrated that gestures *underspecify their relation* to speech and that resolving this relation is logically co-dependent on resolving their meaning in context. The grammar captures this relation by introducing in semantics an underspecified relation between the speech content and the gesture content. The resolution of this relation is dependent on the gestural interpretation assigned via discourse processing. The next level of ambiguity concerns the *gestural "syntactic" ambiguity*—depicting or deictic—which affects the form-meaning mapping in context. Depicting gestures are at heart qualitative in that their form resembles their meaning, and deictic gestures are quantitative in that they demarcate the spatio-temporal coordinates in the physical space [Lascarides and Stone, 2009b]. We also introduced *ambiguity in form* to reflect the fact that the same hand movement can yield predications of different arity within the same context or across contexts. Finally, we used the notion of *attachment ambiguity* to account for the distinct alignment configurations between speech and gesture: similarly to attachment ambiguities in natural language, gestures exhibit a choice with respect to the phrase they can align with. We claim that there is no need to curb the choices of attachment as the overall aim of the grammar is to produce underspecified logical forms that support the various ways gestures can be interpreted in context. Lastly, we used a constructed example to demonstrate that despite gesture ambiguities, the speech and gesture alignments are not unrestricted where the constraints come from the prosody of the temporally overlapping speech phrase.

Now that we have laid out the scope of this thesis, in Chapter 3 we discuss the related literature thereby identifying certain gaps that this work intends to fill.

# Chapter 3

# Related Literature

*. . . cum sit autem omnis actio, ut dixi, in duas divisa partis, vocem gestumque, quorum alter oculos, altera aures movet, per quos duos sensus omnis ad animum penetrat adfectus. . .* [1]

[Quintilianus, 1992, Book XI, Chapter 3, line 14]

The literature on gesture from the $20^{th}$ century is unanimous in its observation that speech and gesture function within a single communicative system with the view of delivering an integrated message. This message is realised through semantically related spoken words and hand movements. The descriptive, cognitive and computational approaches to multimodal interaction use this finding to analyse the semantic interaction of the temporally co-occurring speech signal and gesture signal, to provide models of the pre-verbal processes underlying the speech and gesture production, and also to build systems for multimodal human-computer interaction. Whereas these studies provided insight into the domain-specific use of multimodal communicative actions, we intend to further explore the speech-and-gesture interaction in the following ways: first, while the descriptive studies of gesture describe how the speech signal is semantically related to the gesture signal, we shall refine this finding by providing a formal model of multimodal well-formedness and multimodal ill-formedness, which, by drawing on standard methods from linguistics, explores the form of the gesture, the form of the speech signal and their mode of combination. Second, while the descriptive studies use recorded meetings of a single domain to describe how gesture is related to speech, we use multimodal corpora of distinct domains to provide quantified

[1] "All delivery, as I have already said, is concerned with two different things, namely, voice and gesture, of which the one appeals to the eye and the other to the ear, the two senses by which all emotion reaches the soul."

evidence for the speech-gesture interaction. In so doing, we demonstrate that quantitative and qualitative evidence is of importance for analysing multimodal signals. Third, a considerable amount of the computational approaches to gesture formally model its relation to speech. However, this body of research fails to address the effects of the *form* of the linguistic input on the speech-gesture alignment. In particular, there does not yet exist a formal analysis of the *ambiguity* of gesture that models the one-to-many form-meaning mappings in terms of *non-unique* speech-and-gesture alignments. Another under-researched area is also the *domain-independent* multimodal interaction. This thesis fills these gaps by contributing a formally precise and domain-independent *form-meaning mapping* of multimodal communicative actions based on the form of the linguistic input, the (ambiguous) form of the gesture and their relative timings.

This chapter sheds light on the existing findings related to the speech-gesture interaction from a descriptive, cognitive and computational perspective. The goal of this chapter is on the one hand to demonstrate how our research fits in the broader context of gesture studies, and on the other to strengthen its contribution by spelling out the under-researched areas. The chapter is structured as follows: we begin in Section 3.1 with a brief historical overview of the first written accounts of gesture. In Section 3.2 we review the descriptive studies of speech and gesture integration. Then in Section 3.3 we proceed with an overview of gesture production from a cognitive perspective. Section 3.4 discusses the integration of gesture in multimodal systems for human-computer interaction. Finally in Section 3.5, we present the existing formal approaches to multimodal syntax. Each presented study is intertwined with a discussion of how this thesis addresses the weaknesses and/or strengths of prior work.

## 3.1 First Accounts: Historical Notes

One of the first written accounts of the role of gesture in the spoken delivery is found in Quintilian's treatise of the art of oration, *Institutio Oratoria*, written in the first century AD [Quintilianus, 1992]. The treatise offers a prescriptive account of how to increase the expressive power of the rhetorical discourse by properly employing visually perceived signals such as nodding, glance and hand movements.

Further studies from the Renaissance and the $18^{th}$ century followed the prescriptive paradigm, proposing how gestural manners were indicative of people's social conduct and expressivity in public speeches [Bulwer, 1644].

However, it is not until the 1970s when the first descriptive studies of the integrated

nature of speech and gesture emerged. Independently from each other, Kendon [1972] and McNeill [1979] observed that the utterance is an 'ensemble' composed of speech and gesture—roughly, this means that a single thought is materialised in speech and in gesture—thereby establishing the theoretical framework for studying gestures as part of language.

## 3.2 Speech-Gesture Integration: a descriptive account

The most significant contributions to the gesture studies of $20^{th}$ century were made by Kendon [1980] and McNeill [1992]. They advanced a new descriptive framework for understanding the speech-and-gesture interaction. Through studying naturally occurring conversations, they described gesture in its relation to speech, and more specifically how hand movements are deliberately expressed in synchrony with verbal signals to convey a "single thought".

In this section, we discuss the descriptive work on speech-gesture integration that serves as a starting point for the present study from various perspectives. We shall now describe each of them in turn.

### 3.2.1 Integrated Message of Spoken and Gestural Material

In McNeill's Growth Point theory, the communicative act is viewed as a single system with the gestural manifestation being a byproduct of the vocal delivery [McNeill, 1992; Duncan and McNeill, 2000]. The growth point is a minimal psychological unit—a seed for the utterance—that combines two unlike cognitive modes in the 'thinking-for-speaking' process (term coined by Slobin [1996]) — linguistic thinking and imagistic thinking. The growth of this unit creates an 'imagery-language dialectic' which presupposes an opposition between the two cognitive processes and which hence seeks resolution in stable constructs. The unit gets resolved by unpacking the linguistic thinking into well-formed linguistic units and the imagistic thinking into gesture strokes, thereby producing a complex inseparable multimodal entity.

The dialectic entails the same point of origin of both modalities: the imagistic thinking does not precede or follow the linguistic thinking but rather both cognitive modes are simultaneously initialised. Further to this, the Growth Point theory suggests that speech and gesture are semantically and pragmatically co-expressive within the same temporal interval: preparations happen when forming the growth point, strokes

co-occur with the articulation of the linguistic unit that advances the communication, post-stroke holds entail that the articulation of speech takes longer than the articulation via hand signals. For McNeill [1992], the tightly synchronised nature of the speech-and-gesture actions does not imply that the two modes convey the same meaning. Instead, gesture and speech are two sources for describing *different properties* of the same underlying idea: for instance, while spoken words introduce referents in the discourse, gestures often augment the spoken words with spatial and visual information.

This is the primary finding that the current study is based on: we view the communicative action as composed of spoken and gestural material where the two modalities are connected through a range of semantic relations, similarly to connecting units of discourse. The connecting relation can be redundancy; for instance, Lascarides and Stone [2009b] observed that unlike purely linguistic discourse, redundancy across speech and gesture does not have pragmatically marked effects, and thus communicating the same idea in speech and in gesture does not violate the cooperative principle of Quantity [Grice, 1975]. The gesture can also provide a complementary contribution relative to that of speech. Our goal is thus to provide the mechanisms for producing the adequate range of semantic relations between speech and gesture.

In contrast to McNeill [2005] where speech and gesture co-expressivity is based on temporal simultaneity, we advance a different approach to establishing the semantic relatedness of the spoken signal and the gesture signal. While we consider the temporal performance of gesture relative to the temporal performance of speech a guiding factor, we go deeper by investigating the linguistic properties of the spoken signal—for instance, its syntax and prosody—before deciding on semantic co-expressivity. In so doing, we ensure that the semantic representation produced by the form of the multimodal signal is compatible with constraints on meaning and reference that are captured in current models of the semantics/pragmatics interface—an issue that McNeill does not account for. As we said in Section 1.3.1, these models make meaning and reference sensitive to discourse structure, which in turn is determined by the compositional semantics of the discourse units, and hence their linguistic form. With all that in mind, our contribution is two-fold: we not only cover the range of semantic relations of speech and gesture based on quantitative and qualitative evidence but we also exploit how gesture relates with the linguistic properties of the speech signal thereby providing more insight into the inseparable nature of the two modalities.

### 3.2.2 Gestures as a Global Performance Dependent on Context

In the Growth Point theory, gestures are 'global' affordances since the meaning of a gesture cannot be determined by decomposing it into smaller units of content [McNeill, 2005]. Contrary to the decompositional analysis of lexical items or the semantic compositional approach to natural language phrases, where the meaning of the whole is a function of the meanings of its parts, the meaning of a gesture stroke cannot be derived compositionally. Rather, the meaning of the gesture is obtained in a top-down direction by linking the gesture form features that contribute distinct aspects of its meaning. For instance, the palm open vertical hand shape in (2.10), Figure 15 does not necessarily mean a visual description of the event of giving books or of the books themselves. It is only by linking the position of the hands and the orientation of the palms and finger to the specific speech context that this gesture can be interpreted in this way. Various other interpretations of this hand shape can arise in distinct speech contexts.

And yet, the global nature of gesture does not preclude relating distinct gestures to one another via a hierarchy of semantically related units. We believe that distinct gesture performances can be hierarchically organised, just as segments in linguistic discourse can be. For instance, the second gesture in utterance (2.10) can be subordinately related via Elaboration to the first gesture: that is, the brisk hand movement to the right periphery elaborates on the referent books introduced by the open hands in the first gesture signal. Relating them in this way creates a complex gesture performance, consisting of two gesture strokes connected via Elaboration. Representing the gesture sequence in this way helps for both constructing the representation of meaning in context, and for evaluating that meaning. This is a fundamental feature of models of discourse that are based on coherence (e.g., Kehler [2002], Asher and Lascarides [2003]).

In Section 1.1.2, we stated that the global, non-hierarchical nature of gesture [McNeill, 2005] requires a different form-meaning representation compared to the form-meaning representation of unimodal linguistic input. Capturing gesture form in terms of hierarchical syntax trees and then using this tree to compose the gesture semantics is thus not adequate for gesture. In line with previous work [Kopp, Tepper, and Cassell, 2004; Lascarides and Stone, 2006; Hahn and Rieser, 2010], we shall use a flat format—namely feature structure descriptions—for capturing the various aspects of gesture such as the hand shape, the hand location, palm and finger orientation and hand movement. We postpone a detailed discussion about modelling gesture form to

| Gesture system | Intonation system |
|---|---|
| Stroke | Pitch accent |
| Gesture phrase | Prosodic phrase |
| Gesture unit | Locution |
| Consistent head movement | Locution group |
| Consistent hand and arm movement | Locution cluster |

Table 2: Relationship between Gesture Units and Tone Units [Kendon, 1972]

Section 5.2.

### 3.2.3 Relationship between Gesture and Intonation

There has been a rising number of studies investigating the temporal relation between the unfolding of gesture and the unfolding of the prosodic pattern in speech. We shall now mention a few that formed the basis of our corpora investigation, discussed at length in Chapter 4.

Analysing a recording of one conversation, Kendon [1972] observed a strong relationship between units of speech and units of bodily movements where distinct tone units co-occurred with specific units of bodily movements. This relationship, illustrated in Table 2, appeared from the minimal to the maximal units of speech and gesture as follows: the minimal unit in the spoken system, the pitch accent, co-occurred with the minimal unit in the gesture system, the stroke. Then the hand excursion from lifting the hands to an expressive position to reaching the expressive peak (also referred to as “gesture phrase” in Kendon [1972; 2004]) corresponded to a prosodic phrase in speech. In Kendon’s [1972] terminology, the sequence of gesture phrases constitutes a gesture unit. Its beginning and ending are marked by the highest relaxation points reached by the hands before the gestural execution and after the hands’ retraction to rest. The gesture unit seemed to co-incide with a locution which syntactically maps to a complete sentence. Several gestural units which have one distinctive feature throughout their performance (such as a consistent head movement) could be grouped together. At the level of speech, this gestural group mapped to a ‘locution group’. At the highest level of communication, Kendon [1972] positioned the ‘locution cluster’ which roughly corresponded to a paragraph. This level was marked by consistent body and

| Co-occurrences within 275 msec | Pitch accent | Other (than a pitch accent) |
|---|---|---|
| Gesture apex | 83 | 28 |
| Other (than an apex) | 43 | 46 |

Table 3: Co-occurrence Between Gesture Apex and Pitch Accent [Loehr, 2004]

arm movements. It might include thematic shift which was usually marked by posture shift.

While this analysis contributes to our understanding about the interaction between intonation of speech and body performance at various levels, at this stage we are not familiar with an empirical study that tested this multi-layer mappings over a larger set of recorded conversations.

Driven by the interaction between the gesture organisation and the intonation hierarchy, Bolinger viewed intonation and gesture as "a single form in two guises, one visible and the other audible" [Bolinger, 1986, p. 199]. Using constructed examples, Bolinger [1986] articulated this interaction in the parallel hypothesis which stated that a rising pitch co-incided with a rising head, hand or finger movement and the other way around: a lower pitch would happen along a lowering body movement. It is however not until the studies of McClave [1991] and Loehr [2004] when this hypothesis was tested against empirical data. The results of these studies did not find reliable evidence in support of the parallel hypothesis. Loehr [2004] reported that the H* high pitch tone occurred with 47 upward hand movements and with 62 downward hand movements, and also the L% boundary tone occurred 35 times with a head moving in an upward direction and 27 times in a downward direction. These results indicate that the speakers' intuitions about gestures are often erroneous, and so empirical studies are necessary for drawing reliable generalisations.

A major contribution to the relationship between gesture and intonation was provided in the doctoral dissertation of Loehr [2004] who used labelled corpus data to extract statistical generalisations about the co-occurrence between intonation units and body movements. Some of his findings are relevant to the aims of this thesis, namely:

- The minimal units in the gesture system tend to temporally co-occur with the minimal units of the speech system (see Table 3). The highest peak of effort in the gesture carrying the meaning (in the nomenclature of Loehr [2004], this is the 'apex') typically temporally overlaps with the pitch accent in speech. The

number of 275 msec was empirically established as this was the standard deviation of the distribution of tones near gestures.

- In over two-thirds of the gesture phrases in the analysed data, Loehr [2004] found a temporal co-occurrence between the gesture phrase and the intermediate phrase.

Further evidence in the literature suggests that the temporal co-occurrence of speech and gesture affects perception of multimodal well-formedness and of multimodal ill-formedness. Giorgolo and Verstraten [2008] conducted an experiment where they showed to subjects some misaligned audio-video recordings with a negative or positive delay of 250 msec, 500 msec, 750 msec, 1000 msec, and asked them to judge whether the clips were synchronous. The results showed that the specific delay had an effect on the final judgement: a difference of 250 msec (note that this approximates the average time it takes to say a word) was not significant for the preference, but differences of more than 500 msec were significant. The authors also isolated prosody as the parameter that influenced multimodal perception and multimodal integration.

The empirically validated studies of Loehr [2004] and Giorgolo and Verstraten [2008] provided statistical evidence for the interaction between speech and gesture at the level of *prosody*. For our purposes, these empirical studies are relevant because they suggest that prosody constrains the relative temporal performance of speech signals and gesture signals. We shall therefore use this evidence when setting up our own corpus investigation and also when defining constraints on the alignment of speech and gesture within the grammar.

Note also that the study of Giorgolo and Verstraten [2008] was a statistical analysis at the level of acceptance of multimodal deviation. We hypothesise that the results can also be related to the linguistic structure of the speech signal, namely that the perception of misaligned speech-and-gesture signals as well-formed did not include signals that crossed a constituency boundary. This also dovetails with the observation that gesture phrasing interacts with syntactic phrasing [Loehr, 2004].

### 3.2.4 Relationship between Gesture and Syntactic Constituency

Since we are interested in the relation between gesture and speech on the level of form, we searched for studies reporting on the interaction between gesture performance and syntactic constituency. McNeill [1992] analysed this relation from the perspective of the Communicative Dynamism (CD) which refers to the contribution of a spoken

message to the communicative action.[2] Considering the integrated nature of the multimodal action, the CD theory predicts that more elaborate linguistic phrases presuppose higher level of gesture accompaniment—that is, a heavier grammatical construct pushes the communication forward, and so it is more likely to happen along with a gesture. The example in (3.1) taken from McNeill [2005, p. 55] illustrates an elaborate NP that, by "breaking the continuity" [Givón, 1985] of information delivery, presupposes gesture accompaniment. There is no indication what the gesture in this particular case was.

(3.1) the next thing he did was ...

In McNeill [2005], the CD theory also functions in a reversed way: a syntactic gap does not indicate a communicative peak, and thus there is less likelihood that it would be accompanied by gesture. For instance, in utterance (3.2) [McNeill, 2005, p. 55], the empty subject does not contribute any novel piece of information, making gesture accompaniment less likely. Of course, this observation is reliant on a syntactic theory such as Government and Binding that features empty categories.

(3.2) He ran and ∅ got a bowling bowl and ∅ dropped it down the drainpipe.

Whereas McNeill [2005] was concerned with the relationship between the gestural accompaniment and the syntax of the temporally co-occurring phrase, we hypothesise that the gesture materialisation can be explained in terms of the relation of the utterance to the discourse model: whether to express an assertion, proposition or old/new information. More specifically, it has been previously attested that information structure, focus in particular, constrains the mapping of words onto prosodic structures: nuclear prominent nodes tend to align with foci [Calhoun, 2006]. There is also increasing evidence that gesture strokes overlap with prosodically prominent elements (see the empirical findings of Loehr [2004] and our empirical study in Chapter 4). We therefore hypothesise that there is a relationship between nuclear prominence and gesture on the one hand, and focus on the other. At this stage, we are familiar with one study that investigated whether gestures marked focus: Ebert, Evert, and Wilmes [2011] annotated nuclear accents and marked two types of focus: *new-information-focus (nf)* and *contrastive-focus (cf)*. Whereas the former included units that advanced the communication forward (for instance, (3.3)),[3] the latter was defined as an overt contrast with

[2] The notion of communicative dynamism was coined by Firbas [1992] as a property of a communicative signal to "push the communication forward".

[3] The examples in (3.3), (3.4) and (3.5) are from Dipper, Goetze, and Skopeteas [2007].

other units from the discourse (for instance, (3.4)). These two categories are however not mutually exclusive, and so a single segment can be both a new-information marker and a contrastive focus marker (for instance, (3.5)).

(3.3) $[_{nf}$Who] is reading a book?
$[_{nf}$Mary] is reading a book.

(3.4) $[_{cf1}$Mary] likes $[_{cf2}$apples] but $[_{cf1}$Bill] prefers $[_{cf2}$strawberries].

(3.5) $[_{nf}$What] are your sisters doing?
My $[_{cf1}$older] sister $[_{nf\,cf2}$works as a secretary], but my $[_{cf1}$younger] sister $[_{nf\,cf2}$is still going to school].

The findings of this study can be summarised as follows: 51 out of 276 gesture strokes did not overlap an accent at all, then among the remaining 225 strokes, the stroke started 36 msec earlier than the nuclear accent; the stroke started 31 msec earlier than the new-information focus and no tight correlation was found between strokes and contrastive foci.

This study is important since, to our knowledge, it is the first attempt to demonstrate, through statistical examination of labelled data, that gestures are a means of marking new-information focus. This thesis does not aim to study the interaction between focus and gesture. However, we intend to study the interaction between gesture performance and metrical prosodic structure which is central for marking information structure in English [Calhoun, 2006].

Further analysis of the interaction between gesture performance and syntax is found in Engle [2000]. The empirical material was a conversation about fixing locks, and the coding involved marking the spoken segments temporally co-occurring with gesture as either topically referential—the co-temporal speech referred to fixing locks—or topically non-referential—these included expressions with a meta-communicative function such as “right” or “so that”. Engle [2000] reported that 80% of the topically referential speech phrases formed syntactic constituents (noun phrases or verb phrases, for instance: “push out of the smaller cylinder”, “corresponding cotter pins” [Engle, 2000, p. 77]). The topically non-referential ones included pauses, expressions with a meta-communicative function such as “right”, “so that” and incomplete clauses that did not refer directly to locks such as “match up with those”, “the pins are”, etc. [Engle, 2000, p. 77]. With respect to the coding of the temporal span of gesture, Engle [2000] indicated that the coded gestures covered the hand movement in the interactional gesture

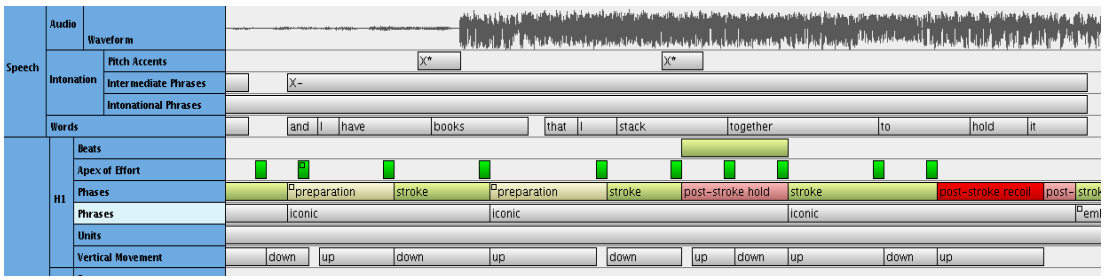



Figure 18: Temporal Alignment between Gesture Phrases and Syntactic Phrases [Loehr, 2004]

space, that is, the performance communicating meaning. Using the terminology of McNeill [2005], we can conclude that the gestures in this study considered only the gesture strokes.

Analysing the temporal alignment between gesture phrases and prosodic intermediate phrases, Loehr [2004] observed that the temporal span of gesture phrase boundaries co-incided with the syntactic boundaries, i.e., the gesture phrase did not extend beyond the syntactic boundary as in "... and I have books / that I stack together / to hold it..." where "/" designates a syntactic boundary [Loehr, 2004]. This has been illustrated in Figure 18 which presents an excerpt from the gesture coding and intonation coding in Anvil, as originally proposed by Loehr [2004]. In this instance, there are three boundaries of *iconic* gesture phrases that correspond to three syntactic boundaries. Strictly speaking, the correspondence is at the chunk level.

At this stage, we are not familiar with other studies of the relationship between the temporal unfolding of gesture and the syntax of the temporally co-occurring speech segment. The studies presented here are vastly inconclusive: they are either driven from intuitions supported by constructed examples or they are based on isolated examples. In any case, there is no statistically relevant evidence suggesting that gestures do in fact coordinate with syntax. This is one of the directions that we shall take in our empirical investigation in Chapter 4.

### 3.2.5 Multimodal Timing

The literature is not conclusive about the temporal performance of gesture relative to speech. Some studies suggest that speech-and-gesture co-expressivity can be defined in terms of simultaneity of both signals [McNeill, 2005]. Likewise, timing has been

proposed as a condition on multimodal integration [Engle, 2000]. However, there are studies suggesting that multimodality does not necessarily involve simultaneity; for instance, Oviatt [1999] reported only a 25% temporal overlap of 'speak and point' commands.

In Section 1.4, we stated that the controversy with respect to the multimodal timing is due to the lack of a clear methodology of how to establish the alignment of speech-and-gesture. For instance, what is the temporal span of the gesture that is semantically related with the speech signal: is it the expressive part of the gesture, the entire gesture excursion or maybe only the interval from lifting the hands to an expressive position to reaching the gesture peak? How do we establish which speech segments gesture aligns with? Is it only the segments temporally co-occurring with gesture, in which case we should expect syntactically incomplete speech segments to align with gesture? Or is it the segments that bear a semantic relationship with gesture, in which case we can expect a break of the one-to-one correspondence between alignment and synchrony? These are some of the questions that we address in the next chapters.

### 3.2.6 Summary

In this section we presented several descriptive approaches to multimodal actions, which unanimously demonstrate that speech and gesture interact to deliver an integrated message. We shall use this observation to demonstrate elsewhere that the same formal model can be used to analyse speech signals and gesture signals. In this way, we refine and extend the previous findings about speech and gesture interaction by applying the same techniques from linguistics for analysing gesture in parallel with speech.

## 3.3 Gesture from a Cognitive Perspective

The current project draws on the finding that speech and gesture are an integrated ensemble conveying a 'single thought' [McNeill, 2005]. In this section, we review some studies that provide cognitive evidence in support of the integrated nature of speech and gesture. A deeper understanding of the cognitive processes that underlie the production of an integrated signal should shed light on the following questions. First, does the cognitive organisation influence the form of gesture? Second, considering that the grammar provides a representation of what is stored in our minds, can we explain

the combination of speech and gesture through the cognitive processes that underlie their production? Finally, is there any cognitive evidence that relates gesture to the context of the synchronous speech?

The main debate in the cognitive approaches to multimodal communication revolves around the temporal onset of speech and gesture coordination: namely, does gesture parallel the pre-linguistic process of conceptualisation, or are speech and gesture two synchronous channels for conveying a message. This problem is closely related to the communicative function ascribed to gestures: are hand movements and facial expressions communicatively intended [Bavelas et al., 2002; Levelt and Melinger, 2004; Bavelas and Chovil, 2006] or do they participate in the conceptualisation process of information packaging and hence facilitate the speaker rather than the addressee [Hadar and Butterworth, 1997; Krauss and Hadar, 1999; Krauss, Chen, and Gottesman, 2000; De Ruiter, 2000]? This argument seems to be blended in McNeill [2005]:

> A gesture is a bridge from one's social interaction to one's individual cognition — it depends on the presence (real or imagined) of a social other and yet is a dynamic element in the individual's cognition McNeill [2005, p. 54].

In the next two sections we detail these two main approaches.

### 3.3.1 Gesture as a Product of Cognitive Processes

The cognitive processes that underlie the production of gestures have been articulated in several hypotheses.

The Lexical Retrieval Hypothesis argues that gestures are initiated pre-linguistically in the working memory to facilitate word retrieval, rather than being a meaningful part of the communicative act [Krauss and Hadar, 1999; Krauss, Chen, and Gottesman, 2000]. This hypothesis was underpinned by the observation that gestures depicting a clockwise rotation—for instance, a gesture describing the Gaelic word "deasil"—were performed from the speaker's perspective, that is, anticlockwise to the addressee and presumably, were not meant to facilitate the listener. This served as evidence for the authors to question the communicative intentionality of gestures, compared to speech — an argument we shall take issue with.

Displaying an anticlockwise rotation from the speaker's viewpoint would contradict the integrated nature of speech and gesture, resulting in a failure for the addressee to perceive the multimodal action as well-formed and coherent. The literature offers

enough evidence that the relations between speech and gesture can be of identity, complementarity but not of contrast since any contrastive effects could violate multimodal production and multimodal perception (cf. Lascarides and Stone [2009a] about the relations licensed by depicting gestures). The anticlockwise rotation is thus not evidence against the communicative intentionality of gestural performances.

Looking for the origins of gesture, De Ruiter [2000] proposed that gestures are involved in the process of information packaging: linguistic concepts as well as gestures are formed at a stage that precedes the formulation of verbal messages. According to this model, gestures are initiated by the conceptualiser where the preverbal message is constructed in terms of imagistic and spatial knowledge coming from the working memory. Apart from the integrated conceptualisation level, there is no further interaction between these two packaging modules (which is contrary to the Growth Point theory where speech and gesture are integrated in a single conceptual unit [McNeill, 2005]).

From the perspective of information packaging, the Interface Hypothesis (IH) assumes that gestures are generated from spatio-motoric processes that interact constantly with the processes underlying speech production [Kita and Özyürek, 2003]. This hypothesis is related to the Growth Point theory, which also assumes a dialectic between linguistic thinking and spatio-motoric thinking. In the IH theory, however, the surface packaging of speech and gesture originates from the on-line interaction between two autonomous systems: the system responsible for forming spatio-motoric representation and the system responsible for forming linguistic expressions.

IH also predicts that the way people package information into gestural and spoken modalities would vary depending on the expressive resources of a language—so gestural behaviour is predicted to differ across languages. This hypothesis was verified through experimental studies in which native speakers of English, Turkish and Japanese described the swinging event from the cartoon about Sylvester and Tweetie when the cat trying to catch the bird ‘swings’ across the street. The way information was packaged varied from language to language: in English, which has a separate lexical item for the swinging event, the gesture formed an arc trajectory. In contrast, in Japanese and Turkish where there is a lexical gap for the swinging event, the gesture did not depict an arc motion. This led Kita and Özyürek [2003] to conclude that since gestural expressions represented cross-linguistic variations, the form of gesture reflected the way people organise linguistic information.

But this finding contradicts other evidence and claims in the literature. McNeill

[1992] compared the iconic gestures of speakers of Swahili, Georgian, Mandarin Chinese and English produced when describing a cartoon and he discovered the same patterns of iconicity across the various languages:

> A remarkable thing about iconics is their high degree of cross-linguistic similarity. Given the same content, very similar gestures appear and accompany linguistic segments of an equivalent type, in spite of major lexical and grammatical differences between the languages. This resemblance suggests that the gesture emerges at a level where utterances in different languages have a common starting point—thought, memory, and imagery [McNeill, 1992, pp. 221–222].

Similarly, cross-linguistic variation was not established in other studies. For instance, Goldin-Meadow et al. [2008] used experimental studies to demonstrate that speakers of certain languages—English, Turkish, Spanish and Chinese—that do not adhere to the predominant word order, all used the actor-patient-event pattern in nonverbal descriptive tasks. The findings of this study are illustrative for the existence of a universal order for depicting events nonverbally which is not affected by the surface realised word order in a specific language.

Whereas the observation of Kita and Özyürek [2003] is an interesting attempt to establish parallelism between gesture realisation and linguistic forms, their study produces much more general conclusions than permitted by their evidence. To draw a reliable generalisation about cross-linguistic variations, we need a more systematic investigation over a larger set of empirical data.

Assuming that cross-linguistic differences and cross-linguistic universals can be found across sufficient data, both findings would not be of contradictory but rather of complementary character. Based on our current knowledge and understanding of gesture, it seems plausible that gesture production is governed by the same principles as language, and so both gesture universals and gesture variations could be found across languages.

### 3.3.2 Communicative Intentionality of Gestures

A radically social approach was offered in the studies of Janet Bavelas [Bavelas and Chovil, 2000; Bavelas and Chovil, 2006; Bavelas et al., 2008] who was interested in how conversational gestures function within face-to-face dialogue. In an experimental study, Bavelas et al. [2008] tested whether the most salient parameters of face-to-face dialogue—*visibility* (participants see each other) and *dialogue* (as opposed to

monologue)—have effects on the gesture performance, including gesture rate, gesture form and the gestures' relation to speech. The results can be summarised as follows: first, the rate of gesture in dialogue (no matter whether in a face-to-face condition or on the phone) was significantly higher than the gesture rate in monologue. Second, visibility had effects on gestural performance: for instance, the size of the gestures was significantly larger when the participants were visible to each other, and also the participants used deictic gestures when they could see each other. These findings refute any claims or intuitions that gestures are not intended for communication but that they function to facilitate the lexical retrieval [Krauss and Hadar, 1999; Krauss, Chen, and Gottesman, 2000].

Further, Bavelas et al. [1995] introduced the notion of *interactive gestures*. These are movements with the hand that do not refer directly to the topic of conversation but rather serve a performative function such as seeking a response, coordinating a turn, delivering information and also citing an addressee's contribution. In our study, we model these gesture types as having particular semantic relations to speech, similar to so called 'meta-talk relations' [Polanyi, 1985] in linguistic discourse — in this respect gesture is analogous to a parenthetical such as "tell me" or "of course" (recall our earlier discussion in Section 2.1).

On the same account, Holler and Beattie [2007] addressed the problem of whether gestures were influenced by the recipient's thinking and understanding. They tested the extent to which people resorted to gestures when conveying ambiguous information. In their experiment, participants were assigned the task of uttering a semantically ambiguous sentence such as "The old man's glasses were filthy" and after that they were asked for some clarifications about any of the two possible meanings for the lexically ambiguous word "glass". The authors reported that 46% of the tested subjects used gestures (with or without accompanying speech) to resolve the ambiguity thereby facilitating the communication.

This finding dovetails with our previous observations from Section 1.1. It provides evidence from a perception experiment confirming that even when there is redundancy across speech content and gesture content, the two modes are not redundant for parsing as the content of one mode serves to disambiguate the content of the other. The fact that gestures have a disambiguating function also confirms our observation that they are not semantically vacuous and that they contribute content to the final meaning of the utterance. We shall encode this meaning in the grammar by composing the gestural semantics with the linguistic semantics. For instance, the form of the gesture accom-

panying “The old man’s glasses were filthy” would map to a meaning representation unambiguously supporting one of the two possible interpretations for “glasses”.

### 3.3.3 Summary

Based on the integrated origin of speech and gesture as a cognitive process, we analyse the relationship between speech and gesture at a *production* level. Whereas the cognitive models explain what happens in the mind of the speaker when producing speech and gesture, a multimodal grammar gives a descriptive account of the plausible and implausible multimodal signals by means of constraints, articulated in terms of their form and relative timing.

Our work also draws on the radically social views on gesture: by means of a grammar, we intend to produce a formal representation of the multimodal action that supports the various interpretations of multimodal signals in their context of use. Essentially, the formal representations are compatible with any semantics/pragmatics framework that captures the evolving context in face-to-face dialogue. With this in mind, this work can be placed somewhere between the cognitive models of gesture—we provide a formal account of the speech-gesture interaction at the production level—and the communicative models of gesture—the representation of the communicative multimodal action generated by the grammar supports inferences that yield pragmatically plausible interpretations.

## 3.4 Computational Models of Gesture

Broadly speaking, the computational approaches to multimodal interaction proceed in two general directions: designing systems for multimodal recognition and parsing, and designing systems for multimodal production. The former involves implementing dynamic maps for communicating using pen and voice (e.g., Oviatt, DeAngeli, and Kuhn [1997], Johnston [1998a], Johnston [1998b], Johnston and Bangalore [2000]), and of robots executing human commands conveyed by multiple channels (e.g., Giuliani and Knoll [2007], Sidner and Lee [2007]). The latter has been accomplished by embodied conversational agents that synchronise speech production with hand gestures, facial expressions, lip- and bodily movements, etc. (e.g., Cassell et al. [1998], Kopp, Tepper, and Cassell [2004], Cassell, Stone, and Yan [2000], Cassell, Vilhjálmsson, and Bickmore [2001], Kopp and Wachsmuth [2004], Bergmann and

Kopp [2007]). Typically, the verbal and co-verbal grammar modules are incorporated within a larger domain-specific multimodal system, and thus a domain-independent grammar of aligned speech and gesture signals has not been accomplished yet.

In the next two sections, we provide an overview of the existing approaches to gesture in systems for multimodal parsing and systems for multimodal generation.

### 3.4.1 Multimodal Parsing

One of the first empirical investigations of the nature of multimodal interaction with a map-based system was conducted by Oviatt, DeAngeli, and Kuhn [1997]. That work then informed the creation of the QuickSet map based system [Cohen et al., 1997; Johnston et al., 1997]. Oviatt, DeAngeli, and Kuhn [1997] studied different aspects of multimodal interaction with the view of designing models, that would be predictive about multimodal interaction; for instance, does the frequency of multimodal usage differ when people issue spatial location commands and selection commands, and so is there a general preference for multimodal communication vs. unimodal communication? What are the temporal relations between the input signals? Also, what is the semantic relation between the input modes: complementarity or redundancy? Using simulation experiments where users navigated on a map, Oviatt, DeAngeli, and Kuhn [1997] reported a preference for multimodal interaction when users issued a spatial location command, and a higher frequency of unimodal interaction when people issued selection commands. Further, this study established that users relied on speech and writing modalities to express *complementary* information rather than duplicating information across speech and writing: for instance, participants used speech to describe the subject, verb and the object, and written commands to describe locative information [Oviatt, DeAngeli, and Kuhn, 1997]. This dovetails with the observation of McNeill [2005] for whom speech-gesture co-expressivity involves complementarity rather than redundancy (see Section 3.2.1). With respect to the temporal synchronisation patterns, Oviatt, DeAngeli, and Kuhn [1997] reported 57% writing precedence vs. 14% speech precedence and 29% no precedence of either mode.

To model the temporal interaction between speech and gesture, we use the findings of Oviatt, DeAngeli, and Kuhn [1997] in the following way: we use the relation *temporal overlap* to constrain the choices of integrating spoken input and gestural input. We view this relation as an abstraction over more fine-grained distinctions such as precedence of gesture start and sequence of gesture ending, precedence of speech start

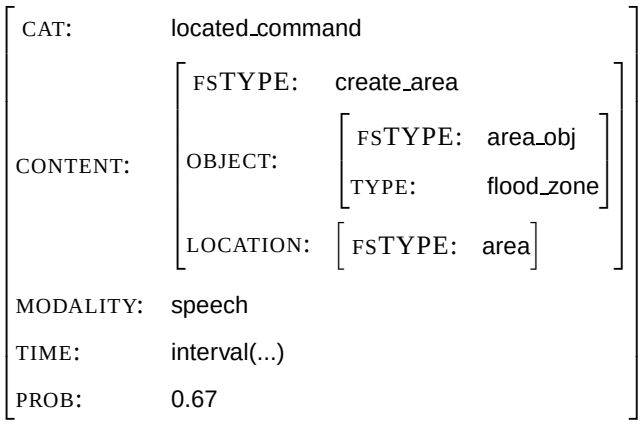



Figure 19: Feature Structure for Natural Language Input [Johnston, 1998a]

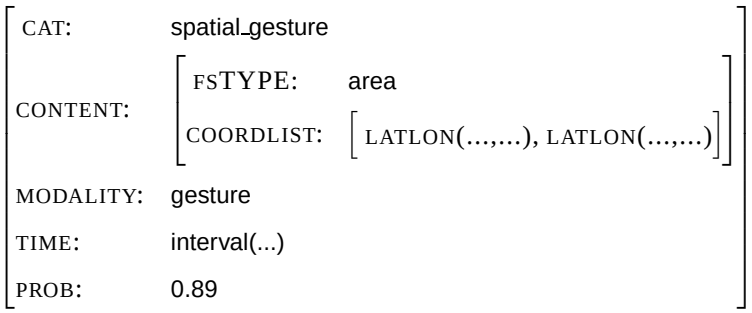



Figure 20: Feature Structure for Gestural Input [Johnston, 1998a]

and precedence of speech ending, etc. This relation is a convenient way of capturing the interaction of distinct modalities that allow a higher degree of temporal relaxation.[4]

Driven by the advantages of multimodal interaction, Johnston et al. [1997] used unification of gesture feature structures and speech feature structures to model multimodal integration of pen and speech. Further, Johnston [1998a; 1998b] introduced the notion of a feature structure based multimodal grammar and a multimodal chart parser. The multimodal architecture in that work was based on unification of Feature Structures (FS) assigned to linguistic and gestural input, as demonstrated in Figure 19 and in Figure 20 (taken from Johnston [1998a]). The CAT(egory) feature encodes the category of the issued command with the possible values being *located_command*—a spoken command—and *spatial_gesture*—a written gesture. The CONTENT introduces the semantic content of the input component. In Figure 19, for instance, this is a *cre-*

[4]Unlike the timing of speech production and lip movement, the temporal performance of gesture is rarely identical to the temporal performance of the semantically related speech units, i.e., the beginning and end of the semantically related speech and gesture rarely happen within the same time frame.

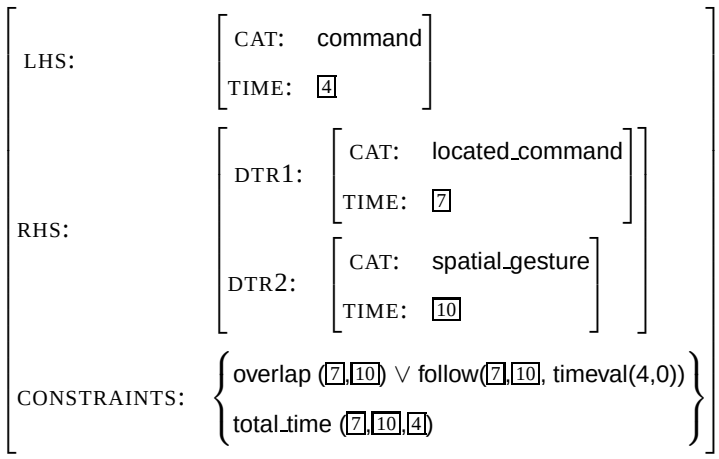



Figure 21: Rule Schema for Multimodal Integration [Johnston, 1998a]

*ate_area* command which creates an object of type *flood_zone* and the location is constrained to a type *area*. The remaining MODALITY, TIME and PROB(ability) features encode additional constraints.

The integration of the input signals follows the production of the form

$$\left[\text{LHS} \quad \rightarrow \quad \text{DTR1 DTR2}\right]$$

where each element—Left-Hand Side (LHS), Daughter1 (DTR1) and Daughter2 (DTR2)—is a feature structure representation [Johnston, 1998b]. Multimodal integration consists in unifying the feature structures under spatial and temporal constraints encoded as rule schemata. An excerpt of an example rule schema is shown in Figure 21 [Johnston, 1998a]. This rule licenses the integration of speech and gesture if the temporal relation between the two modalities is overlap or if speech is performed within a four-second delay from the gesture.

To handle more complex integration patterns such as the combination of speech input with several gestures, Johnston [1998a; 1998b] borrows some ideas for lexicalisation of grammars, namely: the speech input is assigned a feature structure with a multimodal subcategorisation frame indicating the range of gestures it needs to combine with, and then general feature structure schemata are used for the speech-and-gesture integration.

While the modularity of the unification-based approach [Johnston, 1998a; Johnston, 1998b] is clearly advantageous in that it allows for re-using different speech parsers in combination with the multimodal parser, this approach becomes more cumbersome with an increase in the complexity of the speech input since multimodal in-

tegration happens only after processing the speech utterance. Driven by the desire to scale up the unification-based grammars to more complex syntactic grammars [Johnston, 2000] and also driven by the desire to operate directly over lattice input, Johnston and Bangalore [2000; 2000; 2009] introduced a multimodal grammar formalism in which multimodal integration is at the level of terminals—a terminal is a triple composed of speech input, gesture input and their combined meaning—and achieved understanding of multimodal language by operating directly on multiple input streams via a cascade of finite-state operations which incorporate information from speech and gesture lattice inputs, including their single meaning representations.

The approach taken in this thesis is somewhere between the unification-based approach [Johnston et al., 1997; Johnston, 1998a; Johnston, 1998b] and the finite-state approach to multimodal integration [Johnston and Bangalore, 2000; Bangalore and Johnston, 2000; Bangalore and Johnston, 2009]: like the finite-state approach, we achieve multimodal integration in the syntactic grammar, and like the unification-based methodology we use typed feature structures for the speech and gesture input, and rule schemata to implement and constrain the combination.

In line with these previous unification-based approaches, we are going to use the relative timings as a constraint on the multimodal integration. In contrast to this prior work, however, our approach does not rely only on quantitative criteria. We shall use the *form* of the linguistic signal, as well as its relative timing, to determine speech-gesture alignment. We also go further than any prior research by providing a formal model which is predictive of multimodal ill-formedness. Recall earlier examples (1.4), page 13 and (1.5), page 14 and the associated discussions that form rather than timing is essential for the syntactic and semantic well-formedness of the multimodal signal.

Further, we intend to refine Johnston's [1998a; 1998b] grammar by providing a *domain-independent* form-meaning mapping of multimodal actions. Whereas the unification-based grammar of Johnston [1998a] is designed for issuing spatial commands via voice and pen on an interactive map, our formal model of multimodal communicative actions is applicable to any domain. This is also related to the gesture types we are dealing with: in contrast to Johnston [1998a] who models spatial gestures issued through a pen—that is, gestures that mark out an area of an artefact—we shall investigate different gestural dimensions—such as literally depicting, metaphorically depicting and deictic—so as to model their various semantic contributions to the final utterance, and the distinct ways they can relate with the speech.

A grammar-based model for processing multimodal input was proposed by Giu-

liani and Knoll [2007]. In this approach, the various input modes—speech, gesture and gaze—were processed autonomously using Combinatory Categorial Grammar (CCG) rules and were further abstracted over into hybrid logic formulae used for computing the system's reaction. In contrast to Johnston [1998a] where speech and gesture were directly integrated into a single multimodal representation, in the approach of Giuliani and Knoll [2007] the multimodal system operated on a mode-independent level, and so the distinct modes were not fused together in a single derivation tree but were rather mapped against each other on the basis of their timestamps. In contrast, we believe that speech and gesture combine in the grammar into a single parse tree, thereby capturing the constraints coming from the form of the linguistic component.

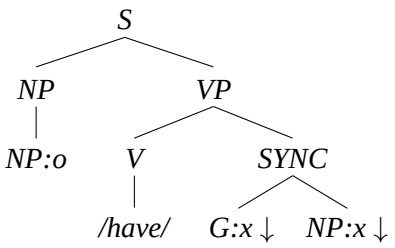
### 3.4.2 Multimodal Generation

Based on the integrated nature of speech and co-speech gestures [McNeill, 1992], Cassell, Stone, and Yan [2000] put this theory into practice for the purposes of a natural language generation system combined with gestures: the embodied conversational agent REA (Real-Estate Agent). The implementation was drawn upon the following theoretical principles: gesture and speech interact at an early conceptualisation stage and the mapping from form to meaning is linked to the context and the communicative intentions of the speaker. The integration of speech and iconic gestures was realised within a single derivation tree called a 'lexicalised description': a lexicon combining speech, gestures, predicate arity and pragmatic context, which declared the semantic and pragmatic coordination between a linguistic expression and gesture. Each lexical entry encapsulated syntactic (linguistic and gestural), semantic and pragmatic information as follows:

- a hierarchical tree in the format of a Tree-Adjoining Grammar where the syntactic constituent of the spoken phrase was paired with a gesture description. For instance,[5] utterance (3.6) illustrates the syntactic component for a transitive construction where a gesture of G category is integrated with an NP constituent producing thus a synchronous multimodal construction of SYNC category.

[5]The examples in (3.6), (3.7) and (3.8) are from Cassell, Stone, and Yan [2000].

(3.6)




- semantic specification. This is formally represented by the predicate and its arguments, as illustrated below:

  (3.7) $have(o,x)$

- pragmatic specification. This component establishes the relation between the utterance and the context, that is, whether the information contained in the *x* and *o* variables is new to the speaker.

  (3.8) *hearer-new(x)* $\wedge$ *theme(o)*

An essential contribution of this approach is the linking of speech and gesture within a single derivation tree which then projects a complex multimodal phrase. This is also the general direction we shall undertake. Note, however, that the model of Cassell, Stone, and Yan [2000] vastly under-generates the possible meanings of gesture by assuming that speech and gesture would always introduce the same referent (as realised by the *x* variable). To illustrate this limitation, consider example (3.9), which is a modified version of the original example in Cassell, Stone, and Yan [2000]:

(3.9) The house has a nice garden.
*Right hand is horizontal with open palm up. It performs a sweeping movement from the centre in front of the speaker's torso to the right periphery.*

It is perfectly acceptable for the gesture in (3.9) to be metaphorically interpreted as a gesture acknowledging a previous contribution. Following the terminology of Bavelas et al. [1995], this gesture would fall in the category of an interactive citing gesture. Essentially, the referent introduced by this hand movement has nothing to do with the topic of the co-occurring speech segment, namely nice gardens. Instead, it refers to the proposition expressed by the speech act and it can thus be interpreted as "I agree with your statement that the house has a nice garden". This interpretation would be possible by formally representing the gesture semantics as $gesture(\pi)$ where $\pi$ labels the proposition expressed by "The house has a nice garden". We therefore

claim that the range of different interpretations rendered by gestures is captured via an *underspecification* mechanism that does not under-generate the gesture meanings in context. We shall come back to this in Chapter 5 where we discuss the formal modelling of gesture form and meaning.

For the integration of speech and gesture, we also follow the model of Kopp, Tepper, and Cassell [2004] who analysed iconic gestures into semantic units—called 'image description features'—directly linked to the gesture form (namely, its trajectory, hand shape, movement, direction, location). The authors' motivation for building a common speech-gesture representation was the different semantics of lexical items vs. gestures:

> While a word may have a limited number of possible meanings, an iconic gesture without context is vague from the point of view of the observer, i.e., it displays an image that has a potentially countless number of interpretations in isolation [Kopp, Tepper, and Cassell, 2004].

We, however, do not support this observation: we believe that the underspecified meaning as revealed by just form indicates that a gesture maps to a certain meaning no matter how partial and incomplete it could be. The same is true for polysemous lexical items and "transfers of meaning" such as (3.10) (originally due to Lakoff and Johnson [1980]) and (3.11) (due to Nunberg [1995]) where the context-of-use generates a specific interpretation. For instance, out of context, the apple-juice seat from utterance (3.10) has no meaning, but in the context of four seats, where in front of one seat there was an apple juice and in front of the other three there was an orange juice, it is clear what the apple-juice seat refers to. Likewise, in (3.11) there is a nominal reference transfer from the person who had ordered french fries to the order itself.

(3.10) Please sit in the apple-juice seat.

(3.11) That french fries is getting impatient.

To capture the non-hierarchical organisation of gestures [McNeill, 1992], Kopp, Tepper, and Cassell [2004] used attribute-value matrices (see Figure 22). In this approach, the gesture interpretation was possible only after linking it to the context. The integration of speech and gestures happened in a single micro-planning stage. Similarly to Cassell, Stone, and Yan [2000], the grammar contained syntactic structures augmented with semantic information. Whereas the REA system used precanned gestures paired with the constituent of the synchronised utterance, here gesture-speech

| | |
|---|---|
| LOCATION | periphery right |
| TRAJECTORY | ⟨horizontal, linear, large⟩ |
| MOVEMENT-DIRECTION | forward |
| FINGER-DIRECTION | —— |
| HAND-SHAPE | 5 (asl) |
| PALM-DIRECTION | toward right |

Figure 22: Gesture's Image Description Features [Kopp, Tepper, and Cassell, 2004]

coordination was achieved by inserting on the fly a gestural feature structure into the constituent tree provided they had the same discourse referents. However, recall our earlier observation that this vastly under-generates the possible alignment configurations since gesture does not always introduce the same referents as speech.

### 3.4.3 Summary

The different approaches to multimodal interaction introduced above were intended as a part of larger multimodal systems for generating or parsing natural language input combined with gesture. The grammar modules were therefore designed for a restricted domain. In contrast, our grammar will operate on a domain-independent level thereby supporting the range of plausible interpretations in context.

In line with previous unification-based approaches [Johnston, 1998a; Johnston, 1998b; Kopp, Tepper, and Cassell, 2004], we intend to analyse multimodal actions in a unification-based grammar framework. Our formal representation of gesture is thus consistent with the observation from the descriptive studies that gestures are not hierarchical [McNeill, 2005], but rather the distinct aspects of gesture form such as hand shape, orientation, trajectory and direction each introduce a feature value pair, the combination of which yields the overall representation of gesture form. Further, we shall combine speech input with gesture input into a single feature-structure representation based on constraints expressed in rule schemata. Similarly to Johnston [1998a], we shall capture the integration in terms of temporal constraints. However, our constraints will also consider the linguistic properties of the speech signal—for instance, its prosody. In this way, we combine quantitative with qualitative criteria when constructing multimodal actions, thereby accounting for the various ways gestures can be interpreted in context.

The approaches introduced above modelled speech-gesture alignment as atomic elements the combination of which produced a common semantic representation. Cassell, Stone, and Yan [2000] combined the spoken utterance with the gestural production in one syntactic tree where the gestural production was paired with the production of the syntactic phrase. Further on, this speech-gesture pair was used to derive the semantic interpretation of the whole utterance. Likewise, in the model proposed by Kopp, Tepper, and Cassell [2004] an iconic gesture obtained its meaning by pairing it on-the-fly with a spoken utterance. Conversely, we assume that gestures convey a meaning out of context, which is derivable via a functional relation from its form. Following Lascarides and Stone [2009b] we assume that gesture form maps to a meaning representation which gives an abstract idea of what the gesture can mean in context. How exactly this abstract representation resolves is determined pragmatically in context.

The grammar-based model of Giuliani and Knoll [2007] used CCG to parse multimodal input which was further augmented with abstract logical formulae to derive their interpretation. We also saw that semantic relatedness was accounted for by matching the timestamps of gesture with that of speech. Similarly to that model, we shall be using a formal grammar framework with the view of extending an existing wide-coverage grammar with formal representations of multimodal actions. However, whereas for Giuliani and Knoll [2007] speech and gesture were aligned as two independent modules and their relatedness was captured via quantitative means of temporal matching, we assume that speech and gesture combine in a single derivation tree, as informed by the form of the linguistic signal. In so doing, we not only shed light on the cognitive accounts of co-verbal communication—imagistic thinking and linguistic thinking are formulated as a single multimodal unit contributing to the communicative action—but we also demonstrate that the speech-gesture interaction can be modelled via well-established techniques for grammar development.

## 3.5 Existing Formal Models of Multimodal Syntax

In the previous section we introduced grammar models of multimodal communication for the purposes of human-computer interaction systems. In this section, we overview the existing domain-independent formal models of multimodal syntax.

To date, we are familiar with the following formal approaches to multimodal actions: Kühnlein, Nimke, and Stegmann [2002] discussed an integration mechanism for speech and deixis in HPSG-grammars, Fricke [2008] reported on a multimodal

grammar of German, and Paggio and Navarretta [2009] demonstrated how the speech-gesture interaction can be formalised in a unification-based grammar.

### 3.5.1 An HPSG-based Integration of Speech and Deixis

The goal of the study of Kühnlein, Nimke, and Stegmann [2002] was to provide an integrated formal architecture of multimodal directives comprising speech and pointing gestures. The two modalities were analysed in parallel, interfacing their syntactic structure, semantic representation and pragmatic force in a unified representation. The model was empirically driven, using experiments over task-oriented dialogues. One of the questions in this study concerned the syntactic position and the semantic representation of the pointing signal. The integrated syntactic representation was achieved by linearising the speech and gesture inputs based on their relative timings where the deixis construct was treated like a head modifier. Kühnlein, Nimke, and Stegmann [2002] used the symbolic representation X $\searrow$ Y to designate that the gesture stroke happened after uttering X and before having uttered Y. For instance, they observed the following syntactic configurations for the multimodal utterance "Take the yellow cube" + deictic gesture:

(3.12) Take [NP $\searrow$ [NP the [N' yellow cube]]]

(3.13) Take [NP the $\searrow$ [N' [N' yellow cube]]]

(3.14) Take [NP the [N' yellow [N' $\searrow$ cube]]]

(3.15) Take [NP the [N' yellow [N' cube $\searrow$]]]

(3.16) Take [NP the [N' yellow [N' cube] $\searrow$]]

(3.17) Take [NP the [N' yellow [N' cube]] $\searrow$]

To build the final logical representation strictly compositionally, Kühnlein, Nimke, and Stegmann [2002] used typed $\lambda$-calculus. The semantics assigned to the $\searrow$ construct differed depending on its function. When the deictic gesture was used to single out an object, its semantic contribution was as shown in (3.18), and when the gesture marked as salient a whole set of objects, i.e., it was used restrictively, its formal representation was as shown in (3.19).

(3.18) $\lambda F \lambda x(x = c \wedge F(x))$

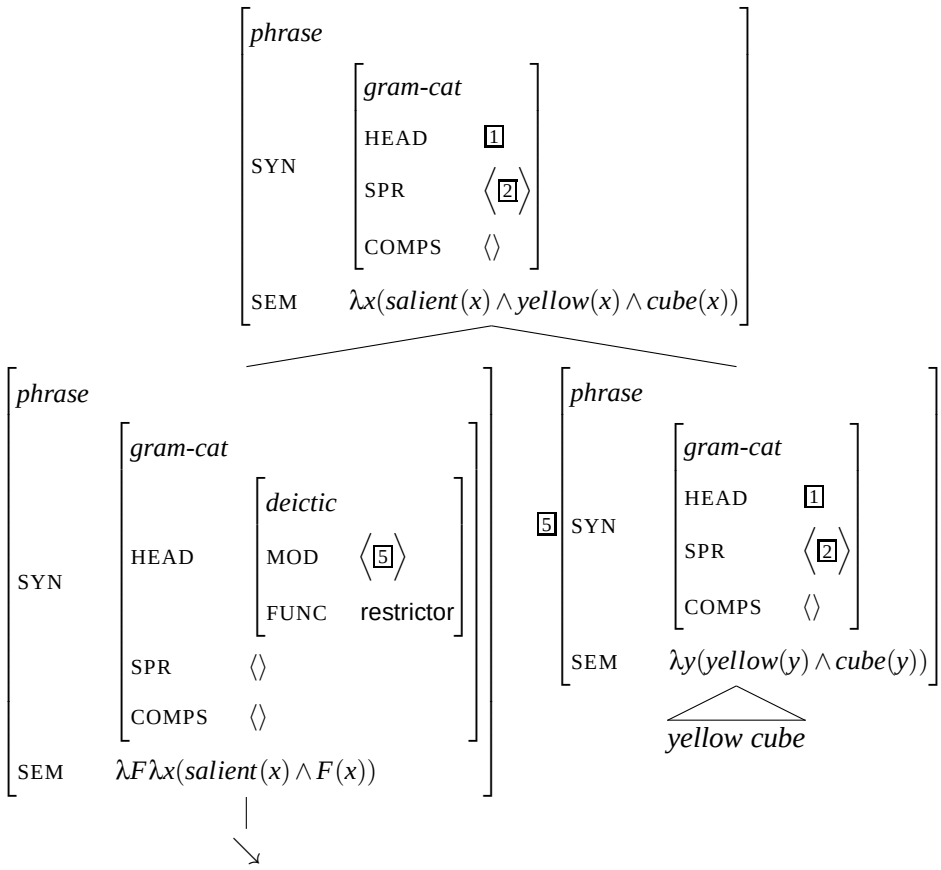



Figure 23: HPSG-based derivation of speech and deixis [Kühnlein, Nimke, and Stegmann, 2002]

(3.19) $\lambda F \lambda x(salient(x) \wedge F(x))$

The authors used the syntactic and semantic representations for the integration of speech and deixis in the HPSG grammar framework (see Figure 23). For instance, the combination of deixis and "yellow cube" was rendered by the standard Head-Modifier Rule by co-indexing the MOD feature of the deixis with the linguistic sign. The semantic representation of the mother N' projection was obtained by functional application of the spoken semantics to the deixis semantics.

This research is significant for understanding the integration of deixis in the syntactic structure of the spoken input and its semantic contribution to the final utterance. However, a certain limitation is that the integration patterns in (3.12) through (3.17) are based only on the timings of speech relative to gesture, and so they fail to be predictive about the range of well-formed multimodal actions. Kühnlein, Nimke, and Stegmann [2002] do not justify using a multimodal grammar on the basis that the *form* of the

| | | |
|---|---|---|
| | | $GP^n$ + Retr |
| | | $GU_1^{n+1}$ + $GU_2^{n+1}$ + ($GU_n^{n+1}$) + Retr |
| $GU^n$ | $\rightarrow$ | $GP^1$ + $GP^2$ + ($GP^n$) + Retr |
| | | $GU_1^{n+1}$ + ($GU_1^{n+1}$) + $GP^1$ + ($GP^n$) + Retr |
| GP | $\rightarrow$ | (Prep) + SP |
| SP | $\rightarrow$ | $S_n$ + $S_{n+1}$ |
| S | $\rightarrow$ | (Hold) + s + (Hold) |

Table 4: Gesture Production Rules [Fricke, 2008]

linguistic component influences the temporal performance of gesture: for instance, Kühnlein, Nimke, and Stegmann [2002] do not argue why deixis does not temporally co-incide with "take". In contrast to that, we use the form of the linguistic signal to provide a model that is predictive of which multimodal actions can be produced and which cannot. In this context where the "yellow cube" was the focus of the utterance, we assume that the prosodic unmarkedness of "take" explains why the deixis did not overlap the verb head. In the alternative (hypothetical) situation of uttering "Take the yellow cube" with a pitch on "take" as a continuation of "Don't throw the yellow cube. Take the yellow cube", it is possible for the deixis to happen along with "take". So, our grammar fills this gap on the basis that the form of the linguistic element effects the timing of the gesture.

In addition to that, their logical forms do not determine a *relation* between the gesture referent *x* and the speech referent *y*: they assume that the speech denotation is identical to the denotation of gesture. However, recall from Section 2.2.1 that speech and deixis can be bound through distinct semantic relations where only one of them is Identity. Assuming that speech and deixis introduce distinct referents, we can infer a wider range of possible relations between them.

### 3.5.2 A Multimodal Grammar for German

Some of the major questions addressed in Fricke [2008] concern the syntactic structure of multimodal actions: do gesture units combine into a complex gesture unit and if so, how do they integrate with the speech constituents? Fricke [2008] proposed to represent gesture units in recursive hierarchical trees based on Kendon's [2004] hierarchical gesture structure as follows (see the production rules in Table 4):

- The highest constituent in the gesture hierarchy is the Gesture Unit (GU) which can be elementary (it dominates an obligatory gesture phrase and a retraction) or complex (it dominates an *n*-number of gesture phrases followed by an obligatory retraction, or an *n*-number of gesture units followed by an obligatory retraction, or even a mixture of gesture phrases and gesture units followed by an obligatory retraction).

- The constituent immediately dominated by GU is the Gesture Phrase (GP). It contains an obligatory stroke and a non-obligatory preparation.

- The Stroke Phrase (SP) is the maximal projection of a stroke that dominates an *n*-number of sister strokes.

- The Stroke (S) dominates a single stroke that can be preceded or followed by non-obligatory holds.

- The elementary terminal nodes include: a Preparation (Prep), Retraction (Retr), Hold (Hold) and a Stroke (S).

The terminal nodes are associated with a list of feature-value pairs which encode the gesture form parameters such as hand form, movement, orientation, position and gravity.

Further to this, Fricke [2008] studied whether gestures can be integrated within the syntax of the spoken signal through some syntactic function. She observed that gestures can be assigned the syntactic function of modification, that is, they can be used to modify noun phrases in the same way as attributive adjectives do. To illustrate the attributive function of gesture, Fricke [2008, p. 206] gives the following examples:

(3.20) *Du gehst hier geradeaus entlang.* (No gesture)
'You walk straight-ahead.'

(3.21) *Du gehst hier geradeaus entlang.* (+ Gesture giving the direction)
'You walk straight-ahead.'

(3.22) *Du gehst hier entlang.* (+ Gesture giving the direction)
'You walk along.'

(3.23) *Du gehst hier entlang.* (No gesture)
'You walk along.'

Assuming that the deictic origo is provided by the body orientation of the speaker, Fricke [2008] observed that the gesture in (3.22) plays the attributive role of specifying the direction, which in (3.20) and (3.21) is rendered by the modifier "geradeaus". The author also specified that this gesture usage is not a replacement for a grammatical gap in the linguistic signal since it was performed along with *hier* 'here' and not along with a speech pause. Fricke [2008] concluded that due to its attributive function, the same grammatical principles can be applied to gesture and speech, and hence gesture can be integrated into the linguistic utterance in the same way as units of speech.

Note that the grammar production rules in Table 4 do not contradict our assumptions that gestures are not hierarchically organised. Whereas Fricke [2008] works on the highest level of gesture performance, the gesture unit, and then proceeds recursively in a top down direction to investigate the possible gestural phases within this unit—for instance, preparation, stroke and retraction—our work is interested in the expressive part of the gesture excursion, the gesture stroke. It is the stroke whose meaning is not derived compositionally.

### 3.5.3 Integration of Speech and Gesture in Unification-based Grammars

An alternative approach to multimodal integration was proposed by Paggio and Navarretta [2009] who analysed the phonological, syntactic, semantic and pragmatic information of the multimodal sign in terms of feature structures. In this model, the integration of the speech signal and the gesture signal was a matter of temporal co-occurrence, and so a gesture signal could be combined with a single co-occurring word or with a sequence of temporally co-occurring words.

For instance, Figure 24 illustrates a multimodal sign where the gesture daughter is a single gesture and the speech daughter is a single word. The multimodal construction combines the contribution of both modalities in parallel: the linguistic sign is defined in terms of its phonological representation (the word *tak* 'thank you'), syntactic information SYNSEM and also its dialogue act DIAL-ACT. The DIAL-ACT feature is structure-shared with the DIAL-ACT feature of the multimodal sign. The gesture sign is of type *FacialDisplay* whose semiotic type is *IndexicalNon-deictic*, and whose function is coded as *FeedbackGive*, which is token-identical with the FUNCTION feature of the mother multimodal sign. The DIAL-ACT and the FUNCTION features capture the fact that the gesture sign and the linguistic sign reinforce each other.

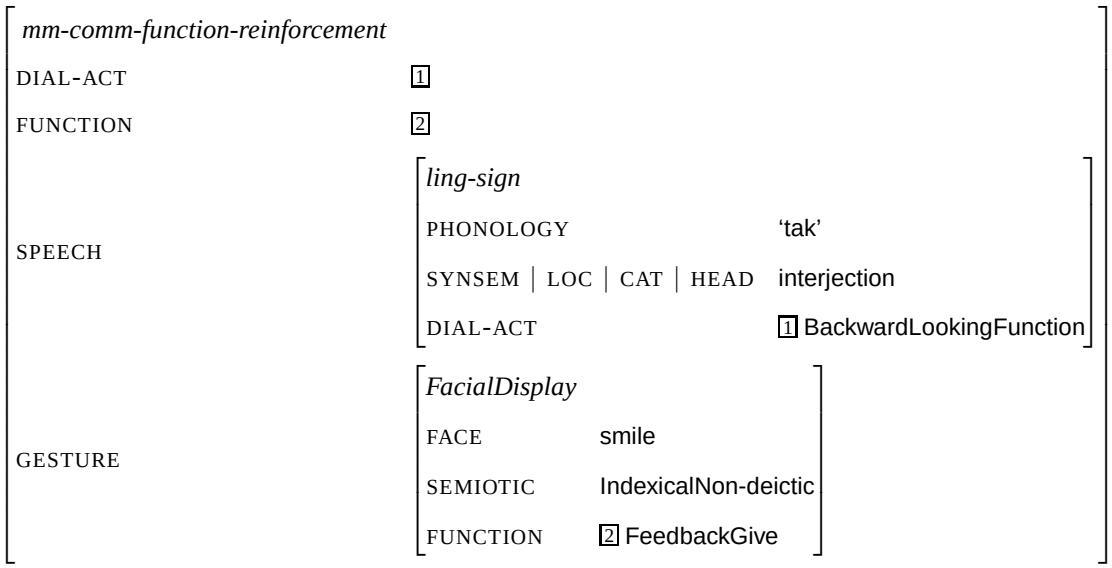



Figure 24: Feature Structure Representation of a Multimodal Sign [Paggio and Navarretta, 2009]

Further to this, Paggio and Navarretta [2009] argue that there is a significant difference in the conceptual level in which distinct types of gestures are integrated with the linguistic signs. For instance, whereas beat gestures interact with linguistic signs at the level of information structure—say, focus—iconic gestures interact with gestures at the level of content. This distinction also presupposes that the distinct gesture types in distinct contexts would encode distinct features, so as to accommodate the various levels of reinforcement between speech and gesture.

The integration of a gesture sign and a sequence of spoken words is also based on the temporal co-occurrence of the two modalities. Syntactically, the synchronous linguistic sign is not necessarily a syntactically complete phrase, it can be a fragment and it could also contain fillers and self-repairs. Since the feature structure representation of the integration of gesture and multiple words follows the principles introduced above, we do not introduce them here.

While the relative timing in the model of Paggio and Navarretta [2009] imposes one-to-one mappings, our grammar is less restrictive with respect to the speech-gesture alignments. This is accomplished by using form as a constraint on the alignment rather than timing alone. In so doing, our model supports the range of interpretations of the multimodal action without enforcing a unique integration. This is also related to the fact that our grammar rules can only license the alignment of gesture to a syntactically

complete structure.

### 3.5.4 Summary

In this section we presented three formal models of multimodal syntax that shed light on the interaction between speech and gesture on the level of form. These studies confirm our starting assumptions that speech and gesture can be analysed in terms of the same formal model while adapting it as necessary. The finding of Fricke [2008] that gestures can take over an attributive function plays an important role when designing the methodology of multimodal analysis.

In line with the previous work on formal syntax [Kühnlein, Nimke, and Stegmann, 2002; Paggio and Navarretta, 2009], we intend to use unification-based grammars to produce a single syntactic tree for the multimodal action which maps an integrated meaning representation. From a formal perspective, unification-based grammars are formalised by feature structure descriptions which are now well-established for expressing gesture form. However, a limitation of all previous formal approaches to multimodal actions is that they integrate the speech signal with the gesture signal on the basis of their timings, forcing thus a unique coordination pattern between the two signals. In contrast, we align speech and gesture in a single tree, where the form of the linguistic signal constrains the alignments (in Chapter 4, we provide empirical evidence to demonstrate that speech-gesture alignment is informed by the form of the linguistic signal). In so doing, we capture the various ways gesture can relate with different parts of the speech signal. In contrast to Kühnlein, Nimke, and Stegmann [2002], we shall not linearise the speech and gesture inputs thereby not constraining the integration to a head-modifier relation. Also, instead of using lambda calculus over fully specific first order formulae to represent gestural semantics and to compose the meaning of the multimodal action, we shall use unification over underspecified semantic formulae so as to account for the various gesture interpretations in their context of use. Whereas deriving fully specific logical forms might be suitable for deictic gestures (but see Chapter 5 for evidence that even deictic gestures have underspecified meanings), we claim that semantic underspecification mechanisms can reliably capture the ambiguous gesture form.

In Section 3.2.2, we used the finding of McNeill [2005] to demonstrate that a gesture (stroke) cannot be represented in terms of a traditional syntax tree. This observation does not contradict the hierarchical analysis proposed by Fricke [2008]. The

non-hierarchical nature of gesture is at the level of its expressive part, the stroke. In contrast, Fricke [2008] proposed to organise gesture trees on the level of the gesture unit. Since we are interested in the form-meaning mapping of gesture, we shall not analyse the syntax of the gesture unit.

## 3.6 Conclusions

In this chapter we presented various studies of multimodal interaction, focussing on how our work fits the broader context of gesture research.

Consistent with recent gesture studies, this thesis builds on the finding that speech and gesture function in tandem so as to communicate a uniform message. In contrast to prior research, we approach the multimodal actions by means of standard techniques from linguistics for the form-meaning mapping. A vastly under-researched area remains the effects of the form of the multimodal action on the final interpretation in context. In all prior work, the integration of speech and gesture into a composite signal was driven from the temporal performance of speech in relation to the temporal performance of gesture, thereby forcing a unique integration pattern; for instance, identity of the beginning and ending of both modalities was sufficient for their semantic relatedness. In contrast, we look at the form of the linguistic component, the form of gesture component, as well as their relative timing to derive a methodology for identifying the semantically related speech and gesture signals. An added benefit of this methodology is that it allows us to account for the one-to-many form-meaning mappings, without undergenerating what the gesture can mean in context. We also aim for representations of meaning that respect constraints on reference that form part of the semantics/pragmatics interface.

By using mechanisms for underspecifying gesture meaning, we are also able to abstract over the full range of gesture interpretations, and hence to abstract over the full range of semantic relations between speech and gesture. Whereas for the domain-specific purposes of previous studies—for instance, Johnston [1998a], Cassell, Stone, and Yan [2000]—a higher abstraction level was not necessary, our programme includes a domain-independent form-meaning mapping of multimodal actions and we therefore need mechanisms for yielding the full range of possible interpretations. Accounting for the form of the linguistic signal brings further the studies of multimodal communication in that it allows us to build a model predictive of multimodal well-formedness and multimodal ill-formedness.

# Chapter 4

# Empirical Investigation

Most approaches to multimodal communication take for granted that the temporal synchrony between speech and gesture is a premise for their alignment — i.e., the speech signal that gesture can be semantically related to is based only on temporal co-occurrence. Given our assumptions about a coherence-based pragmatics model (Section 1.3.1), we defined the notion of speech-gesture alignment (Section 1.4) in terms of the attachment of speech and gesture in a single syntactic tree which maps to an (underspecified) meaning representation. We also stipulated that the attachments are licensed by constraints coming from the form—e.g., prosody and syntax—of the linguistic phrase.

Our assumptions about the relationship between the performance of gesture and the prosody of the temporally overlapping speech signal are based on the intuitions of native speakers and also on previous work in the gesture literature. We believe, however, that native language speakers lack entirely reliable introspective intuitions about multimodal communication: for instance, Loehr [2004] did not find empirical proof for the parallel hypothesis of Bolinger [1986]; the trade-off hypothesis which claims that people use gestures as a compensatory mechanism for disfluency in speech was not empirically validated [De Ruiter, Bangerter, and Dings, 2012]. Therefore, our stance on the issue of multimodal grammaticality is that introspective speculations can be falsified by examining real data. We shall therefore proceed with an empirical investigation that sheds light on the relationship between speech and gesture.

In Section 1.3.2, we presented constructed and real examples that demonstrated the coordination between gesture performance and the prosody and syntax of speech. This chapter builds on these prior observations through a detailed study of multimodal corpora.

While we can find evidence for the 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 *relationship* between gesture and speech plays a central role in combining with either thematic or rhematic utterances. Later in this chapter we shall provide our intuition about how the gesture manifestation reflects the information status of the utterance. Although this knowledge might be needed by a discourse processor, we are not convinced that information structure should constrain the choices of attachment for linguistic phrases and gesture within the grammar. We shall limit ourselves to prosody, syntax-semantics and timing as central factors for combining speech and gesture within the grammar to produce a unified meaning representation.

To spell out constraints on the alignment between gesture and speech, our empirical investigation proceeded in two separate stages: the first one studied the interaction between depicting gesture and speech, and the second one investigated deixis and speech. Our motivation for not conducting a common study is theoretically and practically grounded. The theoretical reason pertains to the diametrical difference between these two gesture dimensions: namely, the form-meaning mapping of depicting gestures is derived from the qualitative characteristics of the hand form features; contrary to that, how the form of deictic gestures maps to meaning includes the spatio-temporal context in which the utterance was uttered. From this perspective, the truth-conditional content of depicting gestures is modelled only through its relation with speech, whereas the content of deictic gesture is understood as a function that maps from its *contextually-specific time and space* to reference and truth values. The practical reason was the availability of the resources, and more specifically the fact that at the time of conducting this study there was no single multimodal corpus that was annotated for speech and gesture and that comprised a range of examples of the various gesture types.

Despite the differences, our investigation addresses the findings in the literature concerning speech-gesture interaction, and the aim is to shed light on the following questions: do gesture strokes happen along with pitch accents in speech? Do gestures coordinate with particular tonal pitch accent events? Do gestures interact with the metrical structure of the speech phrase? Are there gestures that do not overlap with the semantically related speech phrases, and if so how do we explain these instances? Do gestures occur with a particular syntactic constituent, if any at all? Do beat gestures always superimpose other gesture dimensions, or pure beats happen as well?

| | Pitch Accents | | Phrase Accents | | Boundary Tones | | All Tones | |
|---|---|---|---|---|---|---|---|---|
| | # | % | # | % | # | % | # | % |
| Number of words in data | 525 | | 525 | | 525 | | 1575 | |
| Agreement on existence of tone type | 494 | 94% | 489 | 93% | 484 | 92% | 1467 | 93% |
| Agreement on exact tone | 462 | 94% | 477 | 98% | 483 | 100% | 1423 | 97% |
| Absolute agreement | | 88% | | 91% | | 92% | | 90% |

Table 5: Inter-annotation agreement on ToBI [Loehr, 2004]

We start with a description of the corpora that we used, and then we proceed with putting forth our hypotheses and the subsequent empirical study.

## 4.1 Corpora and Annotation

### 4.1.1 Corpus of Loehr [2004]

For depicting gestures, we used a 165-second collection of four recorded meetings annotated for gesture and intonation [Loehr, 2004]. The conversations were natural, spontaneous, and took place among friends. The topics of the conversations were 1. fixing cupboards; 2. teaching music; 3. renovating houses; and 4. common friends. The corpus was labelled with the view of investigating the relationship between intonation and gesture. In particular, Loehr [2004] searched for evidence for the temporal alignment between the gesture peak, called ‘apex’, and the pitch accent, and also between gesture phrases and intermediate intonation phrases.

**Intonation Annotation** The intonation annotation followed the guidelines for Tones and Break Indices (ToBI) [Beckman and Elam, 1997]. The coding was done in Praat [Boersma and Weenink, 2003] and it included an orthographic transcription, the location and specification of L(ow) or H(igh) pitch accents, intermediate phrases (corresponding to break index level 3) and intonation phrases (break index level 4). The inter-annotator agreement, as reported by Loehr [2004], is displayed in Table 5.

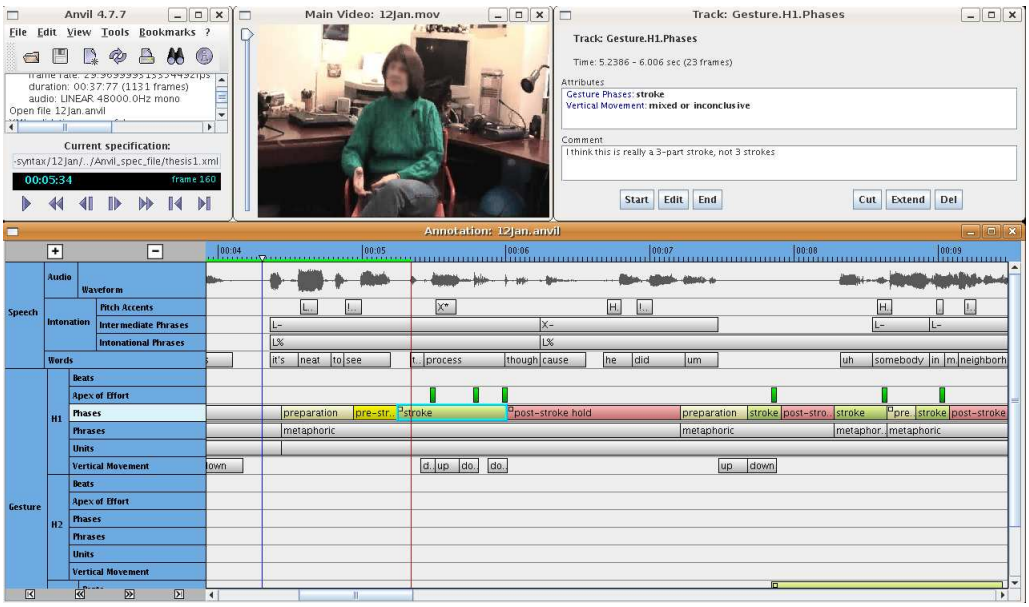



Figure 25: Labelled Utterance in Anvil [Loehr, 2004]

**Gesture Annotation** The gesture annotation was performed in the Anvil labelling tool [Kipp, 2001] and it was largely based on the coding instructions from the McNeill lab [McNeill, 1992]. The annotation (see Figure 25 which illustrates the labelling for the utterance "It's neat to see the process cause he did um…") was split in three subgroups: the main gesturing hand (H1), the non-dominant hand (H2) and the head. We forgo any details about the latter since head movements lie outwith the scope of this thesis.

The H1 and H2 annotation included marking the following events:

- *Beats.* This included coding the beginning and ending of up-down and down-up hand movements.
- *Apex of movements.* Unlike the gesture stroke which spans an interval, the apex is instantaneous and it encodes the peak of the gesture stroke. This was the main deviation from McNeill's [1992] instructions.
- *Gesture phases.* These cover preparation, stroke, retraction, holds, recoils.[1]
- *Gesture phrases.* Consecutive gesture phases with one obligatory stroke make

[1] We have not mentioned recoils since they generally happen outside the gesture; these are slight bouncing back movements which usually happen after the hand had retracted to rest.

| | Gesture (G)-phases | | Gesture (G)-phrases | | |
|---|---|---|---|---|---|
| | G-phase boundaries | G-phase types | G-phrase boundaries | G-phrase types | G-phrase meanings |
| Annotations marked | 136 | 52 | 52 | 21 | 18 |
| Annotations agreed upon | 103 | 46 | 42 | 21 | 16 |
| Percentage of annotations agreed upon | 76% | 88% | 81% | 100% | 89% |

Table 6: Inter-annotation agreement on gesture phase boundaries, gesture phase types, gesture phrase boundaries, gesture phrase types and gesture meanings [Loehr, 2004]

up a gesture phrase. They were exclusively associated with a category: deictic, iconic, metaphoric, emblem, adaptor.[2]

- *Gesture units.* The entire excursion of the hand from leaving the rest position to returning to a rest position is identified as a gesture unit.
- *Vertical movements.* Any vertical head or hand movement was assigned its direction with the view of testing the parallel hypothesis of Bolinger [1986].

Each gesture phrase also contained a short English description of its meaning. These were assigned in the context of speech [Loehr, 2004]. The inter-coding agreement reported by Loehr [2004] can be seen in Table 6.

### 4.1.2 Talkbank and AMI Corpora

For deictic gestures, we used two multimodal corpora: a 5.53 min recording from the Talkbank Data,[3] and observation IS1008c, speaker C from the AMI corpus [Carletta, 2006].[4] The domain of the former is living-space descriptions and navigation giving,

[2] In the gesture community, non-communicative hand gestures are known 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 the nose.

[3] The video clip can be found here `http://www.talkbank.org/media/Gesture/Cassell/kimiko.mov`

[4] `http://corpus.amiproject.org`

and the latter is a multi-party face-to-face conversation among four people discussing the design of a remote control. To address our underlying assumptions concerning the interaction between speech and gesture, we annotated both corpora in two separate stages: annotation of speech which included word transcription, pitch accents pointing to words and prosodic phrases; and gesture annotation which included marking of gesture phrases, gesture phases, and also beats. Both annotations were performed independently from each other so as to guarantee unbiased judgements.

**Prosody Annotation** We adopt the Autosegmental-Metrical (AM) framework (term coined by Ladd [1996]) for the analysis of speech prosody. Our choice is motivated in the fact that in the AM model prosodic prominence is signalled not by the acoustic rise of a stand-alone event, but it is rather viewed as a relational property between two juxtaposed units structurally organised in a metrical tree, which is consistent with the phrase's underlying rhythmical organisation [Calhoun, 2006]. In this way, we can reliably predict the performance of the stroke based on the metrical tree, and we can also interface the hierarchical prosodic structure with the syntactic structure within the grammar [Klein, 2000a; Klein, 2000b].

In the AM framework, nuclear prominence results from the following operations:

- mapping a syntactic structure to a binary metrical tree;
- assigning *strong (s)* or *weak (w)* prosodic weight to the nodes in the metrical tree according to the metrical formulation of the Nuclear Stress Rule shown in Definition 4.1.1 [Liberman and Prince, 1977, p. 257];
- tracing the path dominated by *s* nodes to determine the prominent peak, also known as the Designated Terminal Element (DTE) of the phrase [Liberman and Prince, 1977].

**Definition 4.1.1. Nuclear Stress Rule.** *In a configuration $[_C AB]$, if C is a phrasal category, B is strong.*

In the default case of broad focus, the metrical structure is right-branching, i.e., the nuclear accent is associated with the right-most word. In instransitive constructions, however, the preferred pattern is a nuclear accent on the subject. For instance, Figure 26 illustrates the metrical tree for "your mother called" in its broad focused reading with the nuclear accent being on the word entirely dominated by *s* nodes—"mother". Early pre-nuclear rise on the left of the nuclear node is also possible, and it

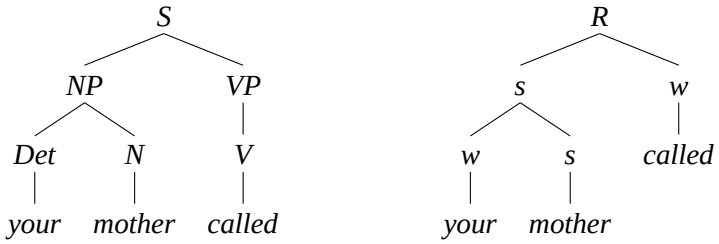



Figure 26: A Syntactic and a Corresponding Metrical Tree

is signalled through its acoustic properties rather than its relative position in the metrical tree. As per Liberman and Prince [1977], we mark the root node of the tree with the generic label *R* to designate that it is not relational, that is, it is neither weak nor strong.

As an annotation tool, we used Praat [Boersma and Weenink, 2003] since it allows visual representation of the pitch and intensity (see Figure 27). Our annotation schema[5] is largely based on the guidelines of the prosody annotation of the Switchboard corpus [Brenier and Calhoun, 2006], which included marking the following layers:

- *Orthographic Transcription.* The Talkbank recording was transcribed manually; the AMI meeting was already equipped with a manually produced transcription which we converted into Praat readable format after stripping off the punctuation marks. Silent pauses, laughter and unintelligible phrases were marked with special tags.

- *Pitch Accents.* Words were unambiguously associated with at least one accent of the following type: *nuclear:* the accent of the whole prosodic phrase that is structurally, and not phonetically perceived as the most important one. Each prosodic phrase has at least one nuclear accent; *pre-nuclear:* an early emphatic high rise characterised by a high pitch contour. Intuitively, if within one phrase two accents were perceived as nuclear, we marked the early one as pre-nuclear. In general, pre-nuclear accents precede a downstepped pitch accent where the downstepped accent marks "inferable or otherwise semi-active information" [Baumann, 2006]; *non-nuclear:* unlike nuclear accents, non-nuclear accents are perceived on the basis of their phonetic properties, and the rhythm of the sentence (they correspond to 'plain' or 'regular' accents in Brenier and Cal-

[5]Many thanks to Sasha Calhoun for the helpful discussions concerning the annotation guidelines.

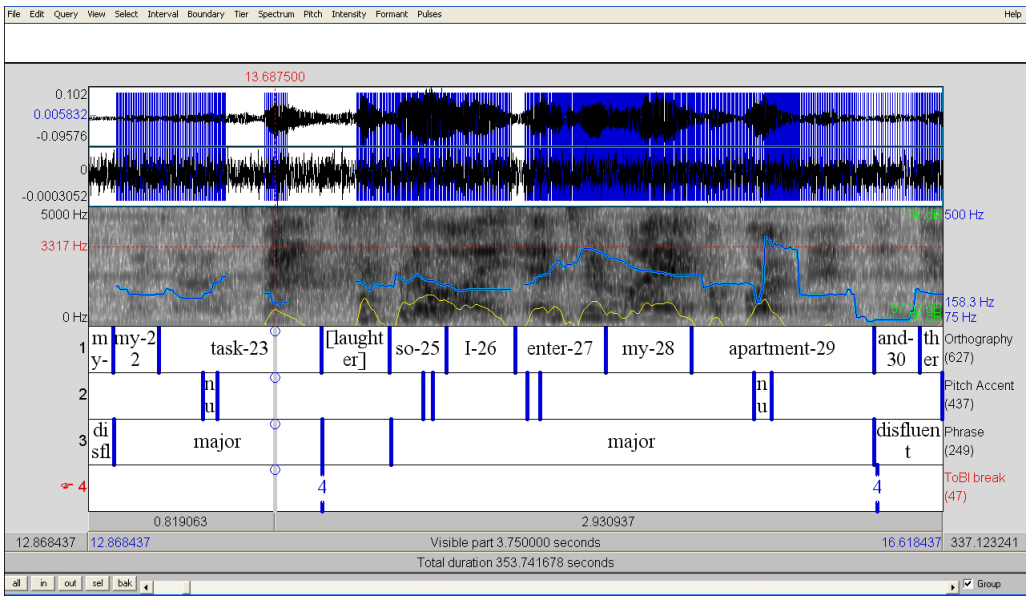



Figure 27: Prosody Annotation in Praat: the annotation tiers included Orthography, Pitch Accents, Prosodic Phrases and ToBI break indices. Praat also displays the waveform, the spectrogram (in grey), the pitch (blue line) and the intensity (yellow line).

houn [2006] and Calhoun [2006]); *none*: a non-discernible accent in a phrase (it corresponds to a 'Z' accent in Brenier and Calhoun [2006]); a question mark *(?)*: uncertainty concerning the presence of an accent.

- *Prosodic Phrases.* A group of words form a prosodic phrase whose type is determined by the break type after the last word in the phrase. We annotated the following phrases: *minor*: phrase where the break after the last word corresponds to ToBI break 3; *major*: phrase where the break after the last word corresponds to ToBI break 4; *disfluent*: phrase where the break after the last word would be marked in ToBI with the *p* diacritic, that is, *1p, 2p, 3p* correspond to disfluent phrases; *backchannel*: short phrases containing only fillers such as "er", "um", "you know", etc.

Past annotation tasks of the Switchboard corpus (see the inter-coder agreement in Table 7) have shown that this annotation strategy is reliable. The Cohen's $\kappa$ coefficient is calculated from the observed pairwise annotator agreement $(Pr(a))$ and the probability of the expected chance agreement $(Pr(e))$ (see (4.1)). It is generally believed that $0.67 < \kappa < 0.80$ is fair, and $\kappa > 0.8$ shows good reliability [Carletta,

| | All Types | Absence/Presence |
|---|---|---|
| Accents | 0.800 | 0.800 |
| Boundaries | 0.889 | 0.910 |
| Words | (752) | |

Table 7: Inter-annotation agreement on accents and phrase boundaries, and also on the presence/absence of accents and boundaries in kappa (κ) [Calhoun, 2006]

1996].

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{4.1}$$

**Gesture Annotation** We used the Anvil labelling tool to annotate the gesture phrases, gesture phases and beats. Along the lines of Loehr [2004], we annotated gestures for the dominant H1 hand, and for the non-dominant H2 hand. Bi-handed gestures where the movement of H1 was symmetrical to H2 were coded in H1.

- *Hand Movement.* The annotation of the hand movement proceeded in two main passes. The first pass involved marking the temporal boundaries of all hand movements, and performing a binary classification on them in terms of *communicative* vs. *non-communicative* signals. The second pass determined what dimensions the communicative signals belonged to: *literally depicting, metaphorically depicting* or *deictic*. Deictic gestures were further subdivided into *abstract, concrete* or *nomination*. To stay consistent with the findings in the literature that gestures are multidimensional [McNeill, 2005], our annotation schema permitted for marking gestures belonging to more than one dimension.

- *Gesture Phases.* This step involved annotating the phases comprising each hand movement: *preparation, pre-stroke hold, stroke, post-stroke hold* and *retraction*. The distinction between pre-stroke holds and post-stroke holds was often not clear: the form of the hand itself was ambiguous as to whether the signal belonged to the new gesture phrase and it was thus a pre-stroke hold, or it belonged to the previous gesture phrase, and it was thus a post-stroke hold. To help with this issue, we looked at linguistic cues, for instance, pre-stroke holds are more

| | **Segmentation Agreement** | | **Coding Agreement** | | |
|---|---|---|---|---|---|
| | Cohen's κ | Corrected κ | Cohen's κ | Corrected κ | Percentage |
| Hand movement | 0.8502 | 0.8659 | 0.8536 | 0.8994 | 93.2943% |
| Deictic gesture | 0.8502 | 0.8659 | 0.8605 | 0.8994 | 93.2943% |
| Literally depicting | 0.8502 | 0.8659 | 0.8663 | 0.8916 | 92.7734% |
| Metaphorically depicting | 0.8502 | 0.8659 | 0.8221 | 0.8623 | 90.8203% |
| Emblem | 0.8502 | 0.8659 | 0.8502 | 0.8659 | 93.2943% |
| Gesture phase | 0.8864 | 0.8971 | 0.662 | 0.7 | 75% |
| Beat | 0.6599 | 0.8203 | 0.6599 | 0.8203 | 91.0156% |
| Gesture's meaning | | | | | 75.9259% |

Table 8: Inter-coder reliability of gesture segmentation and gesture coding in Cohen's κ and corrected κ

likely to occur with discourse connectives, pronouns, relative pronouns, temporal adverbial such as "while", "when" than post-stroke holds [Kita, 1990]. We observed that pre-stroke holds tend to appear with hesitation pauses while the speaker is looking for some stable verbal form, and so recovery of the temporal cohesion is anticipated; in contrast, post-stroke holds are more likely to occur with fluent speech when the speaker elaborates on the content reached during the stroke.

- *Beat.* Beat movements were marked in a separate layer so as to study whether they always superimpose other gestural dimensions, or pure beats also occur.

We used the gesture annotation schema on a single observation of Loehr's [2004] multimodal corpus when we performed the experiment discussed in Section 1.1.1. The inter-annotator agreement is shown in Table 8. The segmentation column shows agreement on the presence/absence of an element within a certain time slice, and the coding column shows agreement on the element type within the time slice. In the corrected κ, the chance probability is replaced by $1/n$, with $n$ being the number of categories [Kipp, 2008]. We checked manually the agreement on the assigned gesture meanings (recall discussion in Section 1.1).

### 4.1.3 Multimodal Corpora in NXT

To extract generalisations from the corpora, we used the Nite XML Toolkit (NXT [Carletta et al., 2005]). NXT supports flexible and intelligent search of the corpus, including the means to find examples that exhibit specified boolean relations of values across the distinct annotation layers. For that purpose, we automatically converted the Anvil annotations of the three collections—Loehr's [2004] corpus, the Talkbank recording and the AMI observation—into an NXT readable format. A corpus in NXT consists of 'observations' where an observation is composed of a video signal and the annotations associated with it. In this case, our multimodal corpus contained six video signals and the corresponding annotations: orthographic transcriptions, pitch accents, prosodic phrases, gesture phrases, gesture phases and beats. Each data object is necessarily equipped with timestamps so that they can be synchronised with other data objects and with the video signals.

Data objects interact with each other by structural or temporal relations. This information is declared in a meta-data file containing the annotation schema of the corpus. The type of relation—structural or temporal—also determines the query that can be executed onto these objects. The annotation of each type of data object is stored in a separate XML file, and so we created separate annotation files for words (that is, transcription), accents, prosodic phrases, gesture phases, gesture phrases and beat. The relations between the annotation objects are defined in terms of stand-off links between the elements. Figure 28 illustrates the relation between the 'accents' and 'words' layers: the accent's attribute nite:pointer points to the unique nite:id of the relevant word. In this way, we can elegantly capture accents not overlapping a word, accents associated with more than one word, and also words associated with two accents: for instance, in case of an accent referring to multiple words, the nite:pointer of the accent points to the nite:ids of the relevant words.

We further specified the relationships between gestures and gesture phases, and between prosodic phrases and words as parent-child relations (see Figure 29). This choice of representation is consistent with the essence of prosodic phrases and gesture phrases: prosodic phrases are made up by a certain number of words, and so the beginning of the first word is identical to the beginning of the prosodic phrase, and the end of the last word is identical to the end of the prosodic phrase. The same mechanism applies to gestures which are made up by at least one gesture phase. We forgo any details about the specification of beats since they are not represented in a structural

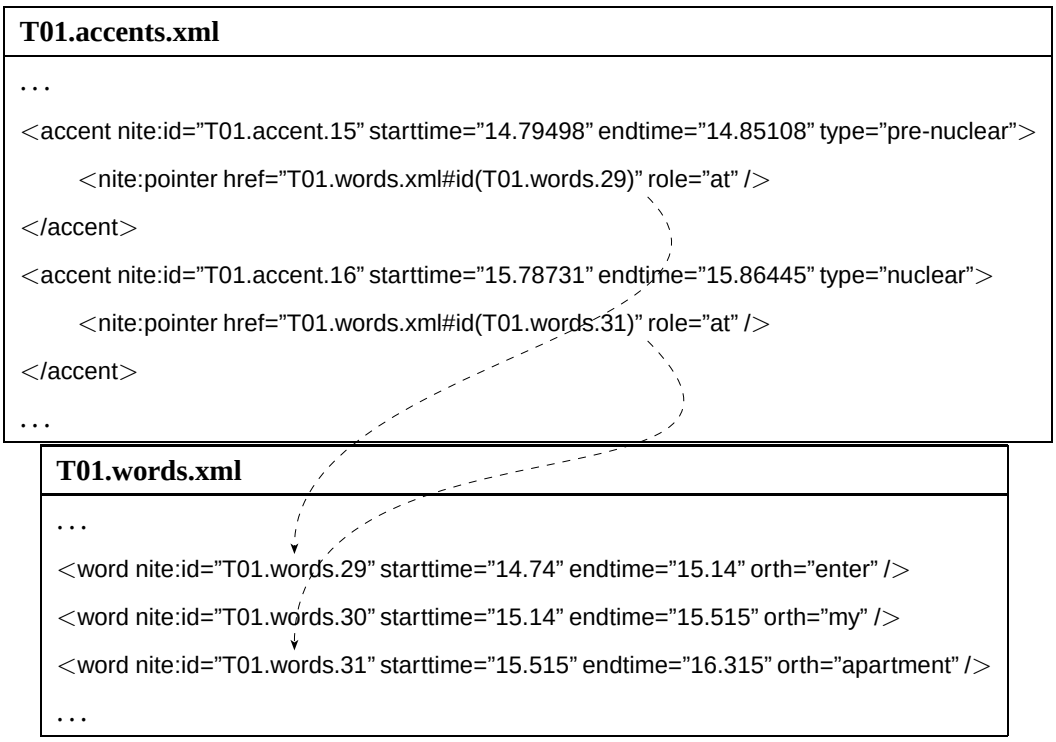



Figure 28: NXT Coding of Accents Associated with Words

relationship with other layers.

## 4.2 Depicting Gestures

### 4.2.1 Aim and Method

Our definition of speech-and-gesture alignment (see Section 1.4) in terms of the syntactic structure of the speech phrase involves analysing the *linguistic properties* of the speech signal that gesture temporally overlaps rather than using timing alone. So the alignment configurations are determined from prosody and syntax, and this is what sets apart our study from previous work where speech-and-gesture integration is a matter of temporal synchrony alone.

With respect to depicting gestures, our central claim and our hypotheses are as follows:

**Central Claim 4.2.1.** Attaching depicting gesture only to the temporally synchronous speech phrase cannot account for the range of possible gestural denotations, and hence for the range of semantic relations between speech and gesture (recall from Section

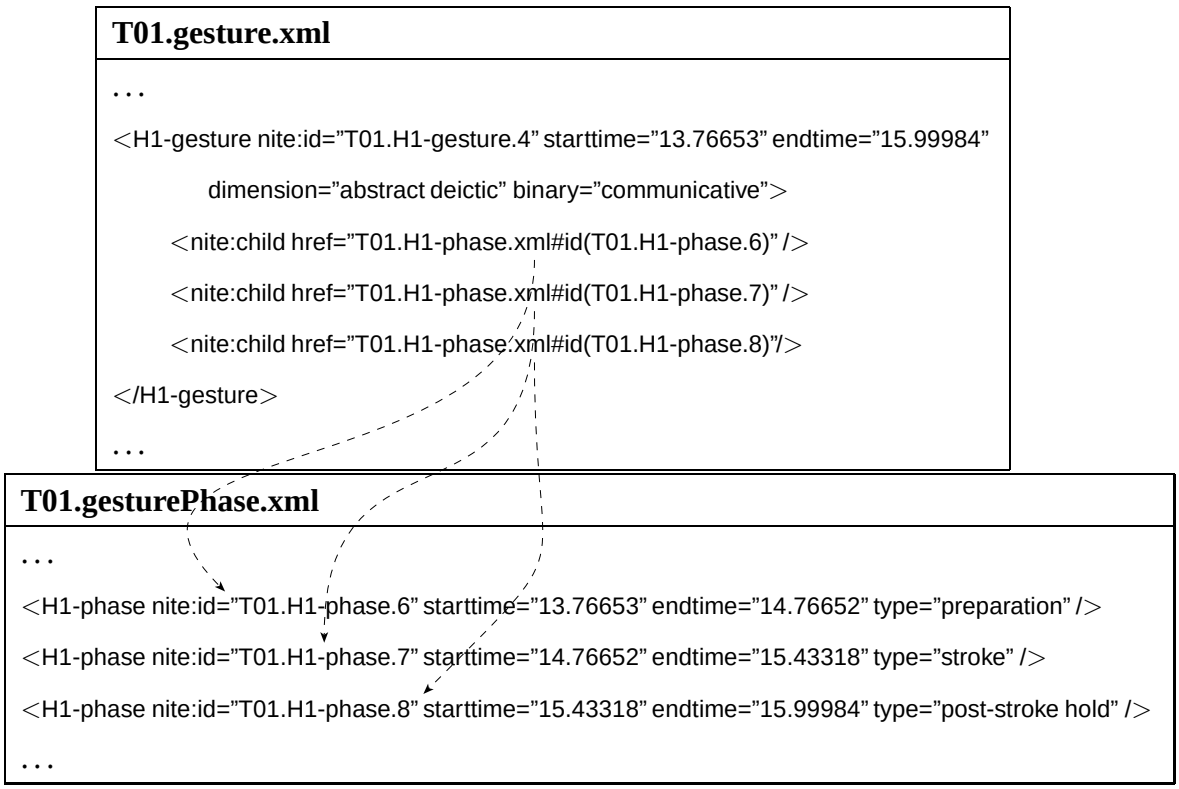



Figure 29: NXT Coding of Gesture and Gesture Phases

2.2.1.1 that computing the semantic relation between speech and gesture and resolving gesture meaning are co-dependent tasks). We approach the range of meanings by allowing for multiple attachments in the syntax tree. Each choice of attachment determines a different speech argument to which the gesture is connected with an underspecified coherence relation—$vis\_rel(s,g)$. This in turn, in the tradition of all coherence-based models of the semantics/pragmatics interface, constrains what the gesture refers to — each entity $g$ denoted by the gesture must be related to an entity introduced in the content $s$ of the aligned speech phrase. Our conjecture is that the choices of attachment and hence ultimately the choices of what the gesture means, are determined by the prosody and syntax of the speech phrase that the gesture temporally overlaps. We assume that there is a major distinction between how timing affects speech-gesture alignment vs. how prosody and syntax affect speech-gesture alignment: timing permits a single attachment in the derivation tree, and prosody and syntax allow for non-unique attachments. In other words, whereas the timing of gesture relative to the timing of speech forces one-to-one mappings, prosody and syntax allow for one-to-many mappings by considering prosodic and/or syntactic constituents

whose temporal performance might not be identical to the temporal performance of the gesture but whose content is still semantically related to the gesture content. We argue for multimodal alignment within the syntactic grammar on the grounds that within the grammar we can spell the constraints coming from various levels of linguistic form.

**Hypothesis 4.2.1.** *The temporal overlap between the speech signal and the gesture signal is not sufficient per se for determining the full spectrum of possible speech-and-gesture alignments. It is the phonetic prominence of the temporally overlapping speech signal rather than the temporally overlapping speech signal alone that guides the decisions of what speech phrase(s) gesture could align with, given our definition of alignment (see Section 1.4).*

**Hypothesis 4.2.2.** *The stroke of depicting gestures can be predicted from the nuclear accent in speech both in the default case of broad focus, and in the case of narrow focus. In case of early pre-nuclear pitch rise, depicting gesture happens along the pre-nuclear pitch accent.*

**Hypothesis 4.2.3.** *The stroke of depicting gestures is not constrained to a particular syntactic category. Likewise, it can attach to distinct syntactic phrases. We hypothesise that like purely linguistic signals, multimodal signals exhibit syntactic (structural) and semantic ambiguity.*

We addressed each hypothesis by a separate search in the NXT tool. The specific queries can be seen in Appendix B, Section B.1. We first searched for the number of gesture strokes temporally overlapping a pitch accent. Since we were interested in the hand movements that communicate meaning, we excluded the pitch accents that happened while performing non-communicative hand movements, the so called adaptors. Since Hypothesis 4.2.1 concerned the presence/absence of prominence while performing the depicting gesture stroke without respect to the accent type, we counted all accents on equal basis.

Hypothesis 4.2.2 stated that it is not the low or high tonal pitch accent, bur rather the nuclear accent that interacts with the gesture stroke. To test that, we augmented Loehr's [2004] corpus with nuclear, pre-nuclear and non-nuclear accents following the prosody annotation schema presented in Section 4.1.2, and made a comparative study between the distribution of tonal pitch accent types and the distribution of nuclear accents across gesture strokes.

Finally, Hypothesis 4.2.3 concerns the interaction between gesture stroke and syntax, and we therefore parsed the linguistic data with the Stanford parser [Klein and

Manning, 2003] so as to assign syntactic labels to phrases synchronous with the gesture strokes. This analysis was preceded by a pre-processing 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 so as to handle incomplete grammatical slots. The assignment of syntactic attributes to gesture was driven by the temporal performance of gesture relative to speech, that is, we looked at the type of overlap relation between gesture and speech. In general, we observed the following three (not necessarily exclusive) temporal relations of a gesture (G) overlapping the relevant spoken word(s) (S):

1. *Inclusion* where $start(G) \prec start(S)$ and $end(G) \succ end(S)$: the temporal interval of the spoken words is included in the temporal performance of the stroke

2. *Precedence* where $start(G) \prec start(S)$ and/or $end(G) \prec end(S)$: the beginning and/or the end of the stroke precedes the beginning and/or the end of the spoken word, i.e., the spoken word begins at some midpoint of the gesture, and/or the gesture ends at some midpoint of the spoken word.

3. *Sequence* where $start(G) \succ start(S)$ and/or $end(G) \succ end(S)$: the beginning and/or the end of the gesture stroke follows the beginning and/or end of the spoken word, i.e, the gesture begins at some midpoint of the spoken word and/or the spoken word ends at some midpoint of the gesture stroke.

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 were captured. Also, if the words overlapping the gesture did not form a syntactic constituent, this was labelled as a "Non-constituent". Finally, a gesture that started at midpoint of $word_1$ and finished at midpoint of $word_2$ was annotated in terms of the label of $word_1$, $word_2$ and their common syntactic label (if available).

### 4.2.2 Results and Discussion

**Experiment 1** The first experiment searched for the number of gesture strokes (G-stroke) overlapped by a Pitch Accent (PA). The results are summarised in Table 9.

| **Temporal Overlap** | **Number** | **Percent** |
|---|---|---|
| Total G-strokes | 95 | |
| G-stroke and PA | 71 | 74.74% |
| G-stroke and PA-word | 79 | 83.16% |
| G-stroke and PA-word $\leq 0.275$ sec | 91 | 95.79% |

Table 9: Temporal overlap between pitch accents and depicting gesture strokes based on the original corpus annotation [Loehr, 2004]

Here we were interested in the presence of a phonetic peak in relation to the temporally overlapping gesture stroke, following the original corpus annotation [Loehr, 2004]. We did not measure the specific timing of the accent relative to the timing of the gesture stroke, that is, we did not calculate whether the stroke started earlier or later than the corresponding accent, and also whether the stroke ended earlier or later than the corresponding accent. We believe that this fine-grained distinction is not informative for the purposes of constraints on the multimodal alignment. This is also in line with previous formal models of multimodal syntax where *temporal overlap* is used as a condition on the integration (e.g., Johnston [1998a] uses temporal overlap or speech following gesture within four seconds).

We found that 74.74% of the gesture strokes were overlapped by (at least one) pitch accent irrespective of its tone. Since, however, gesture often begins/ends at the mid-point of the word associated with it, we then calculated the gesture strokes overlapping pitch-accented words and not the accents alone. Under this condition, we obtained temporal overlap of 83.16%. Finally, we performed a fuzzy match test which involved relaxing the overlap by plus/minus 275 msec. Recall from Section 3.2.3 that this number refers to the proximity of gesture annotation and intonation annotation [Loehr, 2004]: it is the standard deviation of the distribution of accents near gestures, and so this meant that gesture events that occurred within 275 msec from intonation events were considered proximal. Under this condition, we achieved gesture stroke–pitch accented word overlap of 95.79%. The non-matching cases included three strokes that were performed in pauses and/or speech disfluencies, and one event performed with a delay of 290 msec. We examined the data manually, and we established that none of the words performed within these extra milliseconds crossed a constituency boundary: for instance the pitch was on the pre-head modifier or on the complement of the head

synchronous with the gesture stroke.

This test reveals a tendency for gestures to co-occur with the phonetically prominent words in speech, and not with the accent itself. The fuzzy overlap condition supports that the gesture can be reliably associated with a phrase larger than the single prosodically prominent lexical item whose temporal performance is identical to the temporal performance of the gesture stroke. This also dovetails with the findings of Giorgolo and Verstraten [2008] where a delay $\leq$ 500 msec did not affect perceiving speech and gesture as well-formed, and also with the descriptive studies detailing that gestures are synthetic and so the form and meaning of a single gesture corresponds to more than one lexical item in speech [McNeill, 2005].

**Experiment 2** Along the lines of the previous study on prosody [Calhoun, 2006], our second hypothesis concerns the interaction (or lack thereof) between tonal pitch accent types (low and high) and gesture strokes. The results of our experiment investigating the distribution of accents across gesture strokes are displayed in Table 10. The results indicate that although gesture strokes tend to align with H* accents (see the summary in Table 11), no reliable generalisations can be drawn from these data. It is also the very high occurrence of underspecification and uncertainty events (marked with X*) that prevents us from constraining gesture to a specific tonal pitch accent type.

Our second hypothesis therefore stated that the gesture stroke realisation can be predicted from the nuclear prominence in speech, i.e., gesture maps to the metrical structure of the utterance. Table 12 shows the results of this search. Since our hypothesis concerned the interaction between nuclear/pre-nuclear prominence and gesture stroke, we can summarise the findings as follows (see Table 13): 78.67% of the events overlapped at least one nuclear and/or pre-nuclear accent, and 89.33% of the events overlapped at least one nuclear and/or pre-nuclear accented word. We then verified manually the events where the gesture stroke was overlapped by a non-nuclear accent, and we established that there was a nuclear accent that overlapped the post-stroke hold and/or the nuclear accent was on the post-head modifier or the head complement. This is not surprising given the right-branching bias where nuclear accents are usually at the end of the phrase (see, for instance, Calhoun [2010]). We also found one narrow focused event where the nuclear accent was on the verb head preceding the stroke. We finally performed a fuzzy match search within 275 msec, and we reached 100% overlap with this condition.

This raises the question how to interpret those results from the perspective of con-

| Temporal Overlap | Number | Percent |
|---|---|---|
| Total G-strokes and PA | 71 | |
| G-stroke and H* | 27 | 38.03% |
| G-stroke and X* | 14 | 19.72% |
| G-stroke and L+H* | 9 | 12.67% |
| G-stroke and !H* | 5 | 7.04% |
| G-stroke and L* | 3 | 4.23% |
| G-stroke and L+!H* | 2 | 2.82% |
| G-stroke and H*, L* | 1 | 1.41% |
| G-stroke and H*, H* | 3 | 4.23% |
| G-stroke and H*, X* | 2 | 2.82% |
| G-stroke and L+H*, X* | 1 | 1.41% |
| G-stroke and H*, L+H* | 2 | 2.82% |
| G-stroke and !H*, H*, X* | 1 | 1.41% |
| G-stroke and H*,?,L+H* | 1 | 1.41% |

Table 10: Distribution of tonal pitch accent types across depicting gesture strokes based on the original corpus annotation [Loehr, 2004]. The classes are mutually exclusive: the total number of strokes overlapped by an accent equals 100%.

| Temporal Overlap | Number | Percent |
|---|---|---|
| Total G-strokes and PA | 71 | |
| G-stroke and (at least one) H* | 37 | 52.11% |
| G-stroke and (at least one) X* | 18 | 25.35% |
| G-stroke and (at least one) L+H* | 13 | 18.31% |
| G-stroke and (at least one) L* | 4 | 5.63% |
| G-stroke and (at least one) !H* | 6 | 8.45% |
| G-stroke and (at least one) L+!H* | 2 | 2.82% |
| G-stroke and (at least one) ? | 1 | 1.41% |

Table 11: Summary of the distribution of tonal pitch accent types across depicting gesture strokes based on the original corpus annotation [Loehr, 2004]. The classes are not mutually exclusive: the total number of strokes overlapped by an accent is greater than 100%.

| Temporal Overlap | Number | Percent |
|---|---|---|
| Total G-strokes and PA marked with its type | 75 | |
| G-stroke and nuclear PA | 38 | 50.67% |
| G-stroke and pre-nuclear PA | 8 | 10.67% |
| G-stroke and non-nuclear PA | 15 | 20% |
| G-stroke and nuclear, non-nuclear PA | 8 | 10.67% |
| G-stroke and nuclear, nuclear PA | 1 | 1.33% |
| G-stroke and non-nuclear, non-nuclear PA | 1 | 1.33% |
| G-stroke and nuclear, non-nuclear, pre-nuclear PA | 2 | 2.67% |
| G-stroke and nuclear, non-nuclear, non-nuclear PA | 1 | 1.33% |
| G-stroke and nuclear, nuclear, non-nuclear PA | 1 | 1.33% |

Table 12: Distribution of nuclear prominence across depicting gesture strokes. These classes are mutually exclusive: the total number equals 100%.

| Temporal Overlap | Number | Percent |
|---|---|---|
| Total G-strokes and PA marked with its type | 75 | |
| G-stroke and nuclear and/or pre-nuclear PA | 59 | 78.67% |
| G-stroke and nuclear and/or pre-nuclear accented word | 67 | 89.33% |

Table 13: Summary of the distribution of nuclear prominence across depicting gesture strokes. These classes are not mutually exclusive: the total number is greater than 100%.

straining the alignment of speech and depicting gesture in the grammar. First, we shall account for the coordination between nuclear prominence and gesture stroke via a construction rule that constrains the choices of alignment to the nuclear and/or pre-nuclear prominent element, but that does not impose any restrictions on the tonal pitch accent type. Further, our definition of alignment is based on the syntactic structure of the utterance, rather than on relative timing alone. We shall therefore define grammar rules where gesture aligns with a constituent structure where the head or any of its subcategorised arguments are (pre-)nuclear prominent. What the specific accents of the rest of the constituent elements are, and whether they are prosodically marked at all, is not a matter of concern because we still take care to attach gesture to a constituent structure (that is also a metrical tree).

| Syntactic Category of G | Percent | Syntactic Category of G | Percent |
|---|---|---|---|
| S | 5.48% | RB | 8.22% |
| VP | 12.3% | TO | 2.74% |
| V (present and past verb forms, base forms, modal verbs, present and past participles) | 30.14% | JJ (positive and comparative adjectives) | 6.85% |
| | | DT | 13.7% |
| NP | 17.81% | UH | 1.37% |
| NN (singular and plural) | 10.96% | C (coordinating or subordinating conjunction) | 5.48% |
| PRP (personal and possessive) | 20.55% | | |
| IN | 5.48% | Pause | 6.85% |
| PP | 1.37% | Non-constituent | 5.48% |

Table 14: Temporal overlap between depicting gesture strokes and syntax based on the original gesture annotation [Loehr, 2004]. Since every gesture potentially maps to more than one syntactic category (because the gesture may be aligned to a multi-word phrase), the total number of labels is greater than 100%.

**Experiment 3** Our final hypothesis concerned the interaction between gesture and syntax. Our findings concerning the syntactic attributes assigned to the speech elements synchronous with the gesture strokes are summarised in Table 14. The results demonstrate that 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, we observed that when a gesture overlaps a verb head, the ambiguous form of the hand signal often does not fully constrain the attachment of gesture to the head only. This attachment ambiguity is observed with gestures spanning a verb only, a verb phrase, or an entire sentence, thereby allowing for mappings beyond the strict temporal performance. To illustrate this, consider again utterance (2.8), repeated below, where the gesture stroke overlaps an entire sentence.

(4.2) So $[_{H^*}$he mix]es $[_{X^*}$mud] ...
*Speaker's left hand is rested on the knee with palm open supine. The right hand is held loose with fingers facing downwards over the left hand. The speaker performs consecutively four rotation movements with her right hand 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 a salient property of the mud, namely that it was going round, 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. Essentially, these ambiguities would also arise if the gesture was performed while uttering “mixes” only or even “he mixes”.

To account for these multiple possibilities, in the grammar we define rules where the multimodal phrase can be derived by attaching gesture not only to the synchronous prosodically prominent element, but also to its higher projections no matter whether the gesture happened along with, say, the head, its arguments or the entire clause. In this way, we capture two observations: first, the range of possible alignment configurations cannot be obtained solely in terms of timing, i.e., the incomplete meaning of gesture as derived from form allows for ambiguities; second, the attachment to a higher projection 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 multimodal utterance can thus receive more than one correct parse analysis where each one contributes a distinct denotation of the gesture, and hence a distinct relation between the speech and the gesture.

We argue that the same principle of exploring the speech-gesture alignment beyond the identical timings 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 (4.3) [McNeill, 2005, p. 23], Figure 30 demonstrates that gestures can be extended over the preposition head arguments.

(4.3) and he goes up [$_{X^*}$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 multimodal 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

Figure 30: Depicting Gesture along with the utterance “And he goes up through the drainpipe”, example (4.3) [McNeill, 2005]

the synchronous elements by narrowing down to a specific content. Our claim for the non-unique gesture attachment possibilities would also favour an attachment to a larger phrase containing the object, “through the drainpipe”. We argue that both analyses are legitimate and should be obtainable by the grammar so as to provide the necessary underspecified relations resolvable by contextual knowledge.

Similarly, in case of gestures overlapping non-head daughters, the multimodal 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. For instance, recall (1.6), page 16 where the conduit interpretation of the hand signal was available only after linking the gesture with the verb head “teach” even though it was performed while uttering the pre-head modifier “really”. Likewise, in (2.4), page 35, the interpretation where the hand movement represents literally the bottom cupboards can be obtained by attaching the gesture to the overlapping subject daughter, and the gesture denoting the metaphor of completing some process is possible only by an attachment to the S node.

The empirical study also demonstrated that while prosody can make a multimodal utterance ill-formed, in syntax there are 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.

| Temporal Overlap | Talkbank | | AMI | |
|---|---|---|---|---|
| | **Number** | **Percent** | **Number** | **Percent** |
| Total G-strokes | 82 | | 22 | |
| G-stroke and PA | 74 | 90.24% | 18 | 81.82% |
| G-stroke and PA-word | 82 | 100% | 22 | 100% |

Table 15: Temporal overlap between deictic gesture and nuclear prominence

## 4.3 Deictic Gestures

### 4.3.1 Aim and Method

For deictic gestures, our empirical study proceeded similarly to depicting gestures with the main difference being that the intonation annotation did not include the tonal pitch accent types, but it was entirely based on the metrical phonology framework. Our central claim and our hypothesis about deictic gestures are as follows:

**Central Claim 4.3.1.** Deictic gesture realisation is essential for identifying salience in the context of speech and in the context of the communicative situation as well. We consider that salience is expressed by synchronising the meaningful part of the gesture realisation, the stroke, with the meaningful prosodically prominent elements in speech.

**Hypothesis 4.3.1.** *The performance of deictic gestures overlaps the performance of nuclear accented words in speech both in the default case of broad focus, and in case of narrow focus. In case of early pre-nuclear rise, the performance of deictic gestures overlaps the pre-nuclear pitch accented words.*

### 4.3.2 Results and Discussion

The empirical study of deixis and speech was performed using the NXT tool (the queries are included in Appendix B, Section B.2). We first searched for the number of deictic strokes overlapped by a pitch accent of any type. The results, summarised in Table 15, show that in the Talkbank observation, 90.24% of the strokes were overlapped by a pitch accent. This number increased to 100% when we counted the strokes overlapping not simply the pitch accent, but rather a pitch accented word. For the AMI observation, these numbers are 81.82% and 100%, respectively.

| Temporal Overlap | Talkbank | | AMI | |
|---|---|---|---|---|
| | **Number** | **Percent** | **Number** | **Percent** |
| Total G-strokes | 82 | | 22 | |
| G-stroke and nuclear PA-word | 33 | 40.24% | 7 | 31.82% |
| G-stroke and pre-nuclear PA-word | 4 | 4.88% | 2 | 9.09% |
| G-stroke and non-nuclear PA-word | 0 | 0% | 0 | 0% |
| G-stroke and non-nuclear PA-word, nuclear PA word | 26 | 31.71% | 6 | 27.27% |
| G-stroke and nuclear PA-word, nuclear PA word | 11 | 13.42% | 3 | 13.64% |
| G-stroke and pre-nuclear PA-word, nuclear PA word | 2 | 2.44% | 0 | 0% |
| G-stroke and non-nuclear PA-word, nuclear PA word, non-nuclear PA word | 2 | 2.44% | 0 | 0% |
| G-stroke and nuclear PA-word, nuclear PA word, non-nuclear PA word | 3 | 3.66% | 0 | 0% |
| G-stroke and pre-nuclear PA-word, nuclear PA-word, nuclear PA-word | 1 | 1.22% | 2 | 9.09% |
| G-stroke and none, non-nuclear PA-word | 0 | 0% | 1 | 4.55% |
| G-stroke and nuclear PA-word, nuclear PA-word, nuclear PA-word | 0 | 0% | 1 | 4.55% |

Table 16: Distribution of accent types across deictic strokes. These classes are mutually exclusive: the total number is equal to 100%.

We addressed our hypothesis by studying what the distribution of accent types is across deictic strokes. The results, displayed in Table 16, can be summarised with the following observation (see Table 17): in the Talkbank corpus all deictic strokes coincided with at least one nuclear-accented word and/or a pre-nuclear accented word. In the AMI meeting, the overlap was 95.45%. Strokes overlapping a combination of non-nuclear and nuclear accented words were also common.

This experiment confirmed the expected co-occurrence between the nuclear prominent word (not simply the nuclear accent) and the gesture stroke, and also between the pre-nuclear prominent word and the gesture stroke. We will use utterance (4.4), Fig-

| **Temporal Overlap** | **Talkbank** | | **AMI** | |
|---|---|---|---|---|
| | **Number** | **Percent** | **Number** | **Percent** |
| Total G-strokes | 82 | | 22 | |
| G-stroke and at least one nuclear/pre-nuclear PA-word | 82 | 100% | 21 | 95.45% |

Table 17: Summary of the distribution of accent types across deictic strokes

ure 31 and utterance (4.5), Figure 32 to illustrate our findings.

(4.4) I keep [$_N$going] until I [$_{NN}$hit] Mass [$_N$Ave], I think
*Right arm is bent in the elbow at a 90-degree angle, right hand is loosely closed and relaxed, fingers point forward. Left arm is bent at the elbow, held almost parallel to the torso, palm is open vertical facing forward, finger tips point to the left.*

(4.5) And then I [$_N$turn] *[pause]* [$_{PN}$left] on Mass [$_N$Ave]
*Left hand is held in the same position as in* (4.4)*, then along with "left" the right hand moves to the left periphery over the left hand, right hand stays vertically open.*

(4.4) is a broad-focused utterance with the nuclear accent being on the right-most word. Utterance (4.5), which is a continuation of (4.4), displays a pre-nuclear accent on "left" and then there is a downstepped nuclear accent on the right-most word. Recall from the annotation guidelines (Section 4.1.2) that pre-nuclear accents usually precede downstepped pitch accents where the downstepped accent marks an inferable piece of information. With this in mind, the interaction between prosodic prominence and gesture stroke appears to be on the level of information structure: nuclear prominence, along with gesture stroke happens along with the focused (kontrastive)[6] elements that push the communication forward, and not with those available from the background. In case of pre-nuclear accent, there is a gestural re-enforcement along with the pre-nuclear accent and not with the downstepped accent. This finding also dovetails with the observations in the descriptive literature where "a break in the continuity" [Givón,

[6]In the Information Structure literature kontrast designates "parts of the utterance—actually, words—which contribute to distinguishing its actual content from alternatives the context makes available." [Kruijff-Korbayová and Steedman, 2003]

Figure 31: Hand Gesture Placing a Landmark in the Virtual Space, example (4.4)

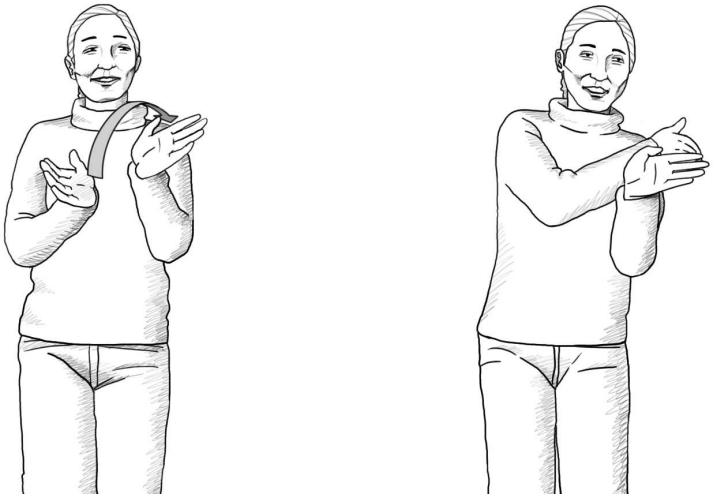


Figure 32: Hand Gesture Pointing at a Landmark in the Virtual Space, example (4.5)

1985] of the narrative implies “highest degree of gesture materialisation” [McNeill, 2005, p. 55] (recall Section 3.2.4).

Likewise, in (2.12), repeated in (4.6), the pre-nuclear prominence was synchronous with the stroke, and then the content was elaborated in speech while holding the hands in an expressive position.

(4.6) 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.*

These results report on the interaction between speech and deixis on the level of form. Our overall aim is to account for the syntactic and the semantic well-formedness of the multimodal signal. In other words, the underspecified logical formulae that we produce from the syntactic tree should provide an abstract description of what the multimodal action means in the particular discourse context. Our empirical investigation therefore proceeded with an analysis of whether a syntactic attachment to the nuclear/pre-nuclear accented word would also produce the semantically preferred interpretation in context. We encountered seven multimodal utterances which, although syntactically well-formed, failed to map to the intended meaning representations due to one of the following reasons:

1. The performance of the deictic stroke takes place before or after uttering the semantically related speech signal; e.g., in (4.7), Figure 33 the deictic gesture is performed along with the prominent "Thank you" when obviously the denotation of the gesture is identical to that of the speech NP "the mouse". The alternative interpretation where the gesture signal and the speech signal are bound through a causal relationship, i.e., the act of the handing the mouse is the reason for thanking the addressee is not possible since "Thank you" is related to what came in the previous discourse—projecting the presentation in slide show mode in response to the speaker's request.

   (4.7) [$_{N}$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*

   Likewise, in utterance (4.8), Figure 34 the pointing of the hand was executed in the direction of the participant denoted by "she", but the temporal performance of this deictic gesture happened along with "cubicle next" and not "she" or any phrase containing "she".

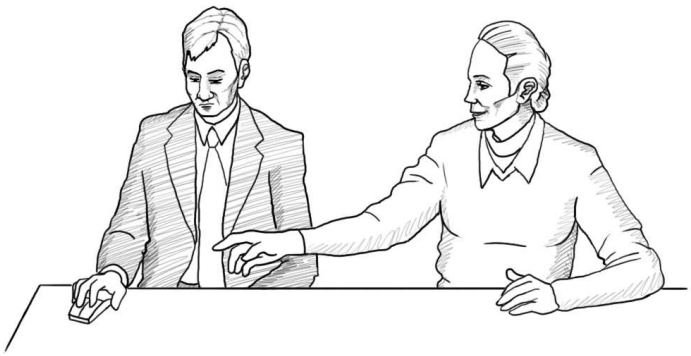


Figure 33: Pointing Gesture towards a Computer Mouse, example (4.7)

(4.8) . . . [$_{PN}$She] works in the [$_{NN}$cubicle] [$_{N}$next] to me . . .
*Speaker's right hand is loosely open, points at the participant diagonally from the speaker.*

Another example that illustrates this mismatch is presented in (4.9), Figure 35. While an alignment with the nuclear prominent "splinters" that the gesture temporally overlaps with could provide an interpretation where the pointing identifies the splinters in the hand rather than the hand directly, this analysis is not the only option. Intuitively in this utterance, the deictic signal is performed in the direction of the hand, and so the semantically preferred element is "hand".

(4.9) . . . you wouldn't wanna have to have [$_{N}$splinters] in your [$_{N}$hand] while you're using your *[disfmarker]*
*Both palms are open flat oblique, right points to the left palm, the speaker is looking at her palm while pointing.*

2. The speech signal that is semantically related to the gesture is not prosodically prominent; e.g., in (2.3), repeated in (4.10), the deictic gesture overlaps tem-

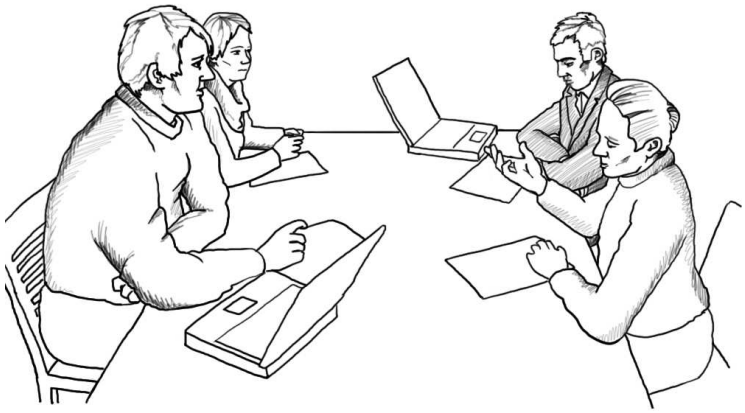


Figure 34: Pointing Gesture towards the Other Participant, example (4.8)

porally the nuclear prominent “said”, when in fact, it identifies the individual pointed at and it thus resolves the pronoun “she” coming from speech.

(4.10) 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 the same way, the gesture in (4.11), Figure 36 is obviously related with the non-prominent “he”.

(4.11) What do you think Ed? Do you [disfl] he [$_{NN}$liked] the [$_N$display] in one of the concepts that you showed
*Speaker’s right hand is loosely open, almost vertical, it points in the direction of the participant in front of the speaker.*

Recall from Section 2.2.1.3 that we use the constant $\vec{p}$ to account for the physical space denoted by deixis form which is used to determine the space $v(\vec{p})$ located by the deictic referent [Lascarides and Stone, 2009b]. Essentially, $\vec{p}$ is not equal to $v(\vec{p})$

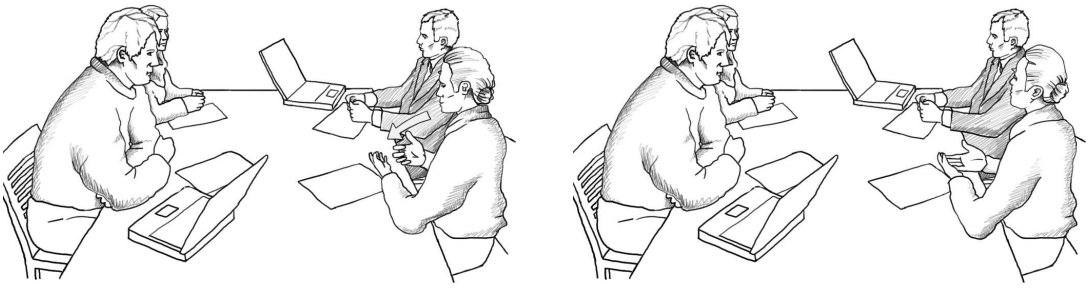


Figure 35: Pointing Gesture with the Right Hand towards the Left Hand, example (4.9)

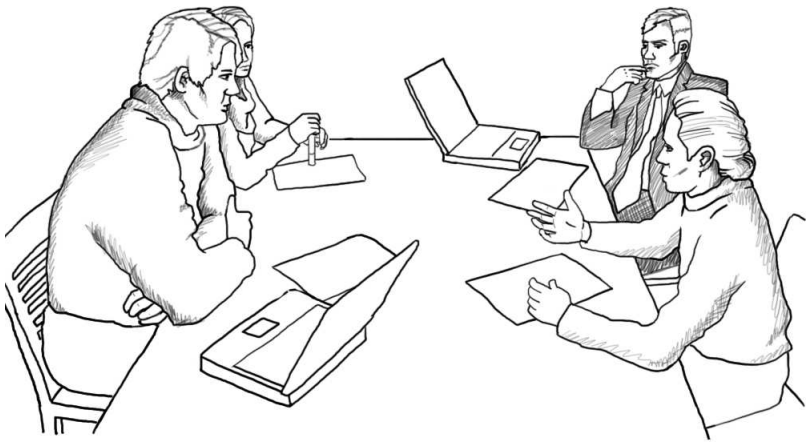


Figure 36: Pointing Gesture towards the Participant in front of the Speaker, example (4.11)

in cases where the referent introduced in the gesture space is not physically present. Conversely, $\vec{p}$ equals $v(\vec{p})$ when the referent introduced by the gesture is at the physical

coordinates identified in the gesture space.

With this in mind, we observed that the temporal/prosodic mismatch occurred only in cases where the visible space $\vec{p}$ designated by the gesture was *equal* to the space $v(\vec{p})$ it denoted, that is, the function $v$ that maps the space identified by gesture to the actually denoted space resolves to equality. In the multimodal grammar, we shall capture this finding by aligning gesture to a spoken word that is not prosodically marked and/or that happened outside the temporal performance of the gesture if the deictic referent is at the exact coordinates identified by the pointing hand. The alternative option of aligning a speech element and a deictic gesture that denotes an individual not present at the exact coordinates in the gestured space would not produce the intended meaning representation in the specific context. To clarify this finding, consider again utterance (2.11), page 48. It is perfectly acceptable for the gesture stroke to be performed a few milliseconds later so that it overlaps "come" or even "tropical countries" without blocking the interpretation where the hand denotes the addressee. In contrast, if the deixis in utterance (2.12), page 49 was performed along with "I", the logical form would fail to resolve to "apartment".

From our starting hypothesis 4.3.1, we have found that the interaction between prosodic prominence and deictic stroke does not necessarily hold for deictic gestures pointing at objects salient in the physical space. This is unlike abstract gestures that create a virtual object in the frontal space. This finding flags up an essential difference between how deictic gestures that identify concrete individuals in the physical space integrate with the semantically related speech elements vs. how deictic gestures that point at virtually created individuals relate with the corresponding speech elements. We assume that from a perception perspective, the flexibility in relating concrete deixis with the semantically related speech element is compensated for by salience of the designated individuals. In contrast, with abstract deixis the designated individuals are not physically present and so establishing which speech elements they correspond to is dependent on the stronger interaction (through timing and prosody) with the speech elements.

Compared to depicting gestures, we do not expect any different behaviour of deixis with the syntax of the temporally overlapping speech signal, and we therefore did not analyse the correspondence between deixis and syntax.

| Temporal Overlap | Number | Total | Percent |
|---|---|---|---|
| G-stroke and Beats | 4 | 36 | 11.11% |

Table 18: Temporal overlap between (depicting and deictic) gestures and beats. The adaptors were excluded from the search.

## 4.4 Beats

Our annotation framework included marking beat events so as to study whether these formless up-and-down movements always superimpose other gesture dimensions, or stand-alone beats are also possible. This question is essential with respect to the formalisation of beats in the grammar. If beats always occur within another gesture dimension, then we can formalise them as a boolean feature on the gesture, that is:

$$\begin{bmatrix} \textit{communicative} & \\ \text{BEAT} & \text{bool} \end{bmatrix}$$

Alternatively, if pure beats occur, then in the gesture type hierarchy we shall represent them as a subtype of the gesture sign, similarly to the other gesture dimensions.

We used the NXT tool to perform our search (the specific queries are included in Appendix B, Section B.3). The results of the search, summarised in Table 18, indicate that only 11.11% of the beat events happened while performing a gesture stroke. We will capture this finding by defining beats as a leaf in the gesture type hierarchy, and not as a boolean feature of the gesture (see Section 6.3).

## 4.5 Conclusions and Next Steps

Through a series of experiments over annotated multimodal data, we reached the following conclusions: first, the prosodic prominence of the speech signal is an essential factor for reliably predicting the performance of the gesture stroke; second, the metrical prosodic structure rather than the tonal pitch accent type interacts with the gesture performance—so information about pre-nuclear and nuclear accents needs to be included in the multimodal grammar. We also found that while prosody strictly constrains the syntactic attachments, in syntax there are non-unique choices of which phrase gesture could attach to so that the multimodal action produces the intended meanings in context. Our study of the interaction between deictic gesture and speech

also revealed that temporal and/or prosodic mismatch between the deictic signal and the semantically related speech signal can occur when the physical space the gesture points at is identical to the gesture denotation, i.e., the individual introduced by the gesture is located at the exact spatial coordinates identified by the pointing hand.

In the next chapter, we provide grammar rules that reflect these generalisations about the corpus data.

# Chapter 5

# A Grammar for Speech and Co-Speech Hand Gesture

*Grammar is to meaning as the law is to good behaviour.*
Adam Kilgarriff

In the last chapter, we demonstrated through a series of empirical studies that gestures interact with the prosodic form of the temporally overlapping speech signal, in turn having effects on the multimodal meaning. The fact that the interaction is on the level of form motivates aligning speech and gesture in the grammar. In this chapter, we lay out the formal theory of the speech-and-gesture alignment which guides the production of Underspecified Logical Formulae (ULFs) supporting the pragmatic interpretations of multimodal actions in context. We begin in Section 5.1 by detailing the resolved logical forms that capture the gesture interpretations in context. This section will motivate the attachment ambiguities, licensed by gesture form, which ultimately constrain how gesture is interpreted given the current models of constructing logical forms in discourse. In Section 5.2 we argue for a symbolic representation of gestural form and its mapping to an underspecified meaning representation. Then in Section 5.3, we motivate and describe grammar construction rules which, driven from the empirical generalisations from Chapter 4, account for multimodal grammaticality, thereby producing (underspecified) logical formulae resolvable to plausible interpretations at the semantics/pragmatics interface.

## 5.1 Gesture Interpretations in Context

Our programmatic approach involves producing (underspecified) logical forms of aligned speech-and-gesture actions that are supported by existing theories of discourse at the semantics/pragmatics interface (e.g., Grosz and Sidner [1986], Kamp and Reyle [1993], Kehler [2002], Asher and Lascarides [2003], and many others). There are two key features to these theories. First, they argue that the meaning of extended discourse is defined in terms of semantic relations that connect the contents of its segments together; furthermore, because an extended segment, consisting of semantically related sub-segments, can itself be an argument to such a semantic relation, this engenders a *hierarchical discourse structure* of semantically related discourse segments and subsegments.[1] The second key feature is that they impose *constraints* on the available references in the logical form to act as antecedents to anaphoric expressions: for instance, an anaphoric pronoun in the current discourse unit must co-refer to an antecedent that is present in the current unit or present in the content of a discourse unit to which the current unit is semantically related in the discourse structure (e.g., Hobbs [1985], Grosz and Sidner [1986], Webber [1991], Asher and Lascarides [2003]).[2]

Recall that *all* individuals and events that are introduced by gesture receive an interpretation only through their context of use—in this sense, they behave like anaphoric expressions. So carried over to multimodal actions, the above constraint that the discourse context imposes on the interpretation of anaphoric elements in a discourse segment means that any individual or event that is a part of the content of a gesture has to be related via some bridging relation to an antecedent that is present in the content of a speech phrase to which the gesture is semantically related. So our meaning representations, which we will derive from the form of the multimodal action, must respect this constraint on pragmatic interpretation. To achieve this, we make the choices of syntactic attachment determine which speech phrase the gesture aligns with, and hence the speech phrase the gesture is semantically related to. This in turn affects the availability of referents that are introduced in speech to act as antecedents for resolving the underspecified content of the gesture (given just its form) to a specific and pragmatically

[1] Theories differ as to the types of semantic relations that connect the contents of segments, and also differ on the types of discourse structure they countenance. But these differences do not matter for our purposes.

[2] In addition, many theories assume constraints on which parts of the discourse context the current discourse unit can be semantically related to: namely, the previous discourse unit or one that dominates it in the preceding discourse structure (e.g., Webber [1991], Asher and Lascarides [2003]). So overall, the interpretation of anaphoric expressions is constrained by the structure of the discourse context, and the contents of the segments that are related in that structure.

plausible interpretation. In other words, we will argue for constraints on attachment in the multimodal grammar on the basis of the pragmatic interpretations of the multimodal action that are plausible and the constraints that those interpretations impose on which speech phrase(s) the gesture can be semantically related to (and hence attached to in the syntax tree).

In addition to that, in Section 2.2.1 we demonstrated that the mappings of (ambiguous) gesture form to meaning are one-to-many. Therefore, the logical forms supporting the plausible interpretations and hence the distinct semantic relations between speech and gesture can be derived from distinct speech-gesture alignments, expressed via different syntactic attachments. Given the current models of the semantics/pragmatics interface of how discourse structure constrains reference, the alignments (and hence the specific attachments) block access to certain referents being used for inferring the specific interpretation of gesture in context.

With this in mind, we will provide examples of resolved logical forms which demonstrate how the different alignments capture the pragmatic interpretations of multimodal actions in context, and hence the distinct relations between the speech content and the gesture content. To fit the current research in the broader context of formal semantics of gesture [Lascarides and Stone, 2009b], the resolved logical forms featured in this section make use of the language of Segmented Discourse Representation Theory (SDRT [Asher and Lascarides, 2003]) for interpreting gesture. Of course, the same information can be expressed in any other model of the semantic/pragmatic interface.

### 5.1.1 Interpreting Depicting Gestures

To illustrate the various ways in which depicting gestures can be interpreted in context, consider the gesture in (2.8), repeated in (5.1).

(5.1) So $[_{H^*}$he mix]es $[_{X^*}$mud] ...
*Speaker's left hand is rested on the knee with palm open supine. The right hand is held loose with fingers facing downwards over the left hand. The speaker performs consecutively four rotation movements with her right hand over the left palm.*

Intuitively, one of the possible denotations of this circular hand movement is the mud being mixed. This interpretation is supported by the logical form in (5.2)[3] which

[3]In Section 5.2.2.4, we will refine the logical form of depicting gesture.

features an Elaboration relation between the content of the speech labelled $\pi_s$ and the content of the gesture labelled $\pi_g$: namely, interpreting the circular hand movement as some substance being rotated elaborates on the denotation of "mud" introduced in speech.

(5.2) $\pi_s : (mud(x))$
$\pi_g : \exists x'(substance(x') \wedge rotate(e', x'), x = x')$
$\pi_1 : \exists h, x(he(h) \wedge mix(e, h, x) \wedge Elaboration(\pi_s, \pi_g))$

Given the constraints on reference imposed by the discourse structure (e.g., Asher and Lascarides [2003]), this logical form is supported by a syntactic tree where the gesture aligns with the NP "mud", and hence the referents *h* and *e* introduced by "he" and "mixes" respectively are not accessible for interpreting gesture. Therefore, neither the denotation of "he" nor the denotation of the event "mixes" feature in the resolved content of the gesture. This ultimately means that the gesture is interpreted as depicting the mud (that is being mixed) going round, and hence that the speech and gesture are coherently related via Elaboration. In the resolved logical form, there is identity between the referent introduced by the gesture and the referent introduced by the speech (i.e., $x = x'$). The resolved logical form of the entire utterance labelled with $\pi_1$ can be paraphrased as "There was someone he and he was mixing mud, the mud that was going round". The relation between the speech and the gesture is thus similar to appositives and non-restrictive relative clauses in language.

The gesture can also be interpreted as depicting the event of mud going round as a *result* of the mixing, supporting thus a Result relation between the speech content and the gesture content. The formal rendition of this interpretation is given in (5.3).

(5.3) $\pi_s : \exists x(mud(x) \wedge mix(e, x))$
$\pi_g : \exists x'(substance(x') \wedge rotate(e', x') \wedge x = x')$
$\pi_0 : \exists h(he(h) \wedge Result(\pi_s, \pi_g))$

Unlike (5.2), the gesture qualifies not only the denotation of "mud", but also the denotation of the verb "mixes", and hence it features in the resolved LF. This is rendered by a hierarchical structure where the gesture aligns with the VP "mixes mud", which blocks access to the agent of the mixing event. Here, the speech referent *e* and the gesture referent $e'$ are bridging related via a causal relationship rather than via identity: namely, the mixing event causes the event of the mud going round.

The alternative interpretation where the circular hand movement enacts the event of mixing mud from the agent's viewpoint is featured in the logical form in (5.4).

(5.4) $\pi_s : \exists h,x(he(h) \wedge mud(x) \wedge mix(e,h,x))$
$\pi_g : \exists h',x'(agent(h') \wedge substance(x') \wedge mix(e',h',x')$
$\wedge horizontal\_motion(e'') \wedge e = e')$
$\pi_0 : Depiction(\pi_s, \pi_g)$

In this instance, the alignment of gesture to the entire clause gives access to the referents introduced by the NPs "he", "mud" and the verb "mixes" for interpreting gesture. They therefore feature in the resolved logical form of gesture. Unlike (5.2) and (5.3), the coherence relation Depiction in (5.4) is inferred with the denotations of "he" and "mixes" contained in the LF of gesture. Again, gesture and speech denote the same referents (i.e., $e = e'$).

These gesture interpretations featured identity between the referent of speech and the referent of gesture. However, in Section 3.4.2 we claimed that identity between the speech referent and the gesture referent is not the only option. To demonstrate this, we will consider a slight modification of utterance (5.1), shown in (5.5).

(5.5) So [$_{H^*}$he mix]es [$_{X^*}$mud]?
*Speaker's right hand is vertically open with palm facing up. The speaker moves it forward to the frontal space.*

This gesture does not denote a salient property of the referents introduced in speech: instead, it qualifies the speech act of questioning (this is possible with the additional assumption that the speech phrase was uttered with a rising intonation). A rough paraphrase of the meaning of the multimodal action in (5.5) would be "Are you telling me that he mixes mud?". Interpreting the gesture in this metaphorical way (see the LF in (5.6)), and inferring a Metatalk relation [Polanyi, 1985] between the gesture and the speech act (rather than the content expressed through this proposition) would be supported via an attachment to the entire clause, which gives access to the speech act performed with this utterance.

(5.6) $\pi_s : \exists h,x(he(h) \wedge mud(x) \wedge mix(e,h,x))$
$\pi_g : question(tell(e,you,p) \wedge p = \pi_s)$
$\pi_0 : Metatalk(\pi_s, \pi_g)$

### 5.1.2 Interpreting Deictic Gestures

The interpretation of deictic gestures is analogous: the syntactic attachments determine the individuals and/or events that are accessible for resolving the gesture's denotation,

and hence for inferring a rhetorical connection between the content of speech and the content of deixis. To illustrate the framework, we will be using utterance (2.3), repeated in (5.7).

(5.7) 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.*

The resolved logical forms are featured in (5.8) and (5.9). Deictic gestures anchor the spatial coordinates of an event or an individual which we capture by the predicate $sp\_ref(x, v(\vec{p}))$ where $x$ is the referent introduced by the deixis and $v(\vec{p})$ is the location in space identified by the pointing signal. Attaching the gesture to the NP "she" would support an interpretation where the gesture demarcates the location of the referent of "she" in the physical space (see (5.8)). This in turn would support an Identity relation between the speech referent $s$ and the deixis referent $x$ which is true in case there is an individual $x$ who was located at the spatiotemporal coordinates $v(\vec{p}_x)$. Conversely, attaching the gesture to "she said" would support a different relation such as MetaphoricalCounterpart between the event of saying and the event identified by the deixis since there is no physical individual present at the physical coordinates identified by the pointing gesture (see (5.9)). This relationship is true in case $e$ is a saying event by some person $s$ of something $u$, and the agent of the saying event was located at $v(\vec{p}_s)$.

(5.8) $\pi_1 : \exists m(material(m) \wedge environmentally\text{-}friendly(m))$
$\pi_2 : \exists s(she(s) \wedge sp\_ref(x, v(\vec{p}_x)) \wedge Identity(s,x))$
$\pi_3 : said(e_0, s, \pi_1)$

(5.9) $\pi_1 : \exists m(material(m) \wedge environmentally\text{-}friendly(m))$
$\pi_2 : \exists s(she(s) \wedge said(e,s,u) \wedge u = \pi_1 \wedge sp\_ref(e', v(\vec{p}_s))$
$\wedge MetaphoricalCounterpart(e', e))$

### 5.1.3 Summary

The distinct interpretations introduced above provide yet more evidence that the existing formal models of gesture of the type discussed in Chapter 3 [Cassell, Stone, and Yan, 2000; Kühnlein, Nimke, and Stegmann, 2002] are too restrictive with respect to the range of possible gesture interpretations and the various way gesture can relate with

speech. In particular, while all previous models assume that speech and gesture introduce identical referents, our model captures the fact that the denotations of speech and gesture are often not identical. This analysis is supported by a higher level of abstraction expressed via underspecification mechanisms (see Section 5.2.2) which neither undergenerates nor overgenerates what the multimodal action can mean in context.

In Section 5.3, we propose grammar construction rules which comply with the constraints on discourse structure via distinct choices of alignment, and which also produce (i) underspecified logical formulae supporting the gesture interpretations in context and (ii) an underspecified semantic relation between the speech content and the gesture content which is resolvable to a specific value (e.g., Depiction, Elaboration for depicting gestures and Identity, MetaphoricalCounterpart for deictic gestures) at the semantics/pragmatics interface. Modelling the inference from underspecified to fully resolved pragmatic logical formulae is extraneous to our aims: this happens at the semantics/pragmatics interface via complex reasoning and world knowledge. What our grammar produces is underspecified logical formulae that are fully compatible with such pragmatic reasoning.

## 5.2 Mapping Gesture Form to Gesture Meaning

### 5.2.1 Modelling Form

Contrary to the decompositional analysis of lexical items and the semantic compositional approach to natural language, the meaning of a gesture cannot be determined compositionally.[4] Rather, the meanings of the gesture parts such as hand shape and hand movement are determined in a top-down direction by first establishing the meaning of the whole. This "global" property of gestures [McNeill, 2005] contrasts with the bottom-up approach of linguistic phrases whose meaning is a function of the meanings of its parts. To illustrate this property of gestures, consider again utterance (4.3) repeated in (5.10). The fact that the cup-like hand shape denotes a cylindrical object such as a drainpipe, and that the upward orientation of the fingers denotes an upward direction is derived from the contextually-specific meaning of the gesture as a whole—an upward movement through an enclosed area. In distinct speech contexts, many more

[4] Recall that we use 'gesture' to refer to the gesture phase carrying meaning, the stroke. It is the stroke that cannot be determined compositionally from the meanings of its parts. In contrast, recall from the previous formal approaches to gesture, in particular Fricke [2008], as discussed in Section 3.5.2, that the entire gesture unit is subject to hierarchical decomposition.

interpretations can arise from this hand shape and finger orientation.

(5.10) and he goes up [$_{X*}$through] the drainpipe

*Right hand is extended forward, palm facing up, fingers are bent in an upward direction. The hand shape resembles a cup.*

A further difference between linguistic signs and gestural signs pertains to the way the signifier, i.e., the form or the shape of the sign, is related with the signifier, the referent. The link between the form of spoken words and their referents is normally arbitrary [Saussure, 1916]; for instance, the fact that the sequence of letters "b-o-o-k" in this particular order signify a set of printed pages bound together containing a story, a novel, etc. results from a convention that is shared among the speakers of a particular language. On the other hand, the link between the form of the gesture sign and its referent is non-arbitrary and is based on iconicity and/or indexicality: in Peirce's terms, the choice of the speaker of (5.10) to use the cup-like hand shape to denote the upward direction through the drainpipe is based on perceptual resemblance between the gesture form and its denotation, i.e., they are constrained by *iconicity*. With deictic gestures, this relation is *indexicality* — deixis indicates the spatio-temporal coordinates of the referent in the physical space.

These properties of gesture thus necessitate a flat description language for the form-meaning mapping that is distinct from the hierarchical description of linguistic phrases. In Chapter 1 we stated that we capture the contribution of each aspect of gesture form by Typed Feature Structures (TFSs) where each aspect of the physical shape of the hand introduces a feature-value pair. Representing gesture form in TFSs is not a novel approach: for instance, Kopp, Tepper, and Cassell [2004] used 'image-description features' to represent the basic form features of the gestural morphology such as hand shape, location, orientation (recall the discussion in Section 3.4.2); Lascarides and Stone [2009b] described gesture form by means of feature structures which contain a list of the attribute-value pairs pertaining to the physical shape of the hand.

Following prior unification-based approaches to multimodal integration [Johnston, 1998a; Johnston, 1998b], we also use feature structures to represent the relative timings of the input modalities and to enforce constraints on them within the grammar. Finally, we formally represent the input speech signals by feature structures containing not only syntax-semantics information (as it is standardly done in formal grammars) but also their prosodic properties [Klein, 2000a]. Our work hence synthesises all previous analyses based on feature structures: i.e., we analyse the form of the gesture, the form

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

Figure 37: TFS representation of the depicting gesture in (5.1)

of the speech signal and the rules for their combination via typed feature structures.

We represent each aspect of gesture form with a feature-value pair. Following earlier research [McNeill, 1992; Kopp, Tepper, and Cassell, 2004; Bressem, 2008], we consider that the shape of the hand, the orientation of the palm and fingers, the hand location and the hand movement are the distinct aspects of form that potentially have semantic effects. In Section 2.2.1.2 we also stated that, outside of context, gestures often exhibit 'syntactic' ambiguity with respect to the gestural dimension—depicting or deictic. This, however, has effects on the form-meaning mapping in that deictic gestures introduce spatial coordinates in the ULF while purely depicting gestures do not. We capture this distinction by typing the gestural feature structure representations.

To illustrate the symbolic representation, consider Figure 37 which shows the TFS of the depicting gesture in utterance (5.1).[5] The gesture was annotated as literally depicting "mixing drywall mud" [Loehr, 2004]. We therefore type its feature structure representation as *depict-literal*, which is a subclass of *depicting*. The typing system is used for distinguishing the qualitative form features that capture the form of depicting gestures from the quantitative features required by deictic gestures. This distinction is necessary as it allows us to compose the logical form appropriate for the gesture itself, with depicting gestures consisting of qualitative semantic predications and deictic gestures including spatial reference [Lascarides and Stone, 2009b] (in Section 5.2.2 we provide motivation and further details).

We record the form features for deictic gestures as well because the shape of the hand determines the region of space that is designated by the hand: for instance, 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 region is rather a plane. Furthermore, the de-

[5] For the time being, we use as simple values as possible. We shall refine those values in Chapter 6.3 where we introduce the hierarchical organisation of gesture types.

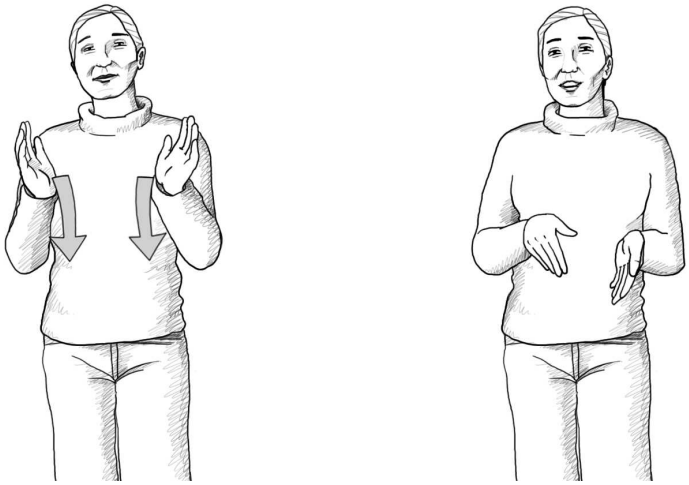


Figure 38: Abstract Deictic Gesture Placing a Hallway on the Virtual Map, example (5.11)

scriptive literature offers evidence that the form of the pointing hand is significant for interpreting its meaning in context: for instance, whereas a hand shape of an extended index finger has the abstract idea of singling out an object, an open hand with a vertical palm refers to a class of objects, rather than to an individuated object [Kendon, 2004]. The features appropriate for deictic gestures include the hand shape, the orientation of the palm and fingers, the movement of the hand, as well as the location of the tip of the index finger at the spatio-temporal coordinates $\vec{c}$ (recall Section 2.2.1.3). With this in mind, let us consider the deictic gesture in utterance (5.11), Figure 38 and the TFS representation of its form in Figure 39.

(5.11) There's like a [$_{NN}$little] [$_{N}$hallway]

*Hands are loosely open, vertical, parallel to each other. The speaker moves them downwards.*

The gesture type *deictic-abstract*, a subclass of *deictic*, sets it apart from *depicting* and hence it determines that the gesture is of quantitative values: namely that $\vec{c}$, along with the deixis form features maps to a meaning representation that designates the spatial coordinate $\vec{p}$. Also, recall from Chapter 2 that deictic gestures often designate individuals or events that are not present at the physical space identified by the pointing hand. We therefore type the gesture as *deictic-abstract* (in contrast to *deictic-concrete*)

$$\begin{bmatrix} \textit{deictic-abstract} & \\ \text{HAND-SHAPE} & \text{flat} \\ \text{PALM-ORIENT} & \text{towards-centre} \\ \text{FINGER-ORIENT} & \text{away-centre} \\ \text{HAND-MOVEMENT} & \text{up-down} \\ \text{HAND-LOCATION} & \vec{c} \end{bmatrix}$$

Figure 39: TFS representation of the deictic gesture in (5.11)

to record the fact that the object demarcated by the pointing—i.e., the hallway—is not physically available. This information is essential for resolving the function *v* that maps the gesture space to the actual space in denotation: with concrete deictic gestures which identify an individual that is salient at the three-dimensional space demarcated by the gesture *v* resolves to equality; that is, the individual identified by the gesture is equal to what the gesture denotes. On the other hand, with abstract deixis or nomination deixis which do not involve the physical presence of the referent *v* is not equality; that is, the individual/object identified in the gesture space (in (5.11), for instance, a virtual placement for a hallway) is not equal to its denotation (in (5.11), the real referent for "hallway"). We need this distinction for constraining the speech-gesture alignment appropriately (recall our findings in Section 4.3 and see the constraints in Section 5.3).

## 5.2.2 Modelling Meaning

### 5.2.2.1 Semantic Underspecification: RMRS

This thesis is centered around the observation that gesture form is massively ambiguous and hence the form-meaning mappings are not one-to-one but rather one-to-many. In Chapter 2, we introduced empirical data to illustrate various aspects of gestural ambiguity, including the gestural denotations mapped from form, the corresponding semantic relations between speech and gesture, and the various gesture "attachments" to the speech phrase. We demonstrated that the ambiguity persists in the context-of-use; for instance, a gesture disambiguated for its category—depicting or deictic—often remains imprecise regarding its specific meaning. A well-established method for handling cases where form is unspecific to derive meaning is semantic underspecification (see discussion in Section 2.2.1.1). All frameworks for underspecified meaning representation—e.g., Quasi-Logical Form [Alshawi, 1992], Underspecified Discourse

Representation Theory [Reyle, 1993], the Constraint Language for Lambda Structures [Egg, Koller, and Niehren, 2001], Hole Semantics [Bos, 2004], Minimal Recursion Semantics [Copestake et al., 2005], Regular Tree Grammars [Koller, Regneri, and Thater, 2008]—construct an abstract representation of what the discourse unit might mean in context instead of constructing logical forms enumerating all possible readings. This is standardly accomplished by *partially* describing the form of a fully-specific logical form, which in turn represents a context-specific and fully resolved interpretation. The underspecification frameworks are based upon the observation that the contextually specific meaning representations share the same parts, and so an underspecified logical form typically omits certain information that is an essential feature of a specific interpretation, though it does include parts that are common to all readings. Generally, underspecified semantics is designed to capture how semantic ambiguities can persist even when syntactic ambiguities are resolved (e.g., anaphoric and semantic scope ambiguities).

Following previous research on the formal semantics of gesture [Lascarides and Stone, 2006; Lascarides and Stone, 2009b], we use the underspecification formalism of Robust Minimal Recursion Semantics (RMRS [Copestake, 2007]) to describe the form-meaning mapping of embodied actions. RMRS—a highly factorised version of Minimal Recursion Semantics (MRS [Copestake et al., 2005])—uses maximally unary predicates and it links the arguments to their predicates via unique anchors (to be explained shortly). Our choice is motivated in the fact that RMRS is fully flexible in the type of semantic underspecification it supports: we can leave the predicate's arity and the type of the arguments underspecified until resolved by the discourse context. This is particularly useful since a semantic predicate such as

$$hand\_movement\_circular(i)$$

mapped from the form of the depicting gesture in Figure 37[6] can resolve to a wide range of fully specific predications in context, and these predications are not necessarily of unique arity. For instance, in Section 5.1 we claimed that one of the interpretations of the circular hand movement in (5.1) was the mud being mixed. This is achieved by resolving $hand\_movement\_circular(i)$ to the one-place predicate $substance(x')$, featured in (5.2). In one of the alternative interpretations where the hand movement is a depiction of the mixing event from the agent's viewpoint, the underspecified predicate $hand\_movement\_circular(i)$ resolves to the three-place predicate $rotate(e', h', x')$,

[6] In Section 5.2.2.4, we detail how this underspecified predicate is mapped from gesture form.

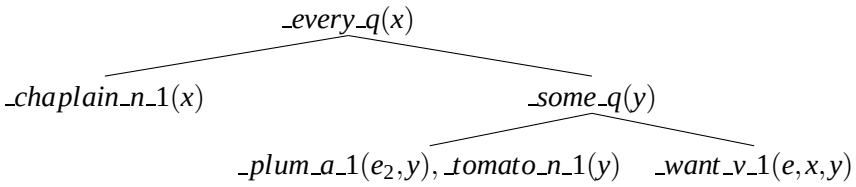
featured in (5.4).

#### 5.2.2.2 Representing information in RMRS

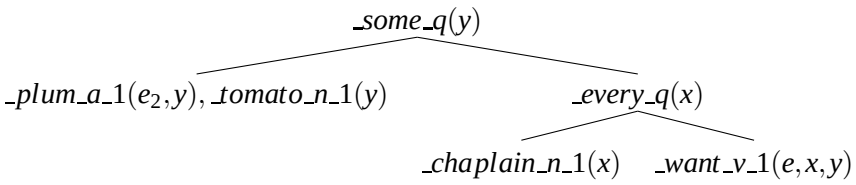
To illustrate the RMRS framework, consider the quantifier scope ambiguity featured in utterance (5.12). The wide scopal reading of the *every* quantifier is displayed in (5.13) and the wide scopal reading of the *some* quantifier is shown in (5.14). Our notation is based on the English Resource Grammar (ERG [Copestake and Flickinger, 2000])[7] where semantic predications corresponding to lexical items are encoded with a leading under-score, a part-of-speech tag, and a sense number (for instance, $\_plum\_a\_1$ means that this is a predicate mapped from a lexical item whose part of speech is adjective of sense 1).

(5.12) Every chaplain wants some plum tomato.

(5.13)




(5.14)




Underspecifying semantics in (R)MRS consists in describing the parts of the possible readings within a flat list of Elementary Predications (EPs) and specifying scopal constraints on them; for instance, the underspecified MRS representation of (5.12) is displayed in (5.15) and its corresponding RMRS variant is displayed in (5.16).

(5.15) $h_0$

$l_1 : \_every\_q(x, h_3, h_1)$

$l_{11} : \_chaplain\_n\_1(x)$

$l_2 : \_some\_q(y, h_4, h_2)$

$l_{21} : \_tomato\_n\_1(y)$

$l_{21} : \_plum\_a\_1(e_2, y)$

[7] http://www.delph-in.net/erg/

$l_3 : \_want\_v\_1(e,x,y)$
$h_3 =_q l_{11}$ $h_4 =_q l_{21}$

(5.16) $h_0$
$l_1 : a_1 : \_every\_q(x_0)\ l_1 : a_1 : RSTR(h_3)\ l_1 : a_1 : BODY(h_1)$
$l_{11} : a_{11} : \_chaplain\_n\_1(x_1)$
$l_2 : a_2 : \_some\_q(y_0)\ l_2 : a_2 : RSTR(h_4)\ l_2 : a_2 : BODY(h_2)$
$l_{211} : a_{211} : \_tomato\_n\_1(y_1)$
$l_{212} : a_{212} : \_plum\_a\_1(e_2)\ l_{212} : a_{212} : ARG1(y_2)$
$l_3 : a_3 : \_want\_v\_1(e_1)\ l_3 : a_3 : ARG1(x_2)\ l_3 : a_3 : ARG2(y_3)$
$h_3 =_q l_{11}$ $h_4 =_q l_{211}$
$x_0 = x_1 = x_2$
$y_0 = y_1 = y_2 = y_3$
$l_{211} = l_{212}$

An MRS elementary predication is associated with a (not necessarily unique) *label* $(l_1 \ldots l_n)$, which is used for resolving scopal positions in the final logical form.[8] Since intersective modification has no effects on scope, intersective modifiers share labels with the modified element; e.g.,

$$\boldsymbol{l_{21}} : \_tomato\_n\_1(y)\ \ \boldsymbol{l_{21}} : \_plum\_a\_1(e_2,y)$$

A more factorised way of representing EPs into unary predicates is achieved by RMRS: an MRS predication such as

$$l_3 : \_want\_v\_1(e,x,y)$$

is re-written in an RMRS notation as

$$l_3 : a_3 : \_want\_v\_1(e)\ l_3 : a_3 : ARG1(x)\ l_3 : a_3 : ARG2(y)$$

where $a_3$ is a unique *anchor*, which serves as a locus for adding arguments to the main predicate so that in case of equated labels, an argument can be uniquely associated with its predication. Thus if syntax simply yields the information $l_3 : a_3 : want(e)$, it is possible, in the absence of further constraints, for the fully specific predication that this resolves to correspond to a predicate with one, two, three or more arguments. This feature is useful for building semantic components for shallow language processors

[8] Labels have the same status as handles in Copestake et al. [2005].

where information about lexical subcategorisation is missing. Further, *holes ($h_i$)* are used to represent scopal arguments whose value is not fully determined by syntax. The admissible pluggings are constrained by *scopal conditions* or "equality modulo quantifiers" (notated as $=_q$ or *qeq*) between holes and labels ($h =_q l$ means that the predication labelled with $l$ is either in the scopal position $h$ or it is outscoped by the scopal position $h$ with only quantifiers intervening between them). Within RMRS, a lexical item introduces an elementary predication of unique variable names into the ULF. *Equations (=)* are used to express unifiability between variables. For instance, if in MRS one uses the same variables to refer to an argument as in

$$l_1 : \_every\_q(\boldsymbol{x}, h_3, h_1)$$

$$l_{11} : \_chaplain\_n\_1(\boldsymbol{x}),$$

in RMRS the same information is conveyed by equality between variables; e.g.,

$$l_1 : a_1 : \_every\_q(\boldsymbol{x_0})\ l_1 : a_1 : RSTR(h_3)\ l_1 : a_1 : BODY(h_1)$$

$$l_{11} : a_{11} : \_chaplain\_n\_1(\boldsymbol{x_1})$$

$$\boldsymbol{x_0 = x_1}.$$

Finally, a global top label $h_0$ is added to the whole formula.

Resolving an underspecified formula involves identifying holes with labels so that the result respects the $=_q$ constraints and leaves no variable that is bound by a quantifier free. For instance, the reading where *_every_q* takes wide scope (as displayed in (5.13)) is obtained from (5.16) as follows:

$$h_0 = l_1\ \ h_1 = l_2\ \ h_3 = l_{11}\ \ h_4 = l_{211}\ \ h_2 = l_3$$

and the reading where *_some_q* takes wide scope (as displayed in (5.14)) is obtained from (5.16) as follows:

$$h_0 = l_2\ \ h_2 = l_1\ \ h_3 = l_{11}\ \ h_4 = l_{211}\ \ h_1 = l_3.$$

No other reading is satisfied by (5.16). The benefit of the (R)MRS representation is that instead of enumerating all distinct readings, we have one flat formula of elementary predications, whose scopal arguments are left underspecified. The logical form is partial in that it contains the parts shared by the two readings and the information about scope is missing.

#### 5.2.2.3 Composing RMRS semantics

Semantic composition with RMRS [Copestake, Lascarides, and Flickinger, 2001] overcomes the shortcomings of λ-calculus in that the composition is *constrained*: a functor cannot arbitrarily pick arguments that are embedded in the logical form. Further, the RMRS semantic composition is *monotonic*, ensuring that the semantics of the daughters are always subsumed by the semantics of the mother [Shieber, 1986]. This is achieved by the append operation on the daughters. Another advantage of RMRS is that it produces directly off the syntax tree a flat description of the possible readings without having to access the distinct readings themselves. This property is particularly useful for composing gestural meaning since even through discourse processing the semantic predications yielded by gestural form may remain unresolved (for instance, recall examples (5.1) and (5.7) and the distinct interpretations of the gestures in context).

Composing an RMRS follows the semantic algebra of Copestake, Lascarides, and Flickinger [2001]. In particular, for each phrase we construct a semantic entity (*sement*) of the following six parts:

1. *Top:* a global label containing the whole formula. There is no label that outscopes Top; for instance, the Top label in (5.15) is $h_0$. During composition, the Top labels of the daughters are equated with that of the mother to demonstrate the derivation of a single logical form.

2. *Hook:* a placeholder for missing information similar to a λ-abstracted term. It is made of three parts:

   (a) *local top* or *ltop*: for a certain ULF, the ltop is identified with the label of the main predication; for instance, the ltop of (5.15) is $l_3$ and of (5.16)—$l_3, a_3$. The ltop of a quantifier is always distinct from the label of its predication since quantifiers 'float' between a hole argument and a labelled argument.

   (b) *semantic index ($i_1, i_2 \ldots i_n$)*: a variable that indicates what the LF is about and it has two subtypes: events ($e_1, e_2 \ldots e_n$) and individuals ($x_1, x_2 \ldots x_n$). For instance, the semantic index in (5.15) corresponds to the main variable of $l_3 : \_want\_v\_1(\boldsymbol{e}, x, y)$—the event $e$.

   (c) *external argument* or *xarg*: for verbs, the external argument corresponds to the subject position. It is used with control verbs such as "attempt", "fail", "try" to bind the subject in the embedded clause to that in the matrix clause as in "John tried to escape" (in that case, the external argument of

the predicate introduced by "escape" would be identified with the subject argument of the predicate introduced by "try"). Since we do not expect that gestures are used in control constructions, we forgo any further details about external arguments.

3. *Slots:* resources that need to be consumed so that a functor becomes semantically saturated and these include specifiers (SPEC), subjects (SUBJ) and complements (COMP). Some predications such as verbs have SUBJ and COMP slots, whereas others such as nouns are semantically saturated and thus have empty slots.[9] These can be viewed as corresponding to λ-abstracted terms: for instance, the λ-terms in a formula such as

$$\lambda x \lambda y \lambda z.give'xyz$$

would be expressed as

$$\{[l_1,x]_{subj}, [l_1,y]_{comp1}, [l_1,z]_{comp2}\}$$

where $l_1$ is the label of the elementary predication

$$l_1 : give(e,x,y,z)$$

To reflect the fact that depicting gestures express entire propositions and so they are roughly semantically equivalent to clauses, we do not ascribe slots to depicting gestures. Likewise, deictic gestures have no slots when used as references to individuals (just as noun phrases have no slots), and also when metaphorically used to point at individuals to offer them the floor, to cite the contribution of their interlocutor, etc (in this case, they can be viewed in analogy to clauses).

4. *Relations* or *Rels*: a bag of elementary predications.

5. *Hole conditions* or *Hcons* (represented as "$=_q$" or "*qeq*"): scopal conditions indicating the admissible pluggings of a labelled subformula $l$ into a hole $h$.

6. *Equations* or *Eqs* (represented as "="): equations between slot and hook variables. Note that some equations are not allowed: for instance, while $x_1 = x_2$, $x_1 \neq e_1$.

[9] Loosely speaking, the slots can be viewed as the required resources located either on the right of the forward slash or on the left of the backward slash in Combinatory Categorial Grammar.

To sum up, an RMRS sement is:

$$\langle Top\,[hook\,[ltop, index, xarg]]\,\{slots\}\{rels\}\,[hcons]\,\{eqs\}\rangle\,.$$

Since we assume that gestures do not take external arguments and slots, an RMRS sement corresponding to gesture is made of the following parts:

$$\langle Top\,[hook\,[ltop, index]]\,\{rels\}\,[hcons]\,\{eqs\}\rangle\,.$$

Semantic composition of the sement of the mother $sement_m$ results from the following binary operations over the sements of the two daughters $sement_{d1}, sement_{d2}$ [Copestake, Lascarides, and Flickinger, 2001]:

- $Top(sement_m) = Top(sement_{d1}) = Top(sement_{d2})$
- $Hook(sement_m) = Hook(sement_{d1})$ where $d1$ is the semantic head daughter[10]
- $Hook(sement_{d2}) = Slot(sement_{d1})$, adding them to $Eqs(sement_m)$
- $Slots(sement_m) = Slots(sement_{d2}) \cup Slots(sement_{d1}) \setminus (Hook(sement_{d2}) = Slot(sement_{d1}))$
- $Rels(sement_m) = Rels(sement_{d1}) \oplus Rels(sement_{d2})$
- $Hcons(sement_m) = Hcons(sement_{d1}) \oplus Hcons(sement_{d2})$
- $Eqs(sement_m) = TransitiveClosure(Eqs(sement_{d1}) \cup Eqs(sement_{d2}))$ where transitive closure over the equations means that in a set $S$ of $x, y, z$, if $x = y$ and $y = z$ then $x = z$

We will illustrate the algebra by computing the meaning representation for "every chaplain". We begin by filling in the *Top, hook, slots, rels, hcons* and *eqs* for the lexical items as demonstrated in (5.17) and (5.18). In sement (5.17), the ltop is co-indexed with the label and the anchor of $l_1 : a_1 : \_chaplain\_n\_1(x_1)$, and the semantic index is co-indexed with the variable $x_1$ bound by the main predication. In sement (5.18), the ltop of the quantifier is distinct from the label of the predication to ensure that the quantifier can be outscoped during composition. Also, the argument of the restrictor (RSTR) $h_1$ is within scopal constraints with the label $l_3$ of the SPEC slot: in this way the argument bound by the restrictor will always be within its scope via the syntax described above.

[10] In Head-driven Phrase Structure Grammar, the semantic head daughter is identified with the adjunct daughter in a head-adjunct phrase, and with the syntactic head daughter in any other headed phrase.

(5.17) *chaplain:*

$$\langle h_0\,[l_1, a_1, x_1]\,\{\}$$
$$\{l_1 : a_1 : \_chaplain\_n\_1(x_1)\}\,[\,]\,\{\}\rangle$$

(5.18) *every:*

$$\langle h_3\,[l_4, a_3, x_0]\,\{[l_3, a_2, x_0]_{spec}\}$$
$$\{l_2 : a_2 : \_every\_q(x_0)\ l_2 : a_2 : RSTR(h_1)\ l_2 : a_2 : BODY(h_2)\}$$
$$\left[h_1 =_q l_3\right]\,\{\}\rangle$$

The composition of the mother sement, displayed in (5.19), follows the operations outlined above: the *Top* labels of the daughters are identified ($h_0 = h_3$) with that of the mother; the SPEC slot for "every" is filled by "chaplain" (hence the empty *slots* in the composed phrase) and the variable equations are recorded within *eqs*; finally the *rels* and the *hcons* of the phrase are obtained by appending the *rels* and *hcons* of "every" to those of "chaplain". Since this is a flat list, the order of the append operation is not important.

(5.19) *every chaplain:*

$$\langle h_0\,[l_4, a_3, x_0]\,\{\}$$
$$\{l_1 : a_1 : \_chaplain\_n\_1(x_1)$$
$$\quad l_2 : a_2 : \_every\_q(x_0)\ l_2 : a_2 : RSTR(h_1)\ l_2 : a_2 : BODY(h_2)\}$$
$$\left[h_1 =_q l_3\right]$$
$$\{h_0 = h_3\ l_3 = l_1\ x_0 = x_1\}\rangle$$

#### 5.2.2.4 Hands-on Mapping of Gesture Form to Underspecified Gesture Meaning

With this machinery in hand, we can proceed with producing underspecified meaning representations from gestural form. As previously discussed (see Sections 2.2.1.2 and 5.2.1), there is a difference between the way depicting gestures denote the referent vs. the way deictic gestures denote the referent. The form features of depicting gestures visualise the qualitative characteristics of the referent, which means that every aspect of the hand form feature (its location, movement, etc) pertains to its meaning in the speech context. The connection is usually iconicity between gestural form and its denotation *in the context of speech*. 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. Through deictic gestures, people anchor their speech signals to the context of the communicative event thereby making the content

of their propositions a function that maps a world in its contextually-specific time and space to truth values. The connection here is indexicality between the deictic form and its denotation in *the context of speech, including the time and space in which the communicative action takes place*.

For depicting gestures, mapping form to meaning involves reading the gestural predications directly off the feature structure, associating them with the corresponding labels, anchors and arguments and identifying scopal constraints wherever necessary [Lascarides and Stone, 2009b]. For instance, the form representation in Figure 37 maps to the underspecified semantic representation in (5.20).

(5.20) $l_0 : a_0 : [\mathcal{G}](h)$

$l_1 : a_1 : \_hand\_shape\_bent(i_1)$

$l_2 : a_2 : \_palm\_orient\_towards\_down(i_2)$

$l_3 : a_3 : \_finger\_orient\_towards\_down(i_3)$

$l_4 : a_4 : \_hand\_location\_lower\_periphery(i_4)$

$l_5 : a_5 : \_hand\_movement\_circular(i_5)$

$h =_q l_n$ *where* $1 \leq n \leq 5$

Recall from Sections 1.3.1 and 5.1 that we assume that there are constraints on the accessibility of the discourse parts that can function as antecedents for resolving the semantic values of the discourse parts. To account for these constraints, Lascarides and Stone [2009b] introduce the scopal $[\mathcal{G}]$ operator which limits the scope of the predicates within the gestural modality, expressed via the scopal condition $h =_q l_n$. This captures constraints on co-reference between speech and depicting gesture so that an individual introduced by a depicting gesture cannot be subsequently co-referred to in speech by "it".

Further, the semantic description is consistent with the principles outlined in Section 5.2.2.1: an EP receives a label ($l_0 \ldots l_5$), an anchor ($a_0 \ldots a_5$) and an underspecified variable argument ($i_1 \ldots i_5$). The gestural EPs underspecify the referent *i* depicted through each predication, and so the variable *i* can resolve to an individual *x* or to an event *e*. This is necessary, as gesture often underspecifies its main argument: for instance, recall from Section 5.1 that the different syntactic attachments can resolve a gestural predication to a fully specific logical form denoting an individual or an event. This is also in line with the experimental study in Section 1.1.1 where the same gesture signal in the same context was interpreted as an event of stacking books by one annotator, and as the books being stacked by another annotator. The use of an underspecified

variable is essential for those cases where ambiguity persists even in the context of use.

The way an RMRS predicate resolves to a fully-specific logical form is a byproduct of discourse processing and so it happens outside the grammar [Lascarides and Stone, 2009b]. Following the approach of Copestake and Briscoe [1995] of constructing a specialised predicate out of a polysemous lexical entry via a type hierarchy of increasingly specific predications, Lascarides and Stone [2009b] propose to interpret the semantic predications contributed by gesture in a similar way. For instance, the underspecified predicates $\_hand\_shape\_bent(i_1)$ and $\_hand\_movement\_circular(i_5)$ featured in the RMRS semantics could resolve in context to $substance(x') \wedge rotate(e,x')$ (see (5.2)) where the fully specific predicates—$substance(x')$ and $rotate(e,x')$—are leaves in the type hierarchy of the gesture predicates licensed by iconicity (assuming resemblance between the depicting gesture's referent and the speech referent) [Lascarides and Stone, 2009b]. Informally, resolving $\_hand\_shape\_bent(i_1)$ to $substance(x')$ means that the bent hand shape resembles a container for some substance, and resolving $\_hand\_movement\_circular(i_5)$ to $rotate(e,x')$ designates that the mud was going round. Not all predications contribute meaning to the final interpretation: for instance, the downward orientation of the fingers in (5.1) has no effects on the interpretation in context. The type hierarchy essentially features the open-ended, but still constrained (by iconicity) ways in which a predicate can be interpreted [Lascarides and Stone, 2009b]: for instance, while a predicate such as $\_hand\_movement\_circular(i)$ can resolve literally to an object that is going round, to the event of rotating something, and even metaphorically to an iterative process, it can never resolve to something denoting a square concept.

We will now turn to the semantic representation of deictic gestures. For that purpose, consider the representation in (5.21) yielded from the deixis form features in Figure 39.

(5.21) $l_1 : a_1 : deictic\_q(i_0)\ l_1 : a_1 : RSTR(h_1)\ l_1 : a_1 : BODY(h_2)$
$l_{21} : a_2 : sp\_ref(i_1)\ l_{21} : a_2 : ARG1(v(\vec{p}))$
$l_{22} : a_3 : \_hand\_shape\_flat(e_0)\ l_{22} : a_3 : ARG1(i_2)$
$l_{23} : a_4 : \_palm\_orient\_towards\_centre(e_1)\ l_{23} : a_4 : ARG1(i_3)$
$l_{24} : a_5 : \_finger\_orient\_away\_centre(e_2)\ l_{24} : a_5 : ARG1(i_4)$
$l_{25} : a_6 : \_hand\_movement\_up\_down(e_3)\ l_{25} : a_6 : ARG1(i_5)$
$l_{26} : a_7 : \_hand\_location\_c(e_4)\ l_{26} : a_7 : ARG1(i_6)$
$h_1 =_q l_{21}$
$l_{21} = l_{22} = l_{23} = l_{24} = l_{25} = l_{26}$

$$i_0 = i_1 = i_2 = i_3 = i_4 = i_5 = i_6$$

The mapping of deixis form to underspecified meaning captures the fact that deixis provides the spatial reference of an individual or event in the physical space $\vec{p}$. This is formalised by the two-place predicate

$$l_{21} : a_2 : sp\_ref(i_1) \;\; l_{21} : a_2 : ARG1(v(\vec{p}))$$

whose first argument is the underspecified variable $i_1$, and the second argument—linked through the anchor $a_2$—is the actually denoted space $v(\vec{p})$ with $v$ being the function that maps the gesture space to the space in denotation (see earlier discussion in Section 2.2.1.3). The ULF is only a partial description of the resolved logical form: for instance, 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.

To demonstrate the resolution of the underspecified deixis predicates, let us consider again utterance (5.7). The logical form contributed by the gesture form would feature the predicate $sp\_ref(i)$ where $i$ underspecifies the main referent introduced by the deixis. The interpretations are dependent on the speech-gesture alignments; for instance, attaching the gesture to the NP "she" would support an interpretation where the spatial referent $i$ resolves to an individual $x$, which in turn would support an Identity relation between the speech denotation and the deixis denotation. This interpretation was accounted for by the logical form in (5.8). Alternatively, a gesture attachment to "she said" would resolve the deixis referent to the event $e$, which in turn would support a MetaphoricalCounterpart relation between the gesture content and the speech content, featured in the logical form in (5.9).

Contrary to the form features of depicting gestures, deixis form features are not a visual display of what the gesture could possibly mean in context. Instead, the way the hand is used in the pointing act has semantic effects on how the underspecified referent $i$, and hence the relation between speech and deixis resolves in context: whereas an extended index finger often means a real or virtual identity between the individual pointed at and the denoted space, an open hand supine often serves a pragmatic function such as offering the floor or citing someone else's contribution to the discourse. We therefore record the form of the pointing hand by mapping a deixis feature-value pair to a two-place predicate, e.g.,

$$l_2 : a_3 : \_hand\_shape\_flat(e_0) \; l_2 : a_3 : ARG1(i_2)$$

with the first argument being an event variable ($e_0 \ldots e_n$) and the second argument ARG1 being the referent identified by the pointing signal ($i_0 \ldots i_n$). This formalisation is similar to the treatment of intersective modification in ERG: 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 via the equations:

$$i_0 = i_1 = i_2 = i_3 = i_4 = i_5 = i_6$$

and whose label is equated with the label of the modified EP, obtained via the equations:

$$l_{21} = l_{22} = l_{23} = l_{24} = l_{25} = l_{26}$$

For consistency with ERG where individuals are all bound by quantifiers, we use the *deictic_q* quantifier to quantify over the spatial referent $i_1$.

Note also that for deixis we do not block co-reference from an individual introduced in deixis to an anaphoric element in speech, and so the referent identified by a pointing signal could be anaphorically related with speech elements. In formal terms, the $[\mathcal{G}]$ modality is not included in the logical form for deixis. To illustrate the co-reference across modalities, consider utterance (5.22), which is a slightly modified version of the original performance in (4.7), page 113. It seems perfectly acceptable for the gesture performed in the direction of the computer mouse in the first discourse unit to serve as an antecedent for the pronoun "it" introduced in the subsequent discourse unit. In this case, the co-reference is across the modalities since the referent of the deictic gesture resolves the pronoun coming from speech.

(5.22) [$_N$Thank] you. [$_{NN}$I'll] take it.

*Speaker's right hand is loosely closed, index finger is loosely extended, pointing at the computer mouse.*

## 5.3 Well-formedness Constraints

In this section, we propose grammar construction rules for aligning the form of the gesture and the form of the speech signal into a single syntax tree that, using the semantic algebra with RMRS, maps to a multimodal semantic representation. The well-formedness rules are driven from our empirical findings detailed in Chapter 4: on the one hand the choices of alignment are constrained by the prosodic prominence of the speech element that gesture (stroke) temporally overlaps with, and on the other we

remain loose about the syntactic category of the speech phrase that gesture attaches to. The well-formedness rules have been designed as independently as possible of a particular grammar formalism since we predict that any constraint-based formalism that interfaces structured phonology, syntax and semantics—e.g., Head-Driven Phrase Structure Grammar, Lexical Functional Grammar or Combinatory Categorial Grammar—is a suitable candidate for developing a formal grammar for multimodal language. The goal of this chapter is hence to enrich our grammatical knowledge with multimodal grammar rules.

### 5.3.1 Prosodic Word and Gesture Alignment

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

**Definition 5.3.1. Situated Prosodic Word Constraint.** *A depicting and/or deictic gesture can attach to a spoken word* $\boldsymbol{w}$ *of a spoken utterance if (a.) there is an overlap between the temporal performance of the gesture stroke and* $\boldsymbol{w}$*; (b.)* $\boldsymbol{w}$ *bears a nuclear or a pre-nuclear pitch accent.*

We use "and/or" to remain as neutral as possible about the gestural dimension, and thus to reflect the fact that certain gestures can inherit information from more than one dimension (recall (2.4), page 35 and the related discussion). We also assume that the temporal relation between the speech (S) and the gesture (G) modality is not of identity but it rather subsumes nine temporal configurations as displayed in Table 19 [Oviatt, DeAngeli, and Kuhn, 1997]. Finally, we do not restrict gesture to a particular syntactic category: as we saw in the empirical investigation, the gesture stroke can happen along a word of any syntactic category (recall Section 4.2.2).

Definition 5.3.1 can be reliably applied to both depicting and deictic dimensions. We will now provide an analysis for both gestural dimensions in turn.

#### 5.3.1.1 Situated Prosodic Word Constraint and Depicting Gesture

For depicting gestures, we tested Definition 5.3.1 on one observation of Loehr's [2004] corpus, and we established that this rule would produce derivation trees for 19 out of 25 gesture strokes. This rule would not produce an analysis for a stroke overlapping a pause. Since pauses and speech disfluencies are not within the purview of the grammar, we do not analyse strokes overlapping pauses and disfluencies. As for the remaining cases, they will be analysed in terms of the Situated Spoken Phrase Constraint (Section 5.3.2) which licenses the alignment of a stroke and a metrical tree. Furthermore,

| Temporal Relations between S and G | | |
|---|---|---|
| $(start(S) < start(G))$ | $\wedge$ | $(end(S) = end(G))$ |
| $(start(S) < start(G))$ | $\wedge$ | $(end(G) < end(S))$ |
| $(start(S) < start(G))$ | $\wedge$ | $(end(S) < end(G))$ |
| $(start(G) < start(S))$ | $\wedge$ | $(end(G) = end(S))$ |
| $(start(G) < start(S))$ | $\wedge$ | $(end(G) < end(S))$ |
| $(start(G) < start(S))$ | $\wedge$ | $(end(S) < end(G))$ |
| $(start(G) = start(S))$ | $\wedge$ | $(end(G) = end(S))$ |
| $(start(G) = start(S))$ | $\wedge$ | $(end(G) < end(S))$ |
| $(start(G) = start(S))$ | $\wedge$ | $(end(S) < end(G))$ |

Table 19: Possible Overlap Relations between the Timing of Gesture and the Timing of Speech [Oviatt, DeAngeli, and Kuhn, 1997]

whenever a gesture overlapped two nuclear prominent elements or a combination of a pre-nuclear and a nuclear accented element, Definition 5.3.1 yields a syntactic ambiguity as to which node in the tree the gesture attaches to. Similarly to "John saw the man with the telescope", we claim that all attachment analyses should be obtainable by the grammar. Whereas the propositional content of a ULF coming from linguistic phrases can be unambiguously recovered using discourse processing, the propositional content of a gestural phrase is almost impossible to reconstruct without using contextual knowledge. We therefore claim that the form-meaning mappings of gesture are one-to-many.

To illustrate how the Situated Prosodic Word Constraint works with depicting gestures, consider example (5.23), Figure 40 [Loehr, 2004]. This utterance is taken from a longer narrative where the speaker described her cupboards and the fact that they were crooked. The gesture annotation was produced by one annotator during our experiment detailed in Section 1.1.1.

(5.23) And I [$_N$couldn't] believe it [$_{PN}$one] of the [$_{NN}$cup]boards is [$_N$off].

*Both hands are in centre, around 50 cm apart, palms are open vertical, finger tips point upwards. Along with "one", both hands rotate to the right to horizontal orientation, with the right hand open up and the left hand over the right hand with open palm pointing downwards.*

(a) I couldn't believe it one…

(b) …of the cupboards…

(c) …is off

Figure 40: Gesture depicting the "off-ness" of cupboards [Loehr, 2004], example (5.23)

The Situated Prosodic Word Constraint in Definition 5.3.1 licenses the following attachments: (1). "couldn't" + depicting gesture, (2). "one" + depicting gesture and (3). "off" + depicting gesture. The attachment ambiguities are visualised in Figure 41. Both syntactic trees (2) and (3) would yield ULFs supporting the final interpretations in context: attaching the gesture to "one" would support an interpretation where the hands place an object in the frontal space which, by the hands shape is rectangular, and by the dynamic change is somewhat crooked (or turned upside-down). Then an attachment to "off" would rather support an interpretation where the hands depict a salient feature of

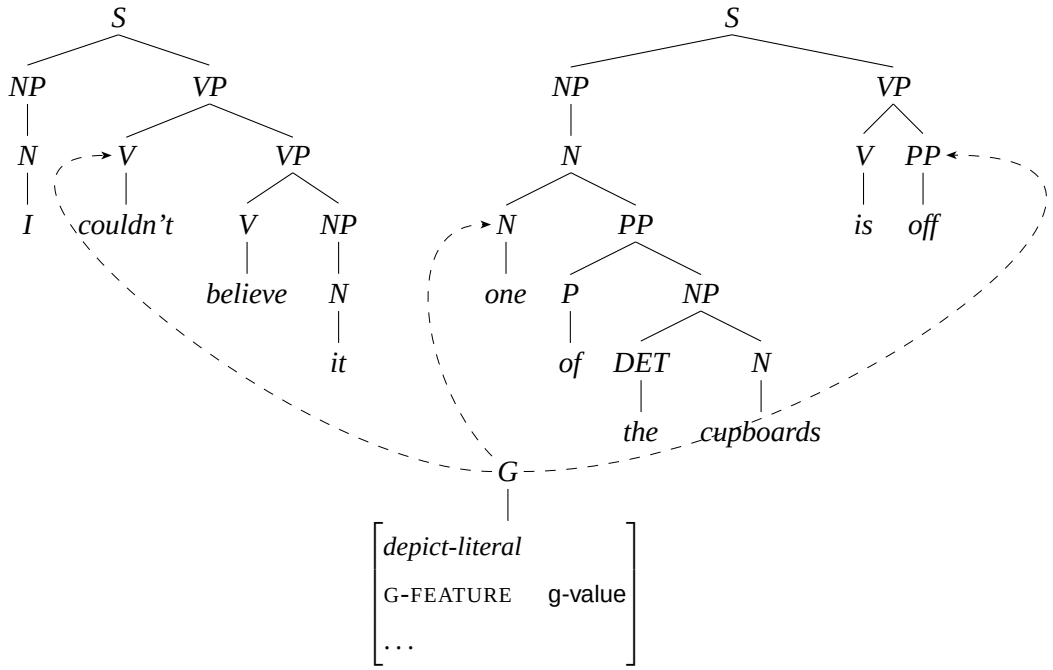



Figure 41: Depicting Gesture–Prosodic Word Attachment Ambiguities

the cupboard doors: namely, the fact that they were "off". Both analyses are supported by the dynamic change of the hand. However, the attachment to "couldn't" is solely based on the overlap relation between the timing of gesture relative to the timing of the nuclear accented element. We claim that although syntactically well-formed, this phrase cannot produce the intended meaning in the communicative act: the content conveyed by the hand shape and its movement cannot be semantically related with the negation.

An alternative analysis based on the symbolic representation of the second annotator from our experiment would not derive such ambiguities: the annotator treated this hand movement as two distinct strokes where the first one spans "one of the cupboards", and the second one spans "cupboards is off". Each stroke, in this case, will be unambiguously associated with "one" and "off", where both attachments support the contextual interpretations proposed above.

The derivation of the situated word proceeds by first representing gesture form in a feature structure of type *depict-literal* as shown in Figure 42. Since the values of the right hand and the left hand are not symmetrical (e.g., the right palm is oriented towards up, whereas the left palm is oriented towards down), we use separate features for the left hand and the right hand: for instance, a feature such as R-HAND-SHAPE designates

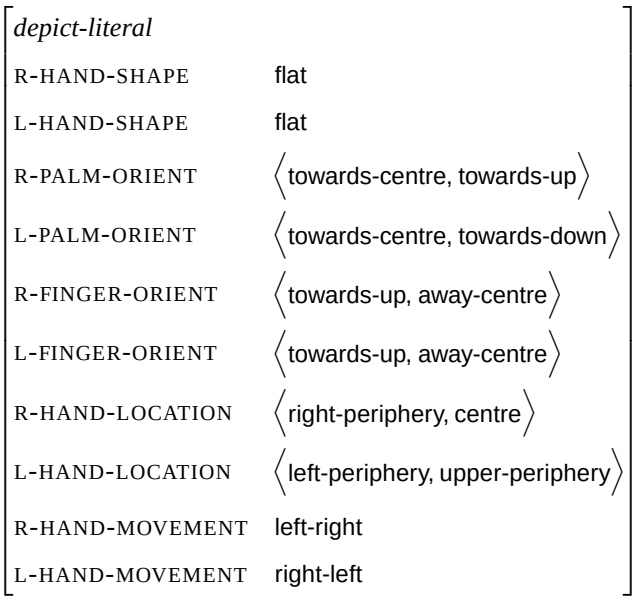



Figure 42: TFS representation of the depicting gesture in (5.23)

the shape of the right hand, and a feature such as L-HAND-SHAPE designates the shape of the left hand. This contrasts with the representation of symmetrical hands where there is no need to distinguish between the left hand and the right hand (e.g., Figure 39, page 131). Further, we account for the dynamics of the hand by recording the arising kinetic change by a separate value: for instance, R-PALM-ORIENT introduces both *towards-centre* and *towards-up,* which are subtypes of *orient* (see Section 6.3). The alternative way of encoding the change in hand via a single value that is multiply inherited from the appropriate supertypes (for instance, introducing a single type *towards-centre-towards-up* inherited from both *towards-centre* and *towards-up*) would be quite cumbersome as the possible combinations are in fact open-ended. We therefore choose to treat the values as a bag of equally ranked attributes that can be combined via conjunction, similarly to intersective modification in language. A limitation of this approach is that the information about the temporal unfolding of the values gets lost. In Figure 42, for instance, the orientation of the right palm changed from *towards-centre* to *towards-up*. The reverse change from *towards-up* to *towards-centre* would yield the same symbolic representation and a corresponding meaning representation. A possible solution to this problem would be to make each feature a complex feature structure that introduces TIME_START and TIME_END values. Whereas this information might effect the interpretation of the logical form, we believe that the kinetics of the

hand shape has no effect on the possible speech-and-gesture alignments. We therefore adopt as simple a formal representation as possible. In contrast, multiple inheritance is used to encode a static hand for which multiple values are appropriate: for instance, a value such as *12345f-bent* would designate a hand shape where all fingers (from the thumb to the little finger) were engaged in the gesture performance and their shape was bent (see Section 6.3, in particular Figure 65).

We now turn to the gesture form-meaning mapping. The gestural semantics is illustrated in (5.24).

(5.24)
$$\begin{aligned}
&\langle h_0\,[l_0, a_0, i_{1-10}]\,\{\}\{l_0 : a_0 : [\mathcal{G}](h)\\
&\qquad l_1 : a_1 : _r_hand_shape_flat(i_1)\\
&\qquad l_2 : a_2 : _l_hand_shape_flat(i_2)\\
&\qquad l_{31} : a_3 : _r_palm_orient_towards_centre(i_{31})\\
&\qquad l_{32} : a_4 : _r_palm_orient_towards_up(i_{32})\\
&\qquad l_{41} : a_5 : _l_palm_orient_towards_centre(i_{41})\\
&\qquad l_{42} : a_6 : _l_palm_orient_towards_down(i_{42})\\
&\qquad l_{51} : a_7 : _r_finger_orient_towards_up(i_{51})\\
&\qquad l_{52} : a_8 : _r_finger_orient_away_centre(i_{52})\\
&\qquad l_{61} : a_9 : _l_finger_orient_towards_up(i_{61})\\
&\qquad l_{62} : a_{10} : _l_finger_orient_away_centre(i_{62})\\
&\qquad l_{71} : a_{11} : _r_hand_location_right_periphery(i_{71})\\
&\qquad l_{72} : a_{12} : _r_hand_location_centre(i_{72})\\
&\qquad l_{81} : a_{13} : _l_hand_location_left_periphery(i_{81})\\
&\qquad l_{82} : a_{14} : _l_hand_location_upper_periphery(i_{82})\\
&\qquad l_9 : a_{15} : _r_hand_movement_left_right(i_9)\\
&\qquad l_{10} : a_{16} : _l_hand_movement_right_left(i_{10})\\
&[h =_q l_1, h =_q l_2, h =_q l_{31}, h =_q l_{41}, h =_q l_{51}, h =_q l_{61},\\
&\qquad h =_q l_{71}, h =_q l_{81}, h =_q l_9, h =_q l_{10}]\\
&\{l_{31} = l_{32}, l_{41} = l_{42}, l_{51} = l_{52}, l_{61} = l_{62}, l_{71} = l_{72}, l_{81} = l_{82}\}\rangle\\
&\{\}\rangle
\end{aligned}$$

Following the principles outlined in Section 5.2.2, the form representation of gesture maps to an RMRS sement composed of a Top label, a hook, a set of elementary predications, scopal constraints and equations, as follows:

1. Top: We assign $h_0$ as a Top label of the formula.

2. Hook: Recall that for gestures, the hook consists of an ltop and a semantic index. We assume that the label and the anchor $l_0, a_0$ of the $[\mathcal{G}]$ operator is the ltop since it outscopes all gestural predications. Further, the logical form is too underspecified to know which of the elementary 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 $i_{1-10}$ as a shorter notation for a disjunction of co-indexations to reflect the fact that the underspecified variable $i_1 \ldots i_{10}$ of each EP could potentially resolve to the main variable: event $e$ or individual $x$.

3. Rels: The gestural predications are directly read off the feature structure shown in Figure 42. Note also that the feature value pairs of multiple values introduce separate predications of equated labels (see 5. below).

4. Hcons: Every gestural predication is within the scope of the $[\mathcal{G}]$ operator:

$$h =_q l_1, h =_q l_2, h =_q l_{31}, h =_q l_{41}, h =_q l_{51}$$

$$h =_q l_{61}, h =_q l_{71}, h =_q l_{81}, h =_q l_9, h =_q l_{10}$$

   Here we omit the equated labels (see below).

5. Eqs: For consistency with ERG, predications that have no effects on scope have equated labels, and so we equate the labels of the predications mapped from features of multiple values:

$$l_{31} = l_{32}, l_{41} = l_{42}, l_{51} = l_{52}, l_{61} = l_{62}, l_{71} = l_{72}, l_{81} = l_{82}$$

Producing a situated word out of "off" and the depicting gesture as licensed by the rule in Definition 5.3.1 involves the following steps: 1. propagating the prosodic PHON and syntactic SYN information of the speech head daughter 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; 2. recording the timing of the situated utterance, that is, the entire interval of the speech-gesture duration from the lower TIME_START value to the higher TIME_END value. This information is necessary in case the situated word further aligns with a gesture; 3. producing a semantic representation SEM of the multimodal utterance which includes (a.) the underspecified semantics of the gesture

daughter as shown in (5.24), (b). the semantics of the speech daughter as shown in (5.25) and (c). an underspecified relation $vis\_rel(l_8, l_0)$ between the ltop $l_0$ of gesture and the ltop $l_8$ of speech. The derivation tree is displayed in Figure 43.

(5.25) $off{:}\langle h_1\,[l_8, a_{12}, e_2]\,\{$
$\{l_8 : a_{12} : \_off\_adv(e_2)\; l_8 : a_{12} : ARG1(u_1)\}[]\{\}\rangle$

For the sake of readability, we have used only two feature-value pairs and the corresponding elementary predications to represent gesture form and meaning. The rest of the depicting features are as shown in Figure 42, and the full set of elementary predications is displayed in (5.24). In composition, the global Top $h_1$ of the speech daughter (see (5.25)) is identified with the global Top $h_0$ of the gesture daughter (see (5.24)). Consistent with our programmatical approach to identify a semantic relation between the speech daughter and the aligned gesture daughter, which demonstrates that the speech signal is coherently related with the gesture signal, the construction rule therefore introduces in semantics an underspecified relation *vis_rel* (visualising relation) between the ltop $l_8$ of the speech signal and the ltop $l_0$ of the gesture signal. 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 detailed in Section 2.2.1.1, *vis_rel* is resolvable at the semantics/pragmatics interface to a specific value—for instance, Depiction, Elaboration, Narration—that is dependent on resolving the gestural denotation (recall from Section 2.2.1.1 that Lascarides and Stone [2009b] assume that resolving the semantic relation between speech and gesture, and interpreting gesture in context are logically co-dependent).

The semantics of the mother is obtained by appending the relations of the gesture daughter to the relations of the speech daughter, which in turn are appended to the *vis_rel* relation. The hole conditions are accumulated from the gesture daughter since the speech daughter's hcons are empty. The ltop of the mother is identified with the label of *vis_rel* contributed by the rule $(l_9, a_{13})$. Finally, the relation *vis_rel* introduces an M-ARG (multimodal argument) attribute which serves as a semantic index of the integrated multimodal signal and so it can be taken as an argument by any external predicate. For instance, in the utterance "He threw the ball" with a gesture attaching to "the ball", the verb "throw" would take two arguments: ARG1—corresponding to the subject argument—would be identified with ARG0 of "he", and ARG2—corresponding to the object argument—would be identified with M-ARG of the situated phrase "the

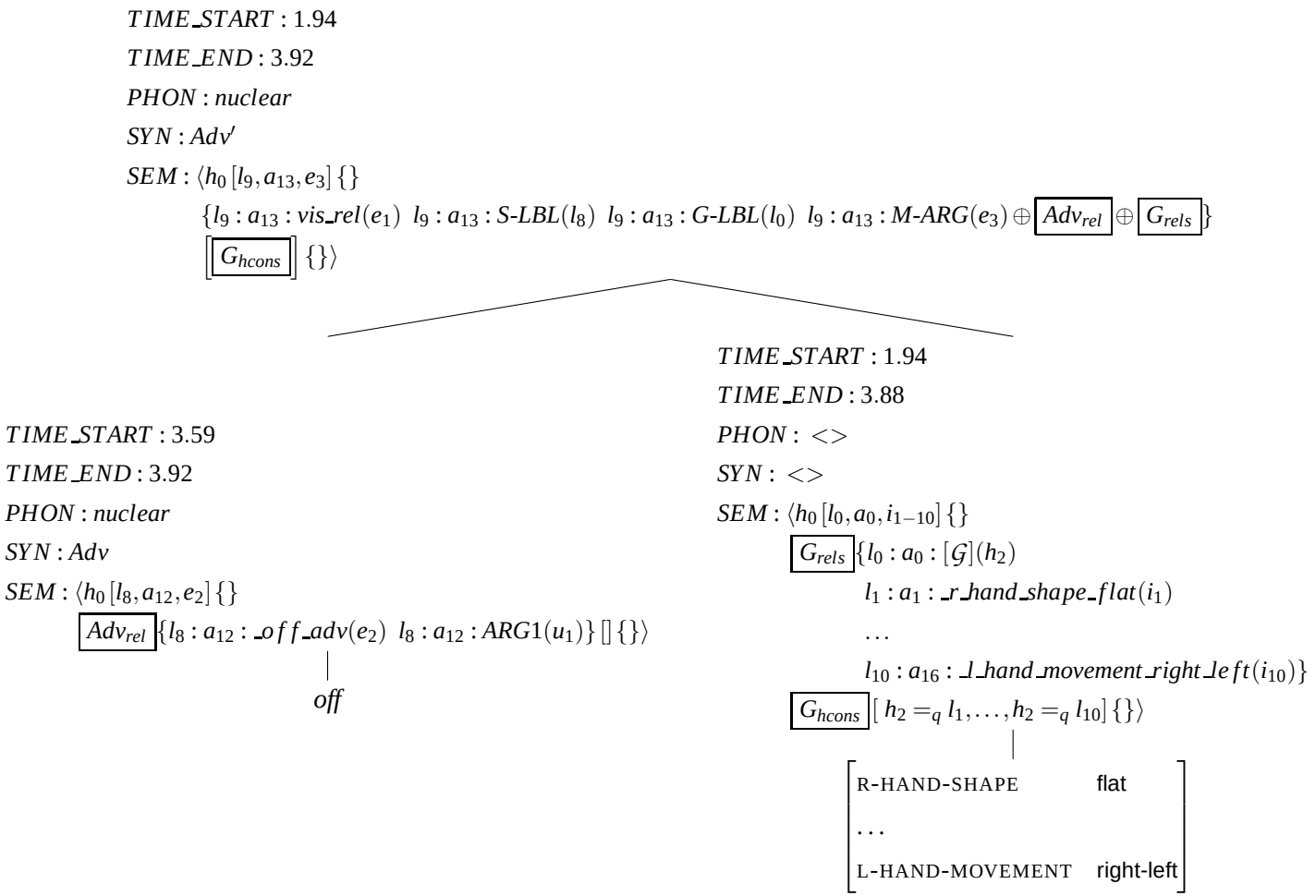



Figure 43: Derivation Tree for Depicting Gesture and the Adv “off”

ball" + gesture. 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 [Flickinger, Bender, and Oepen, 2003].

#### 5.3.1.2 Situated Prosodic Word Constraint and Deictic Gesture

We illustrate the syntactic derivation and the semantic composition for deixis and a spoken word using utterance (5.11). The Situated Prosodic Word Constraint in Definition 5.3.1 licenses an attachment of the deictic gesture to the nuclear prominent "hallway". We have already provided the mapping of the gesture form in Figure 39 to meaning (see (5.21)). We therefore start by extending the logical form in (5.21) with a Top label and a hook as displayed in (5.26).

.

(5.26) $\langle h_0\,[l_3, a_8, i_1]\,\{\}$
$\{l_1 : a_1 : deictic\_q(i_0)\ l_1 : a_1 : RSTR(h_1)\ l_1 : a_1 : BODY(h_2)$
$l_{21} : a_2 : sp\_ref(i_1)\ l_{21} : a_2 : ARG1(v(\vec{p}))$
$l_{22} : a_3 : \_hand\_shape\_flat(e_0)\ l_{22} : a_3 : ARG1(i_2)$
$l_{23} : a_4 : \_palm\_orient\_towards\_centre(e_1)\ l_{23} : a_4 : ARG1(i_3)$
$l_{24} : a_5 : \_finger\_orient\_away\_centre(e_2)\ l_{24} : a_5 : ARG1(i_4)$
$l_{25} : a_6 : \_hand\_movement\_up\_down(e_3)\ l_{25} : a_6 : ARG1(i_5)$
$l_{26} : a_7 : \_hand\_location\_c(e_4)\ l_{26} : a_7 : ARG1(i_6)\}$
$[h_1 =_q l_{21}]$
$\{l_{21} = l_{22} = l_{23} = l_{24} = l_{25} = l_{26}$
$i_0 = i_1 = i_2 = i_3 = i_4 = i_5 = i_6\}\rangle$

The compositional semantics of the deictic gesture is composed of the following components:

1. Top: Similarly to depicting gesture, we add $h_0$ as a Top label outscoping the whole formula.

2. Hook: Since this is a scopal phrase, the ltop of the phrase is a label distinct of the quantifier *deictic_q*. The semantic index is the underspecified variable $i_1$ bound by *sp_ref*.

3. Rels: A bag of elementary predications mapped from deixis feature-value pairs (recall (5.21) and the subsequent discussion about the difference between the

elementary predications contributed by depicting gesture and those contributed by deictic gesture).

4. Hcons: A scopal condition indicating that the referent introduced by the deictic gesture is bound by the quantifier *deictic_q*.

5. Eqs: Consistently with the treatment of intersective modification in ERG, we define equations between the labels of the elementary predications mapped from the deixis form features, and also between ARG1 of those EPs and the first argument of the main predicate *sp_ref* (for details, see the discussion following (5.21)).

The derivation tree is displayed in Figure 44. Like the derivation tree for depicting gesture, we used only two feature-value pairs and the associated elementary predications to represent deixis form and meaning. In semantic composition, the deixis relations are appended to the semantic relation of the speech daughter, the predicate $l_4 : a_4 : \_hallway\_n\_1(x_2)$. In so doing, the underspecified semantic index of the deixis unifies with the semantic index of the speech, and so the underspecified variable $i_1$ of $sp\_ref(i_1)$ resolves to $x_1$. Like depicting gestures, deictic gesture relates with the aligned speech through some (underspecified) relation that demonstrates that they are coherently related. The resolution of this relation is logically co-dependent with how the gesture is interpreted in context [Lascarides and Stone, 2009b]. The construction rule therefore introduces an underspecified relation $deictic\_rel(x_2, x_1)$ between the semantic index $x_2$ of the speech EP and the semantic index $x_1$ of the deixis EP. Recall from Section 5.1 that the pragmatic interpretation of the multimodal action involves reasoning about the semantic relation between speech and deixis. So a possible resolution of $deictic\_rel(x_2, x_1)$ in Figure 44 would be VirtualCounterpart: namely, the deictic gesture places a virtual counterpart of the hallway just in front of the speaker.[11] Similarly to the treatment of intersective 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, this can be paraphrased as "the hallway pointed at" where "pointed at" modifies "hallway" and hence ARG2 of the predication introduced by "point at" would be identified with the main argument of the predication introduced by "hallway". In this way, in any further composition *deictic_rel* would be outscoped

[11] Unlike (5.8), the referent identified by the gesture is not physically present and therefore inferring an Identity relation is not possible.

$TIME\_START$ : 25.5
$TIME\_END$ : 26.73
$PHON$ : *nuclear*
$SYN$ : $N'$
$SEM$ : $\langle h_0\,[l_4, a_9, x_2]\,\{\}$
$\{l_4 : a_9 : deictic\_rel(e)\ l_4 : a_9 : ARG1(x_2)\ l_4 : a_9 : ARG2(x_1)$
$\oplus \boxed{N_{rel}} \oplus \{l_1 : a_1 : deictic\_q(x_0) l_1 : a_1 : RSTR(h_1) l_1 : a_1 : BODY(h_2)$
$l_{21} : a_2 : sp\_ref(x_1) l_2 : a_2 : ARG1(v(\vec{p}))$
$l_{22} : a_3 : \_hand\_shape\_flat(e_0)\ l_{22} : a_3 : ARG1(x_2)$
$\ldots$
$l_{26} : a_7 : \_hand\_location\_c(e_4)\ l_{26} : a_7 : ARG1(x_6)\}\}$
$[\boxed{D_{hcons}}]\ \{l_{21} = l_{22} = \ldots = l_{26};\ x_0 = x_1 = x_2 = \ldots = x_6\}\rangle$

$TIME\_START$ : 25.7
$TIME\_END$ : 26.73
$PHON$ : *nuclear*
$SYN$ : $N$
$SEM$ : $\langle h_0\,[l_4, a_9, x_2]\,\{\}$
$\boxed{N_{rel}}\{l_4 : a_9 : \_hallway\_n\_1(x_2)\}\}\,[]\,\{\}\rangle$
*hallway*

$TIME\_START$ : 25.5
$TIME\_END$ : 25.97
$PHON$ : $<>$
$SYN$ : $<>$
$SEM$ : $\langle h_0\,[l_3, a_8, i_1]\,\{\}$
$\{l_1 : a_1 : deictic\_q(i_0)\ \ l_1 : a_1 : RSTR(h_1)\ \ l_1 : a_1 : BODY(h_2)$
$l_{21} : a_2 : sp\_ref(i_1) l_2 : a_2 : ARG1(v(\vec{p}))$
$l_{22} : a_3 : \_hand\_shape\_flat(e_0)\ l_{22} : a_3 : ARG1(i_2)$
$\ldots$
$l_{26} : a_7 : \_hand\_location\_c(e_4)\ l_{26} : a_7 : ARG1(i_6)\}$
$\boxed{D_{hcons}}[\ h_1 =_q l_{21}]\,\{l_{21} = l_{22} = \ldots = l_{26};\ \ i_0 = i_1 = i_2 = \ldots = i_6\}\rangle$

$$\begin{bmatrix} \textit{deictic-abstract} & \\ \text{HAND-SHAPE} & \text{flat} \\ \ldots & \\ \text{HAND-LOCATION} & \vec{c} \end{bmatrix}$$

Figure 44: Derivation Tree for Deictic Gesture and the N “hallway”

by the quantifier. We forgo any further details about the derivation of the situated word because the rest is the same as in the derivation of depicting gesture.

### 5.3.2 Speech Phrase and Gesture Alignment

One of the central claims in this thesis is that the mappings from gesture form to gesture meaning are one-to-many rather than one-to-one: the ambiguity of gesture form gives rise to multiple interpretations in context which are supported by distinct alignments of speech and gesture. Here we remain neutral regarding the temporal relations between the aligned speech and gesture: the two modalities may happen at more or less the same time[12] or the gesture may align with speech elements whose temporal performance was completely outside the temporal performance of the gesture. For instance, for utterance (1.8), repeated in (5.27), it makes sense for the gesture to align with the whole VP "throw ground rice" and also with "throw ground rice over it" even though the gesture was performed along the verb head "throw" only. So not only that these alignments would support the pragmatic interpretations discussed in Section 2.2.1.1, but also the produced LFs respect constraints on reference that have been proposed in the literature on discourse semantics (see Section 5.1).

(5.27) He used to go down there and throw ...[$_{X^*}$grou]nd rice over it.

*The speaker moves his right hand forward; fingers are flexed inward in contact with the palm; the tip of the thumb is resting on the first joint of the index finger. The hand is moved twice by extending the wrist. The gesture resembles scattering a handful of dust/powder over some surface.*

Following Pustejovsky, who defines the argument structure of a lexical item as 'the minimal specification of its lexical semantics' [Pustejovsky, 1995, p. 63], we consider that gesture visualises some aspect(s) of the meaning of a lexical item that has been bound with its arguments. This means that gesture determines an abstract proposition that can semantically relate (and hence align) with a single spoken item, and also with a spoken item whose minimal contextual requirements have been (fully or partially) met. We claim that gesture can align not only with the synchronous, prosodically prominent element (per Definition 5.3.1), but also with an entire constituent, that is, a head combined with its arguments. From a descriptive perspective, the inclusion of more

[12] Of course, strict identity of performances is almost impossible, so by happening at the same time we rather mean a positive or negative delay of 275 msec, the distance for speech and gesture to be considered near each other (see Sections 3.2.3 and 4.2.2).

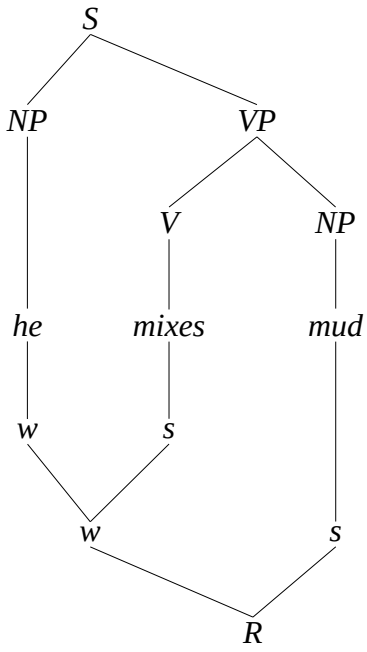



Figure 45: Discrepancy between Prosodic Constituency and Syntactic Constituency

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 the 'throwing-rice' example (5.27) the information about the direction of the throwing event (that is, downwards), the content of what is being thrown (that is, small particles) and even the location (the 'sieving-through' gesture implies that the action was performed over an extended surface) is denoted by a single visual performance and by several linearly ordered lexical items ("throw", "ground rice", "over it"). Furthermore, 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 cannot break the constituent that is attaches to.

A further complication arises from the fact that the syntactic structure is not necessarily identical to the prosodic structure: in Figure 45, for instance, a prosodic boundary separates "he mixes" and "mud" in two prosodic units. Note that this prosodic structure is based on the prosodic realisation in utterance (2.8) where the relative prosodic strength of "mud" is stronger than that of "he mixes". By tracing the path

dominated by strong nodes, we can established that the DTE is "mud". Of course, this does not mean that a prosodic structure that is identical to the syntactic one is not possible.

With all this in mind, we can now proceed by laying down the methodology of how to establish the possible attachments in the derivation tree. Attaching gesture to a constituent larger than a single prosodic word (as per Situated Prosodic Word in Definition 5.3.1) is licensed by the rule in Definition 5.3.2.

**Definition 5.3.2. Situated Spoken Phrase Constraint** *A depicting and/or deictic gesture can attach to a syntactic and/or a prosodic constituent* $\boldsymbol{xp}$ *of a spoken utterance and any of its higher projections no matter what the syntactic label is if (a.) there is an overlap between the temporal performance of the gesture stroke and* $\boldsymbol{xp}$*; (b). there is no (explicit or implicit) discourse connective at an inter-sentential or intra-sentential boundary crossed.*

We will now explain this definition in details. 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 denote different aspects of the speech signal, where the different denotations are supported by distinct attachments in the syntax tree. We will illustrate the construction rule using utterance (5.1). Here we assume a right-branching prosodic structure with the nuclear accent, i.e., the DTE, on "mud". The resolved LFs for this multimodal action (recall Section 5.1.1) featured coherence relations between the NP's denotation and the 'mixing' gesture, between the VP's denotation and the depicting gesture, and also between the clause and the gesture. Given Definition 5.3.2, these interpretations are supported by the following syntactic attachments:

1. gesture attachment to the nuclear prominent NP "mud". Given constraints on reference on the semantics/pragmatics interface, this attachment would block the gesture referring to anything that is bridging related to "mixes" or "he". This attachment would produce the LF in (5.2).

2. gesture attachment to the VP "mixes mud". This attachment enables the gesture to qualify "mixes", and hence the interpretation of the multimodal action is the one shown in (5.3).

3. gesture attachment to the clause “he mixes mud”, which enables reference to both the denotations of the NP “he” and the verb “mixes”. This attachment supports the interpretation in (5.4).

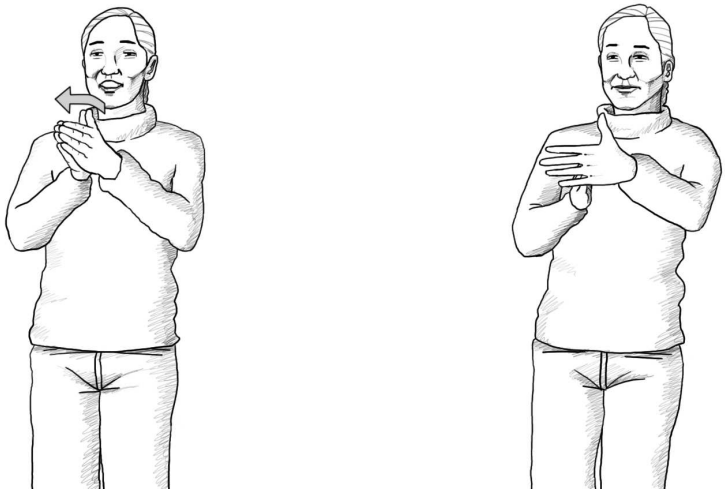
Furthermore, we block attaching a gesture to a constituent that contains a discourse connective since this would prevent us from exploring the relation between the two conjuncts. To motivate this, consider the real utterance from the Talkbank collection displayed in (5.28), Figure 46. Suppose the first gesture attaches to “keep going straight until I hit Broadway” and the second gesture attaches to the resulted multimodal tree, then this syntactic analysis would produce a logical form featuring the predications in (5.29) where *vis_rel_a* and *vis_rel_b* are underspecified semantic relations between speech and gesture, *s* is a supersort label of the speech content, $g_1$ labels the content of the first gesture and $g_2$ labels the content of the second gesture. Despite the fact that it is possible to resolve the underspecified relations, for instance, *vis_rel_a* might be a depiction of going straight and *vis_rel_b* might be a virtual identity between the gestural denotation and the denotation of the speech, we would not exploit how $g_1$ and $g_2$ are connected, namely through the temporal relation *until*: “I keep going straight *until* I hit Broadway” is different from “I keep going straight *after* I hit Broadway”.

(5.28) I keep going straight until I hit Broadway.
*Both hands are in the upper centre, palms are vertical touching each other, finger tips point forward (Figure 5.46(a)). Along with “Broadway”, the left hand with a vertically open palm is placed orthogonally in front of the right hand with finger tips pointing to the right (Figure 5.46(b)).*

(5.29) $l_1 : vis\_rel\_a(s, g_1)$
$l_2 : vis\_rel\_b(l_1, g_2)$

A preferred syntactic analysis would be therefore to attach the first stroke to “keep going straight” and the second stroke to “hit Broadway” producing thus a logical form featuring the predications in (5.30) where *vis_rel_a* and *vis_rel_b* are underspecified semantic relations, $s_1$ is the label of the content of the first clause, $s_2$ labels the content of the second clause, and $g_1$ and $g_2$ are labels of the contents of the first and the second gesture stroke, respectively.

(5.30) $l_1 : vis\_rel\_a(s_1, g_1)$
$l_2 : vis\_rel\_b(s_2, g_2)$



(a) I keep going straight until I hit (b) Broadway

Figure 46: Abstract Deictic Gesture Placing a Landmark in the Virtual Space, example (5.28)

$l_3 : \_until\_x\_h(e, h_1, h_2)$
$h_1 =_q l_1\ h_2 =_q l_2$

In a similar way, the attachments licensed for utterance (5.27) include the head combined with its complement ("throw ground rice") and also the head combined with both the complement and the PP ("throw ground rice over it"). Note, however, that the gesture cannot attach to the entire clause, that is, the head cannot be combined with the subject daughter "he" since this configuration would involve crossing the discourse connective "and".

Note also that Definition 5.3.2 used 'syntactic and/or prosodic constituent' to refer to any phrase of a hierarchical organisation: it being prosodic or syntactic. Assuming an analysis where there is no isomorphism between syntax and prosody, this flexibility is necessary whenever there are mismatches between the prosodic structure and the syntactic structure (recall Figure 45). Since gesture interacts with the rhythmical organisation of the utterance, there is no need to block an alignment to a prosodic constituent that does not correspond to a syntactic constituent in its traditional sense.[13]

[13]Combinatory Categorial Grammar is a notable exception in this sense, since it assumes an isomorphism between prosodic and syntactic constituency [Steedman, 2000].

We will now demonstrate the attachment possibilities yielded by the rules in Definition 5.3.2 when applied to utterances (5.23)—for depicting gestures—and (2.12)—for deictic gestures.

#### 5.3.2.1 Situated Spoken Phrase Constraint and Depicting Gesture

The application of the Situated Spoken Phrase Constraint, Definition 5.3.2, to utterance (5.23) would license the attachment of the depicting gesture to "couldn't believe it" and "I couldn't believe it" (see the illustration in Figure 47): the gesture stroke temporally overlaps the nuclear prominent "couldn't", which is also the syntactic head, and the gesture can thus attach to (i) the VP node after the head had been combined with its compliment only, the VP "believe it" and (ii) the S node after the head had been combined with its compliment and external argument, the subject "I" (despite its timing being outside that of the stroke). These attachments would produce an interpretation where the hand signal elaborates on what is expressed by the aligned speech, namely, the speaker could not believe the fact that the cupboard doors were crooked. Alternative parse trees are rendered by attaching the gesture to (i) the NP "one of the cupboards" (this configuration is licensed by the temporal overlap of the gesture stroke and the nuclear prominent head "one" saturated with its PP complement), (ii) the VP "is off", and (iii) the entire S "one of the cupboards is off". High attachments to the root node are particularly useful when the gesture serves the pragmatic function of qualifying the speech act rather than denoting salient features of the speech content. In such cases the gesture can be interpreted as a 'metatalk' [Polanyi, 1985] content paraphrasable as "I am requesting that", "I am giving you the floor", or something similar.

The syntactic analysis and the semantic composition is displayed in Figure 48, which features the situated phrase "one of the cupboards" + depicting gesture. For the sake of space, we have omitted the semantics of the single speech parts, providing only the composed underspecified meaning of the NP "one of the cupboards". Essentially, the logical form introduces an underspecified relation *vis_rel* between the ltop $l_3$ of the speech daughter and the ltop $l_0$ of the aligned gesture daughter and it demonstrates that the speech and the gesture are coherently related. The resolution of this relation is extraneous to the grammar and is subject to commonsense reasoning. It suffices to mention that in this context, *vis_rel* may resolve to Depiction — the hand movement provides the visual characteristics of the referent: the shape and size of the cupboard and the fact that it is not straight. Since the semantic composition follows the principles

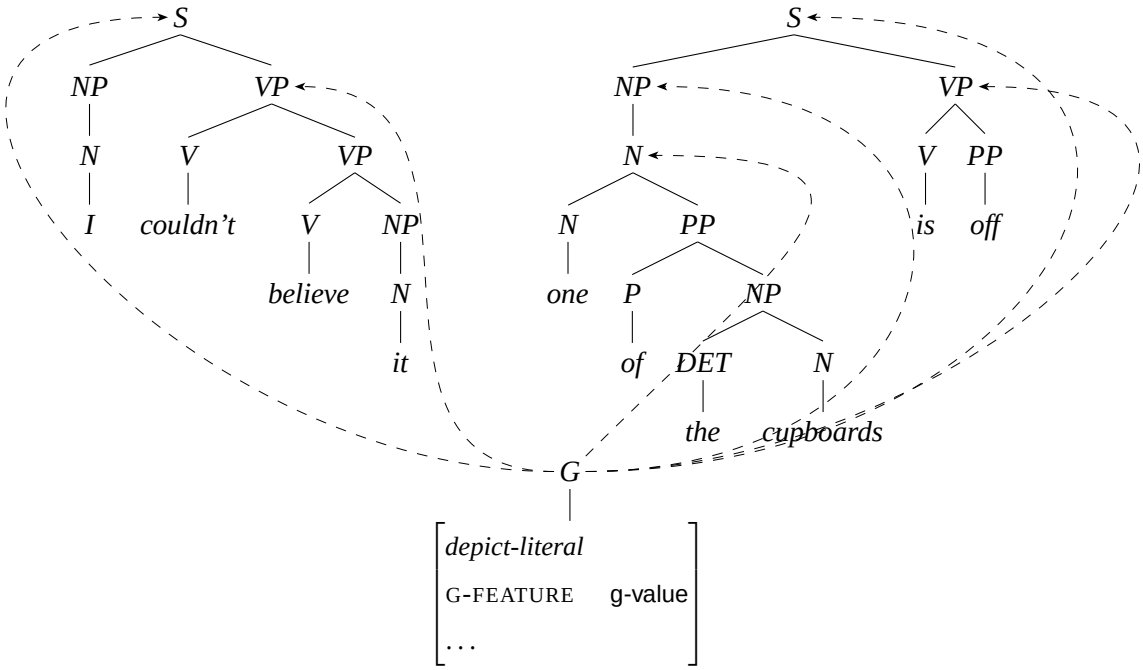



Figure 47: Depicting Gesture–Prosodic Constituent Attachment Ambiguities

introduced above, we forgo any further details.

Note also that there is an implicit connective between the two clauses that bars an attachment to the entire speech segment "I couldn't believe it: one of the cupboards is off" despite its temporal performance spanning the temporal performance of the stroke.

Recall from Section 2.2.2 that the grammaticality of the multimodal actions depends on the gesture temporally overlapping the nuclear accented item or any higher projections. We claimed that utterance (1.4) was ill-formed since the gesture was performed along a non-accented item in an all-rheme utterance. Having introduced the constraints in Definitions 5.3.1 and 5.3.2, we are now in a position to account for the utterance's well-formedness: if the gesture was performed along with "mother", these two rules would license attachments to the N "mother", the NP "your mother" and even to the S "your mother called". Still, nothing licenses an attachment to "called".

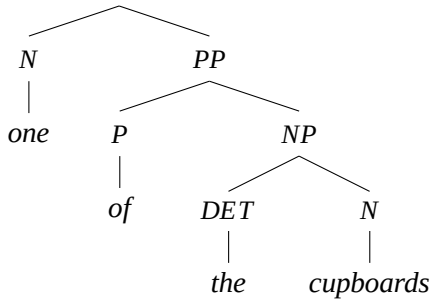
#### 5.3.2.2 Situated Spoken Phrase Constraint and Deictic Gesture

We start off by representing the deictic gesture from (2.12), repeated in (5.31), in a typed feature structure as shown in Figure 49. In the next chapter, we will refine this feature structure after we introduce the distinct gesture types and feature values.

(5.31) I [$_{PN}$enter] my [$_{N}$apartment]

$TIME\_START$ : 1.94
$TIME\_END$ : 3.92
$PHON$ : $R'$
$SYN$ : $NP'$
$SEM$ : $\langle h_0\,[l_6, a_6, e_8]\,\{\}$
$\{l_6 : a_6 : vis\_rel(e_1)\ l_6 : a_6 : S\text{-}LBL(l_3)\ l_6 : a_6 : G\text{-}LBL(l_0)\ l_6 : a_6 : M\text{-}ARG(e_8) \oplus \boxed{NP_{rels}} \oplus \boxed{G_{rels}}\}$
$[\boxed{NP_{hcons}} \oplus \boxed{G_{hcons}}]$
$\{\boxed{NP_{eqs}}\}\rangle$

$TIME\_START$ : 3.07
$TIME\_END$ : 3.92
$PHON$ : $R$
$SYN$ : $NP$
$SEM$ : $\langle h_0\,[l_3, a_3, e_7]\,\{\}$
$\boxed{NP_{rels}}\{l_4 : a_4 : \_one\_n\_1(x_4)$
$l_5 : a_5 : \_pronoun\_q(x_5)\ l_5 : a_5 : RSTR(h_3)\ l_5 : a_5 : BODY(h_4)$
$l_3 : a_3 : \_of\_p(e_7)\ l_3 : a_3 : ARG1(x_2)\ l_3 : a_3 : ARG2(x_3)$
$l_2 : a_2 : \_the\_q(x_0)\ l_2 : a_2 : RSTR(h_1)\ l_2 : a_2 : BODY(h_2)$
$l_1 : a_1 : \_cupboard\_n\_1(x_1)\}$
$\boxed{NP_{hcons}}[h_3 =_q l_4, h_1 =_q l_1]$
$\boxed{NP_{eqs}}\{l_3 = l_4, x_0 = x_1 = x_3, x_4 = x_5 = x_2\}\rangle$




$TIME\_START$ : 1.94
$TIME\_END$ : 3.88
$PHON$ : $<>$
$SYN$ : $<>$
$SEM$ : $\langle h_0\,[l_0, a_0, i_{1-10}]\,\{\}$
$\boxed{G_{rels}}\{l_0 : a_0 : [\mathcal{G}](h_2)$
$l_1 : a_1 : \_r\_hand\_shape\_flat(i_1)$
$\ldots$
$l_{10} : a_{16} : \_l\_hand\_movement\_right\_left(i_{10})\}$
$\boxed{G_{hcons}}[\,h_2 =_q l_1, \ldots, h_2 =_q l_{10}]$
$\{\}\rangle$

| | |
|---|---|
| R-HAND-SHAPE | flat |
| ... | |
| L-HAND-MOVEMENT | right-left |

Figure 48: Derivation Tree for Depicting Gesture and the NP “one of the cupboards”

$$\begin{bmatrix} \textit{deictic-abstract} & \\ \text{HAND-SHAPE} & \text{flat} \\ \text{PALM-ORIENT} & \text{towards-centre} \\ \text{FINGER-ORIENT} & \text{away-body} \\ \text{HAND-MOVEMENT} & \text{up-down} \\ \text{HAND-LOCATION} & \vec{c} \end{bmatrix}$$

Figure 49: TFS representation of the abstract deictic gesture in (5.31)

*Speaker's hands are in centre, palms are open vertically, finger tips point upward; along with "enter" they move briskly downwards.*

Based on Definition 5.3.2, the deixis could attach to "enter my apartment" or to the whole clause "I enter my apartment" (see the illustration in Figure 50) even though the temporal performance of the arguments happens outside the temporal performance of the gesture stroke.

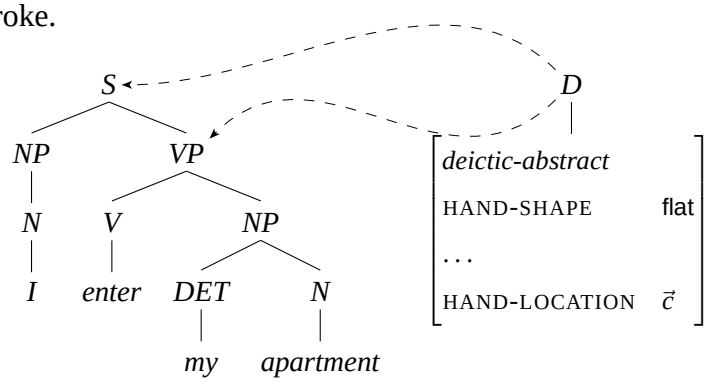



Figure 50: Deictic Gesture–Syntactic Phrase Attachment Ambiguities

Both attachments are licensed by the following factors: (a). the timing of the gesture stroke overlaps the timing of the pre-nuclear accented "enter"; (b). the gesture could attach to any of its projections, namely: the pre-nuclear accented word is the syntactic head in the speech phrase, which can be saturated with its internal argument only licensing thus an attachment to the verb phrase "enter my apartment", or it can be saturated with both the external and the internal arguments licensing thus an attachment to the entire clause "I enter my apartment". There is no reason to block any of these syntactic configurations since the logical forms they map to are supported in the final context of use: we argue that an attachment to the VP node would support a gestural interpretation from the 'observer's viewpoint' [McNeill, 2005] where the hands place the object in the virtual space *proximal* in relation to the speaker; and an attachment to

the root S node would support an interpretation from the first-person, or 'character's viewpoint' [McNeill, 2005] where the speaker is *inside* the gesture space, and she thus uses the hand signal as an enactment of her entering the apartment door.

Once we have established the nodes of attachment, we can proceed with composing the semantics of the multimodal utterance. Figure 51 exemplifies the derivation tree and the corresponding semantics yielded by the alignment of the deictic gesture to the entire clause. We forgo any details about the TIME_START and TIME_END values since their behaviour is the same as with depicting gestures. Since the individual steps of composing the meaning representation of the spoken segment are not relevant for the current analysis, we introduce the already composed underspecified semantics. For the deictic daughter, we omit again the entire feature structure representation, and the corresponding elementary predications. Composing the sement of the speech daughter and the sement of the deictic daughter involves combining the semantics of the speech daughter, the semantics of the deictic daughter and introducing an underspecified relation $deictic\_rel(e_{13}, e_6)$ between the semantic index $e_{13}$ of the speech daughter and the semantic index $e_6$ of the gesture daughter. Since in this instance we align an entire clause with a gesture, the semantic index of the speech daughter is identified with the event variable $e_{13}$ of the ULF's main predication:

$$l_9 : a_{13} : \_enter\_v\_1(\boldsymbol{e_{13}})\ \ l_9 : a_{13} : ARG1(x_7)\ \ l_9 : a_{13} : ARG2(x_8).$$

Note also that in composition, the underspecified index $i_1$ of the gesture daughter's main predication

$$l_{21} : a_2 : sp\_ref(\boldsymbol{i_1})\ \ l_2 : a_2 : ARG1(v(\vec{p}))$$

unifies with the event variable of the speech segment, i.e., $i_1$ resolves to $e_6$. In other words, the deictic gesture denotes the event of entering the apartment door on the virtually created map just in front of the speaker. The label of the deictic relation is shared with the ltop of the speech daughter. In this way, any predication that takes as argument the speech daughter, would also take as argument the deictic daughter; for instance, in "John believes that I enter my apartment", the second argument of "believe"—i.e., the argument filling the object slot—will identify with the label of "enter" and hence with the label of *deictic_rel*.

Now, we will use this example to demonstrate that gesture can align with a prosodic constituent that does not correspond to a syntactic constituent. Following Definition 5.3.2, the gesture in utterance 2.12 could attach to the prosodic constituent "I enter" as shown in Figure 52. In this case of a mismatch between prosody and syntax,

$PHON : R'$

$SYN : S'$

$SEM : \langle h_0\,[l_9, a_{13}, e_{13}]\,\{\}\{l_9 : a_{13} : deictic\_rel(e_{13})\ \ l_9 : a_{13} : ARG1(e_{13})\ \ l_9 : a_{13} : ARG2(e_6)$

$\oplus \boxed{S_{rels}} \oplus \{l_1 : a_1 : deictic\_q(e_5)\ \ l_1 : a_1 : RSTR(h_1)\ \ l_1 : a_1 : BODY(h_2)$

$l_{21} : a_2 : sp\_ref(e_6)\ l_2 : a_2 : ARG1(v(\vec{p}))$

$l_{22} : a_3 : \_hand\_shape\_flat(e_0)\ l_{22} : a_3 : ARG1(e_7)$

$\ldots$

$l_{26} : a_7 : \_hand\_location\_c(e_4)\ l_{26} : a_7 : ARG1(e_{12})\}\}$

$\left[\boxed{S_{hcons}} \oplus \boxed{D_{hcons}}\right]\ \{\boxed{S_{eqs}} \oplus \{l_{21} = l_{22} = \ldots = l_{26}\ \ e_5 = e_6 = e_7 \ldots = e_{12}\}\}\rangle$

$PHON : R$

$SYN : S$

$SEM : \langle h_0\,[l_9, a_{13}, e_{13}]\,\{\}$

$\boxed{S_{rels}}\{l_4 : a_8 : \_apartment\_n\_1(x_1)$

$l_5 : a_9 : pron(x_2)$

$l_6 : a_{10} : pronoun\_q(x_3)\ \ l_6 : a_{10} : RSTR(h_2)\ \ l_6 : a_{10} : BODY(h_3)$

$l_7 : a_{11} : poss(e_2)\ \ l_7 : a_{11} : ARG1(x_4)\ \ l_7 : a_{11} : ARG2(x_5)$

$l_8 : a_{12} : def\_explicit\_q(x_6)\ \ l_8 : a_{12} : RSTR(h_3)\ \ l_8 : a_{12} : BODY(h_4)$

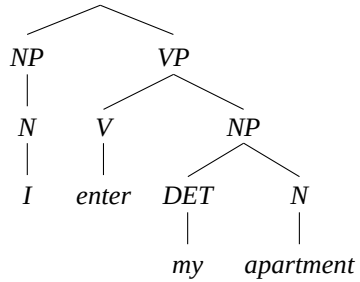
$l_9 : a_{13} : \_enter\_v\_1(e_{13})\ \ l_9 : a_{13} : ARG1(x_7)\ \ l_9 : a_{13} : ARG2(x_8)$

$l_{10} : a_{14} : pronoun\_q(x_9)\ \ l_{10} : a_{14} : RSTR(h_5)\ \ l_{10} : a_{14} : BODY(h_6)$

$l_{11} : a_{15} : pron(x_{15})\}$

$\boxed{S_{hcons}}[h_2 =_q l_5\ \ h_3 =_q l_4\ \ h_5 =_q l_{11}]$

$\boxed{S_{eqs}}\{x_2 = x_3 = x_5\ \ x_1 = x_4 = x_6 = x_8\ \ l_4 = l_7\ \ x_7 = x_9 = x_{15}\}\rangle$




$PHON : <>$

$SYN : <>$

$SEM : \langle h_0\,[l_3, a_1, i_1]\,\{\}$

$\{l_1 : a_{13} : deictic\_q(i_0)\ \ l_1 : a_{13} : RSTR(h_1)\ \ l_1 : a_{13} : BODY(h_2)$

$l_{21} : a_2 : sp\_ref(i_1)\ \ l_2 : a_2 : ARG1(v(\vec{p}))$

$l_{22} : a_3 : \_hand\_shape\_flat(e_0)\ l_{22} : a_3 : ARG1(i_2)$

$\ldots$

$l_{26} : a_7 : \_hand\_location\_c(e_4)\ l_{26} : a_7 : ARG1(i_6)\}$

$\boxed{D_{hcons}}[\ h_1 =_q l_{21}]$

$\{l_{21} = l_{22} = \ldots = l_{26}\ \ i_0 = i_1 = i_2 \ldots = i_6\}\rangle$

$$\begin{bmatrix} deictic\text{-}abstract & \\ \text{HAND-SHAPE} & \text{flat} \\ \ldots & \\ \text{HAND-LOCATION} & \vec{c} \end{bmatrix}$$

Figure 51: Derivation Tree for Deictic Gesture and the S “I enter my apartment”

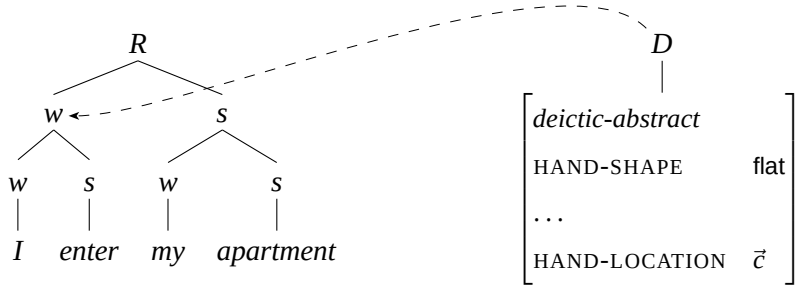



Figure 52: Attachment of a Deictic Gesture to a Prosodic Constituent

we attach the gesture to a higher projection within the prosodic tree. Syntactically this means that the head of the phrase "enter" is saturated with its external argument, the subject "I".

Assuming an architecture where the prosodic structure interfaces the semantic information, we can derive a structure as shown in Figure 53. Informally, the alignment of a prosodic constituent and a gesture of type *deictic-abstract* projects a prosodic constituent that is semantically related with a gesture. In contrast to the structure in Figure 51, there is no syntactic information concerning the linguistic phrase (hence the SYN value is empty). To demonstrate the fact that we combine a phrase that is not fully saturated syntactically, we have a non-empty COMP1 slot:

$$\{[l_3, a_3, x_{11}]_{comp1}\}$$

says that the functor with anchor $a_3$ (and scopal label $l_3$) anticipates a complement of type individual. This also explains why ARG2 of the predicate *_enter_v_1*—the argument that corresponds to the complement—is left underspecified (expressed through $i$). In composition with the complement "my apartment", the non-empty slot would fill with the missing complement thereby emptying the slot list and identifying ARG2 of *_enter_v_1* with ARG0 of the complement. This would be accounted for by equating $x_{11}$ from the slot with $x_8$, and also the underspecified index $i$ with $x_8$ (based on Figure 51 where $x_8$ is the complement argument of the predicate *_enter_v_1*).

**Preliminary Conclusions.** The construction rules so far account for well-formed multimodal utterances where the gesture aligns with a prosodic word or a constituent whose element(s) overlap the temporal performance of the gesture. These constructions, however, are not sufficient as they do not reflect an important finding from our empirical investigation. Recall from Section 4.3 that deictic gesture does *not* necessarily overlap the prosodically prominent word and/or that it can happen outside the

$PHON : R$

$SYN : <>$

$SEM : \langle h_0\,[l_3, a_3, e_{12}]\,\{[l_3, a_3, x_{11}]_{comp1}\}$

$\{l_3 : a_3 : deictic\_rel(e_{13})\ \ l_3 : a_4 : ARG1(e_{12})\ \ l_3 : a_4 : ARG2(e_6)$

$\oplus \boxed{s_{rels}} \oplus \{l_4 : a_4 : deictic\_q(e_5)\ \ l_4 : a_4 : RSTR(h_3)\ \ l_4 : a_4 : BODY(h_2)$

$l_{51} : a_5 : sp\_ref(e_6)\ l_{51} : a_5 : ARG1(v(\vec{p}))$

$l_{52} : a_6 : \_hand\_shape\_flat(e_0)\ l_{52} : a_6 : ARG1(e_7)$

$\ldots$

$l_{56} : a_{10} : \_hand\_location\_c(e_4)\ l_{56} : a_{10} : ARG1(e_{11})\}\}$

$\left[\boxed{s_{hcons}} \oplus \boxed{D_{hcons}}\right]$

$\{\boxed{s_{eqs}} \oplus \{l_{51} = l_{52} = \ldots = l_{56}\ \ e_5 = e_6 = e_7 = \ldots = e_{11}\}\}\rangle$

$PHON : R$

$SYN : <>$

$SEM : \langle h_0\,[l_3, a_3, e_{12}]\,\{[l_3, a_3, x_{11}]_{comp1}\}$

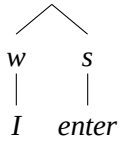
$\boxed{s_{rels}}\{l_1 : a_1 : pron(x_5)$

$l_2 : a_2 : pronoun\_q(x_6)\ \ l_2 : a_2 : RSTR(h_1)\ \ l_2 : a_2 : BODY(h_2)$

$l_3 : a_3 : \_enter\_v\_1(e_{12})\ \ l_3 : a_3 : ARG1(x_7)\ \ l_3 : a_3 : ARG2(i)\}$

$\boxed{s_{hcons}}\,[h_1 =_q l_1]$

$\boxed{s_{eqs}}\{x_5 = x_6 = x_7\}\rangle$




$PHON : <>$

$SYN : <>$

$SEM : \langle h_0\,[l_6, a_{11}, i_1]\,\{\}$

$\{l_4 : a_4 : deictic\_q(i_0)\ \ l_4 : a_4 : RSTR(h_3)\ \ l_4 : a_4 : BODY(h_2)$

$l_{51} : a_5 : sp\_ref(i_1)\ \ l_{51} : a_5 : ARG1(v(\vec{p}))$

$l_{52} : a_6 : \_hand\_shape\_flat(e_0)\ l_{52} : a_6 : ARG1(i_2)$

$\ldots$

$l_{56} : a_{10} : \_hand\_location\_c(e_4)\ l_{56} : a_{10} : ARG1(i_6)\}$

$\boxed{D_{hcons}}\,[\,h_3 =_q l_{51}]$

$\{l_{51} = l_{52} = \ldots = l_{56}\ \ i_0 = i_1 = i_2 = \ldots = i_6\}\rangle$

$$\begin{bmatrix} \textit{deictic-abstract} & \\ \text{HAND-SHAPE} & \text{flat} \\ \ldots & \\ \text{HAND-LOCATION} & \vec{c} \end{bmatrix}$$

Figure 53: Derivation Tree for Deictic Gesture and the Prosodic Constituent “I enter”

temporal performance of the semantically related word provided that the deixis referent is salient in the communicative event. In the next section we propose a rule that takes this condition into account.

### 5.3.3 Spoken Word and Gesture Alignment: Temporal and Prosodic Relaxation

**Definition 5.3.3. Deictic Prosodic Word with Defeasible Constraint.** *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.*

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 (2.3), (4.8), (4.9) and (4.11), the relaxation is a matter of making individuals in the surrounding space salient and it is thus necessary only in utterances where the gesture's 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's 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 felicity of the interpretation. 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, Rieser, and Staudacher [2006]).

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 (2.3), (4.8), (4.9) and (4.11) takes care for aligning the gesture and the semantically related element—i.e., "she" in (2.3), "she" in (4.8), "hand" in (4.9), "he" in (4.11))—the gesture in (4.7) 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 gesture in (4.7) 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 (4.8), repeated in (5.32).

(5.32) ... [$_{PN}$She] works in the [$_{NN}$cubicle] [$_{N}$next] to me ...
*Speaker's right hand is loosely open, points at the participant diagonally from the speaker.*

The Situated Prosodic Word in Definition 5.3.1 would license attachments to the temporally overlapping prosodically prominent "cubicle" and "next". Similarly, the Situated Spoken Phrase Constraint, Definition 5.3.2 would license attachments to "next to me", "cubicle next to me", "works in the cubicle next to me", etc. Although syntactically well-formed, we claimed that these attachments would produce logical forms that are not supported in the specific context: combining the deictic gesture (which identifies an individual seated diagonally from the speaker) and the temporally overlapping "cubicle" would not support 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 Definition 5.3.3: the deictic gesture attaches to "she" thereby providing an interpretation where the gesture's denotation is identical to the denotation of the pronoun "she".

At this stage, we remain agnostic whether this relaxation is also applicable to phrases.

## 5.4 Summary

In this chapter, we provided evidence how the discourse structure (that is, what elements attach to what other elements) is highly constraining on the final interpretations, given the current models of the semantics/pragmatics interface. The choices of alignment are the basic building blocks for constraining reference, which ultimately affect how the multimodal actions are interpreted in context. The grammar construction rules proposed in Section 5.3 not only comply with constraints on reference in discourse, but they also comply with well-established principles for semantic composition in purely linguistic grammars. In this way, the result of our grammar is entirely compatible with the current models of pragmatic interpretation.

The construction rules capture our findings from the empirical investigation in Chapter 4. Our basic rule, the Situated Prosodic Word Constraint accounts for syntactic well-formedness by integrating a nuclear or a pre-nuclear prominent word and

the temporally overlapping gesture stroke into a multimodal derivation tree, that, using the semantic algebra with (R)MRS maps to a unified multimodal meaning.

Further to this, to capture our central claim in this thesis, namely—gesture ambiguous form gives rise to one-to-many form-meaning mappings—we introduced the Situated Spoken Phrase Constraint which licenses attachments to the temporally overlapping phrase no matter whether its constituent elements happen along the temporal performance of the gesture stroke or outside of it. This rule also accounts for cases where the prosodic phrasing is not identical to the syntactic phrasing.

Finally, we introduced a rule with a defeasible constraint which licenses an attachment to a word that is not prosodically prominent such as (2.3) or to a word whose temporal performance does not overlap the temporal performance of the gesture such as (4.8). This rule accounts for the fact that certain multimodal utterances where the gesture precedes or follows the semantically related speech element are still perceived as well-formed. We established that this happens with referents that are salient in the physical space identified by the deixis.

In this chapter, we also introduced the symbolic representation of gesture form and its mapping to (underspecified) meaning. In Section 5.2, we presented well-established methods for formalising gestural form with typed feature structures. This representation captures the fact that the meaning of the gesture is not composed from the meanings of its parts, but rather the meanings of the parts are derivable from the meaning of the whole. The feature structures are typed as *depicting* or *deictic* which helps us to build the adequate underspecified logical formula for the gesture: whereas depicting gestures require qualitative values, deictic gestures are anchored to the space and time of the communicative action and they thus require quantitative values. Further, to capture the incomplete meaning derived from gesture form, we use the semantic formalism of Robust Minimal Recursion Semantics since it allows us to build underspecified logical formulae that give a very abstract idea of what the signal means in any context. This is particularly essential for gestures since they, unlike linguistic input, underspecify meaning even in the final context of use.

The theoretical framework was presented in a grammar formalism neutral way so as to raise our understanding about the different ways gesture can align with speech. In this way we fleshed out our main argument that gestures are part of language and are thus suitable for modelling using standard methods from linguistics.

In the next section, we formalise the construction rules in a constraint-based grammar framework.

# Chapter 6

# An HPSG-based Account

Having articulated the well-formedness constraints on speech and co-speech gesture independently from a specific grammar theory, in this chapter we formalise them into Head-Driven Phrase Structure Grammar (HPSG [Pollard and Sag, 1994]). The chapter is organised as follows: in Section 6.1 we put forth our motivation for the HPSG framework, then in Section 6.2 we present background information about how metrical trees and MRS underspecified semantics are represented in typed feature structures; we proceed with Section 6.3 where we propose the gesture type hierarchy and finally in Section 6.4 we provide an HPSG-based analysis of the grammar construction rules proposed in Section 5.3.

## 6.1 Why HPSG?

The ultimate goal of this work is an implemented grammar for multimodal language that offers extensive coverage over distinct syntactic constructions for the linguistic component and a range of gestures. We therefore intend to augment an existing wide-coverage grammar with the construction rules from Chapter 5, formalised into a constraint-based grammar framework. With this in mind, the suitable grammar formalism should be used for developing computational grammars that meet the following requirements:

1. *Extensive coverage of linguistic constructions.* We intend to demonstrate the distinct ways gestures can be bound with various syntactic constructions, and so implementing from scratch a grammar with a coverage of various linguistic phenomena would be a rather demanding and time-consuming effort.

2. *Hand-crafted precision grammar.* While the automatic learning from annotated treebanks has proven efficient for language development tasks (e.g., Hockenmaier and Steedman [2005]), we are not familiar with existing multimodal treebanks that can be used for inducing grammars for speech and co-speech gesture. The grammar development effort will therefore involve manual specification of rules for speech and gesture alignment. In this respect, the grammar formalism suitable for multimodal grammar engineering should be used for manually implementing precision grammars in an appropriate grammar engineering platform.

3. *Constraint-based grammar framework.* As previously attested, the alignment between speech and gesture is constraint-based where the constraints come from the prosodic and syntactic properties of the speech signal that the gesture temporally overlaps with. The grammar formalism should therefore provide mechanisms for encoding prosodic information in parallel with the syntax/semantics component. Moreover, the prosodic information should be derivable in a *structured* way: for instance, attaching gesture to "I enter" (recall Figure 53) is a matter of deriving structured phonology where a construction rules licenses the attachment of gesture to a prosodic constituent that is not a syntactic constituent in its traditional sense. We do not assume that isomorphism between prosodic structure and surface syntactic structure is a necessary condition for encoding well-formedness constraints on multimodality [Butt and King, 1998; Klein, 2000a; Bögel et al., 2009]. From this perspective, any grammar framework that encodes prosodic properties and derives a structured prosodic representation built on these properties may be a suitable candidate for analysing multimodality.

4. *Semantic underspecification.* The semantic component should support underspecification of the predicate's arity, the predicate's main argument and scope, of the type discussed in Section 5.2.

5. *Support of operations over typed feature structures.* Recall from Section 5.2.1 that we use typed feature structures to first define the gesture type and so to determine whether the gesture should be represented in terms of qualitative or quantitative values, and second to capture the different aspects of gesture form that potentially have effects on the semantics of the multimodal action. Therefore, the description language of the specific grammar theory should support

operations over typed feature structures and also a hierarchical organisation of types.

With these criteria in mind, we chose to analyse the multimodal construction rules in the Head-driven Phrase Structure Grammar theory — a unification-based grammar framework where the linguistic objects are represented in typed feature structures with the types organised in a subsumption hierarchy. Our choice stems from the fact that HPSG was used for the development of the broad-coverage grammar for English—the English Resource Grammar (ERG [Flickinger, 2000])—which can be scaled up to new types, syntactic rules and lexical rules for gesture. An added bonus is that ERG along with the grammar engineering environment used for developing unification-based grammars—the Linguistic Knowledge Builder (LKB [Copestake, 2002])—are free and open-source products. While there exist industry-availale platforms for developing grammars (e.g., XLE),[1] the accessibility of ERG and the supporting engineering components was an important factor.

Further to this, HPSG offers mechanisms to derive structured phonology in parallel with syntax, as proposed by Klein [2000a; 2000b]. In this approach, the prosodic analysis is based on the metrical phonology framework, and so the mapping from our prosodic annotation to strings in TFSs equipped with prosody and suitable for parsing is a straightforward task. Moreover, the semantic component in recent work on HPSG implementations (e.g., the English Resource Grammar [Flickinger, 2000], the LinGO Grammar Matrix [Bender and Oepen, 2002]) is expressed in Minimal Recursion Semantics (MRS [Copestake et al., 2005]) which is entirely compatible with RMRS, the framework we use for capturing the highly underspecified content of gesture given its form (see Section 5.2.2). This is particularly useful since we can easily formalise the underspecified semantic representations of our choice into feature structures, the formal building block for articulating a grammar in the HPSG framework. 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. At this stage, we remain agnostic as to whether information structure should be encoded in the grammar with the view of restricting the choices of speech-and-gesture alignment.

[1] http://www2.parc.com/isl/groups/nltt/xle/

$$\begin{bmatrix} \text{PHON} & \text{*list*} \\ \text{SYNSEM} & \text{synsem} \end{bmatrix}$$

Figure 54: Representation of PHON and SYNSEM attributes in HPSG

Our choice for encoding the grammar constructions rules in HPSG can be viewed as a proof of concept that formal grammar theories can scale up to analysis of multimodal communicative actions. Needless to say, other grammatical formalisms and frameworks can also be suitable for analysing speech and gesture signals (cf. Giorgolo and Asudeh [2011]).

## 6.2 Background

In this section we provide background information about how metrical trees are mapped to feature structure representations, and also how the MRS semantics is formally expressed in feature structures.

### 6.2.1 Metrical Trees in Typed Feature Structures

In HPSG, the phonological component has been standardly represented within the PHON attribute in parallel with the syntax-semantics information SYNSEM, as illustrated in Figure 54. Traditionally, the PHON value of the mother is obtained by appending the daughters' phonologies. This operation produces a flat list of objects with no hierarchical structure. For illustration, Figure 55 features the PHON value for the string "your mother called" which involves appending (represented by the $\oplus$ symbol) the PHON values of its parts: the head daughter (HD-DTR) and the non-head-daughter (NON-HD-DTR).

This non-hierarchical representation is not sufficient for deciding which speech elements gesture can align with since it lacks information about the relative prosodic prominence of the subelements: for instance, we need some mechanisms of encoding that "called" is unstressed and hence prosodically weaker than "your mother" so as to bar an alignment of a depicting gesture to "called" (recall from Section 1.3.2.1 that this would be ill-formed). Alternatively, an alignment to "called" would be allowed in cases where the verb bears the nuclear accent (for instance, as an answer to "Did my mother stop by today?").

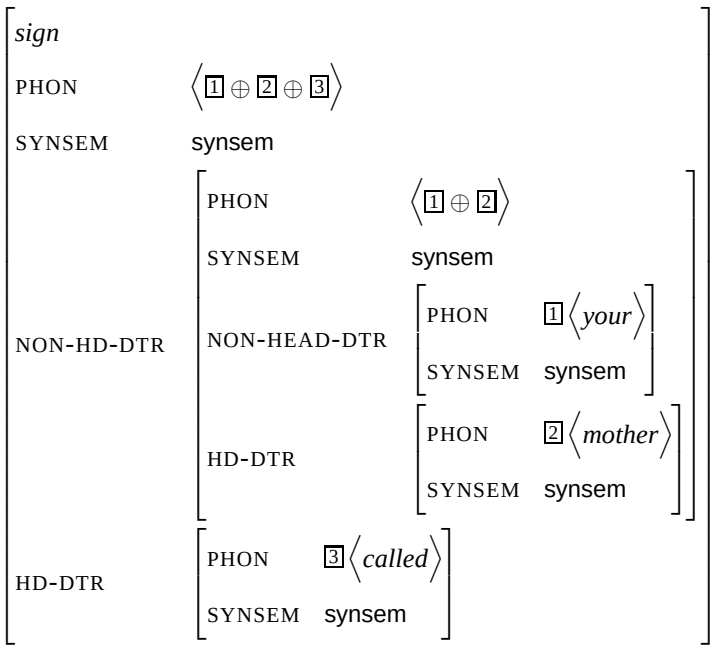



Figure 55: An example of the standard HPSG-based representation of PHON as an unstructured list of objects

We adopt the approach of Klein [2000a; 2000b] since it modifies the unstructured phonological representation by deriving a *structured* phonology of nested objects. This involves mapping the metrical tree to a feature structure in the following manner: the element dominated by strong *s* nodes maps to the Designated Terminal Element (DTE) [Liberman and Prince, 1977], and the elements dominated by weak *w* nodes map to the list of domain objects within the DOM attribute. In the adopted framework, the domain objects receive the generic label *prosodic constituents (pros)* which subsume all objects from prosodic words to intonational phrases. Furthermore, each input element is assigned a type from the prosodic type hierarchy displayed in Figure 56 to reflect the fact whether the prosodic object is a single prosodically marked element, a single prosodically unmarked element, a prosodic structure of prosodically marked elements or a prosodic structure of prosodically marked and unmarked domain elements. As it is standard for type hierarchies, the types constraints are inherited from their supertype(s).

The prosodic domain objects of type *pros* can be either of type *lnr (leaner)* or type *full*. Klein [2000a] adopts the *leaner* class to refer to those lexical items that

> . . . form a rhythmic unit with the neighbouring material, are normally unstressed with respect to this material, and do not bear the intonational peak

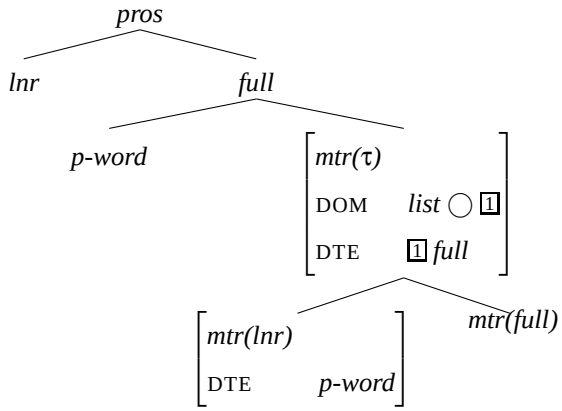



Figure 56: Prosodic Type Hierarchy [Klein, 2000a]

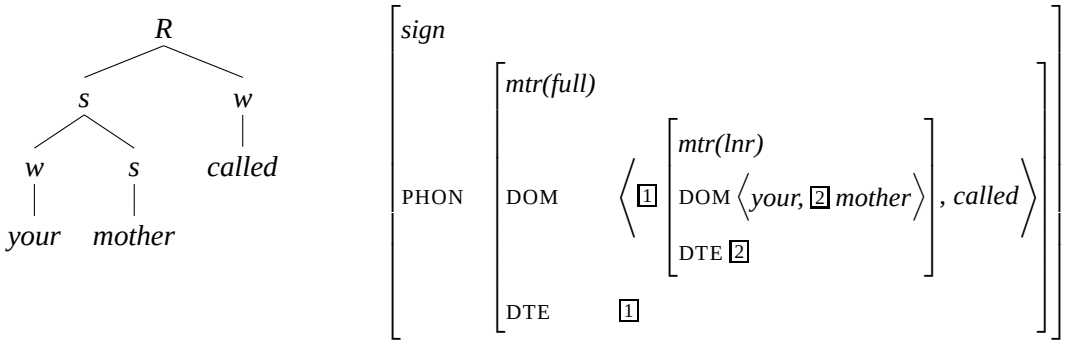



Figure 57: Metrical Tree and a Corresponding Feature Structure

> of the unit. English articles, coordinating conjunctions, complementisers, relative markers, and subject and object pronouns are all leaners in this sense [Zwicky, 1982, p. 5].

Then the type *full* subsumes prosodically marked words (type *p-word*) and metrical trees (type *mtr(τ)*). A metrical tree is specified in terms of its DOM(ain) and DTE values. The domain union relation ◯ is used to combine a list of objects with an element of type *full* which is also the DTE of the phrase.[2] Further, metrical trees subsume a leaner group of type *mtr(lnr)*, where one or more domain objects can be of type *lnr* and where the DTE is necessarily prosodically marked, i.e., a *p-word*. Metrical trees also subsume prosodic constituents of type *mtr(full)* where all daughters are *full*.

With this in mind, we can now map the metrical tree from Figure 26, page 93 to

[2] Haji-Abdolhosseini [2003] compares the domain union relation to shuffling of cards: the relative order of the elements is preserved in the final list but the order of adjacent elements might be disrupted. For us this means that the object of type *full* is not restricted to be the last member in the group.

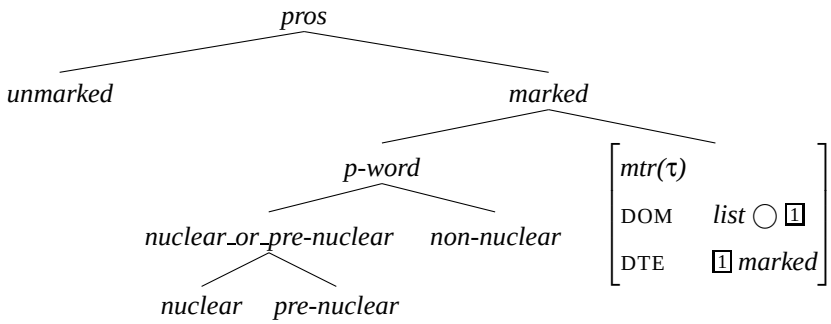



Figure 58: Our Prosodic Type Hierarchy

a feature structure representation, as displayed in Figure 57. In contrast to the flat list of phonological objects (displayed in Figure 55), Klein's [2000b] approach involves the derivation of prosodic constituents based on the prosodic prominence of one object relative to the other. We shall use this analysis to encode the prosodic type information and hence to constrain the attachments of gesture in the syntactic tree.

To reflect our empirical findings from Chapter 4, we are going to use a slight modification of Klein's [2000a] type hierarchy (see Figure 58). Like before, the root node of the type hierarchy is a prosodic constituent of type *pros* which subsumes prosodically marked (type *marked*) and prosodically unmarked objects (type *unmarked*). We also assume that *pros* is a generic label that subsumes all kinds of prosodic objects from words to intermediate and intonational phrases. In contrast to Klein [2000a] who restricts prosodic unmarkedness to the class of leaners, our type hierarchy uses the more general type *unmarked*. Furthermore, the marked object can be realised as a single element or as metrical trees (type *mtr(τ)*). We use *p-word* as a supertype for all prosodically marked singletons. Since only nuclear prominence and pre-nuclear prominence plays a role, we differentiate a type *nuclear_or_pre-nuclear* which is then inherited by *nuclear* and *pre-nuclear*. Note that the type hierarchy of Klein [2000a] does not distinguish between the type of the accent. Also, we do not differentiate between metrical trees of type *mtr(full)* and *mtr(lnr)* since this has no effects on the speech-gesture alignments. In other words, we consider that gestures can align with any prosodic phrase (of course, if licensed by other constraints) no matter whether all of its domain objects are prosodically marked or some are prosodically unmarked; for instance, there is no reason to encode the different metrical status of say, "your mother" vs. "John's mother" as long as they form a single prosodic constituent.

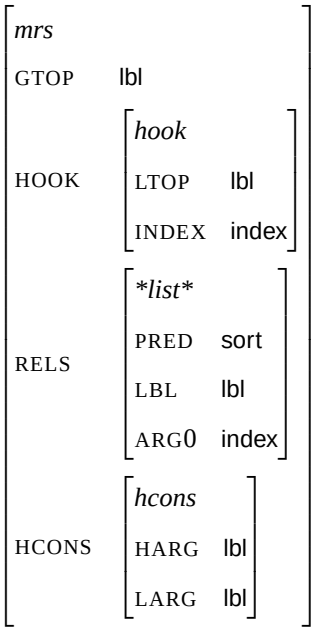



Figure 59: MRS Feature Structure Representation

### 6.2.2 Minimal Recursion Semantics in Typed Feature Structures

MRS and its highly factorised representation RMRS, reported in Section 5.2.2.1, were designed for semantic representation and semantic composition within feature structure grammars. We shall now demonstrate how the underspecified semantics is formalised within HPSG feature structure grammars [Copestake et al., 2005].

In HPSG, the semantic component is expressed within the SYNSEM | CONT attribute in parallel with the SYNSEM | CAT syntactic information. The CONT feature is of type *mrs*, a feature structure where the appropriate features are directly read off the underspecified semantics detailed in Section 5.2.2.1. The MRS feature structure representation has been demonstrated in Figure 59. The GTOP attribute encodes the global label Top: this is the top label that outscopes the entire logical formula contributed by a phrase; the HOOK is of type *hook* and it introduces the local top label and the semantic index (see Section 5.2.2.1). The LTOP and INDEX features are useful as they designate the element to be picked up as an argument during semantic composition. RELS introduces a list of feature structures where each feature structure encodes the PRED, LBL and ARG0 values. PRED introduces the elementary predication's symbol, LBL introduces the label of the predicate and ARG0 encodes the predicate's main argument. Other RELS attributes, apart from those specified in Figure 59, are also possible

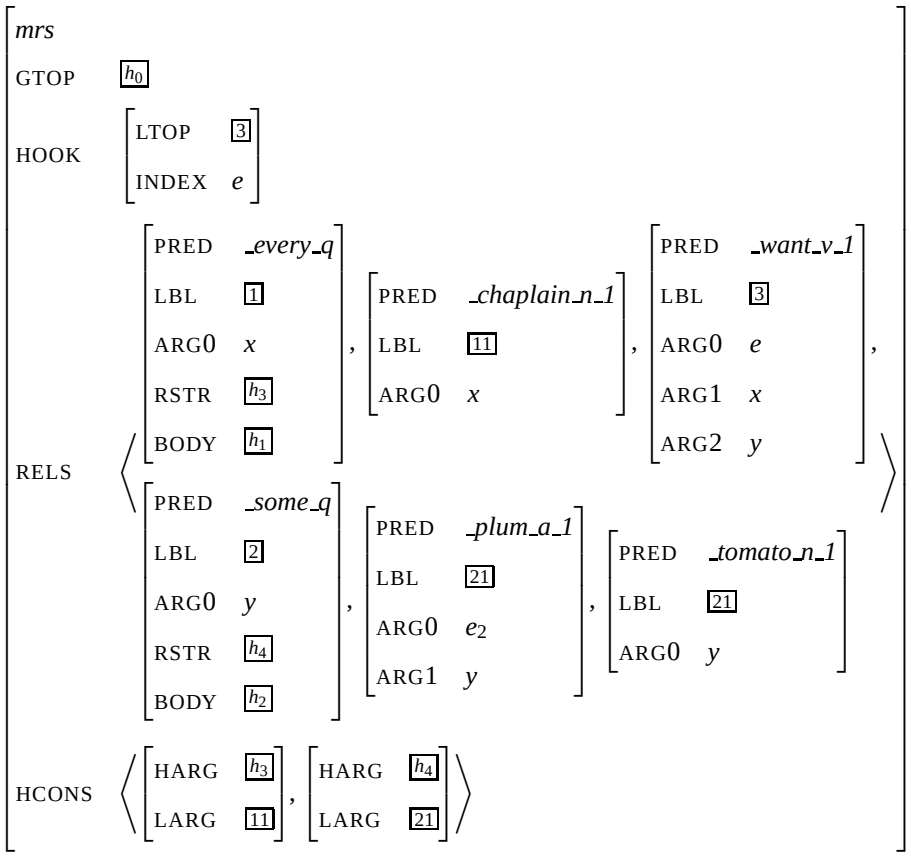



Figure 60: MRS Representation in Feature Structures

for the different classes of elementary predications. For instance, scopal relations also introduce a RSTR (restrictor) attribute and a BODY attribute where RSTR refers to the hole argument whose scope is fixed by the quantifier and BODY refers to the hole argument that the quantifier takes scope over; relations contributed by intransitive verbs also introduce an ARG1 attribute where ARG1 is identified with ARG0 of the subject EP; relations contributed by transitive verbs would have an additional ARG2 attribute that is identified with ARG0 of the direct object's predication; and relations of ditransitive verbs would also encode an ARG3 attribute equated with ARG0 of the indirect object's predication. Finally, HCONS specify the scopal conditions where HARG is the hole argument and LARG—the label of the argument plugged into the hole.[3]

For an illustration, consider Figure 60 which shows the typed feature structure representation for "every chaplain wants some plum tomato" mapped from the non-feature structure format in (5.15), repeated in (6.1).

[3]In line with the MRS implementation in the English Resource Grammar, we shall not represent the slots and equations in feature structures.

(6.1) $h_0$

$l_1 : \_every\_q(x, h_3, h_1)$

$l_{11} : \_chaplain\_n\_1(x)$

$l_2 : \_some\_q(y, h_4, h_2)$

$l_{21} : \_tomato\_n\_1(y)$

$l_{21} : \_plum\_a\_1(e_2, y)$

$l_3 : \_want\_v\_1(e, x, y)$

$h_3 =_q l_{11}$ $h_4 =_q l_{21}$

The composition of MRS semantics in feature structures is consistent with the principles outlined in Section 5.2.2.1 [Copestake et al., 2005]. We shall illustrate it using the sentence "Every chaplain probably ate" (see Figure 61). The global top GTOP labels of the daughters are identified to designate the derivation of a single logical form. Since both "every chaplain" and "probably ate" are scopal phrases, their HOOK is identified with the hook of the semantic head daughter (in this case, different from the syntactic head daughter): the scopal adverb "probably" in "probably ate" and the quantifier "every" in "every chaplain". Then the HOOK of "Every chaplain probably ate" is identified with the HOOK of the semantic head daughter which here is identical to the syntactic head daughter—"probably ate". Further, the RELS of the mother are obtained by appending the relations of the daughters and finally, the HCONS of the mother result from appending the HCONS of the daughters. Note that the final logical form is underspecified — its resolution happens within modules for scope determination external to the grammar.

After having introduced background information about representing metrical trees and underspecified semantics within typed feature structures, in the next section we put forth the gesture type hierarchy which guides the formalisation of the grammar rules from Section 5.3 into HPSG-based construction rules.

## 6.3 Gesture Type Hierarchy

We begin with Figure 62 which illustrates a fragment of the type hierarchy for gestures (for the complete hierarchy, refer to the implementation included in Appendix C). Whereas the traditional HPSG type hierarchy accounts for linguistic signs delivered through speech, our hierarchy distinguishes signs of spoken modality and signs of ges-

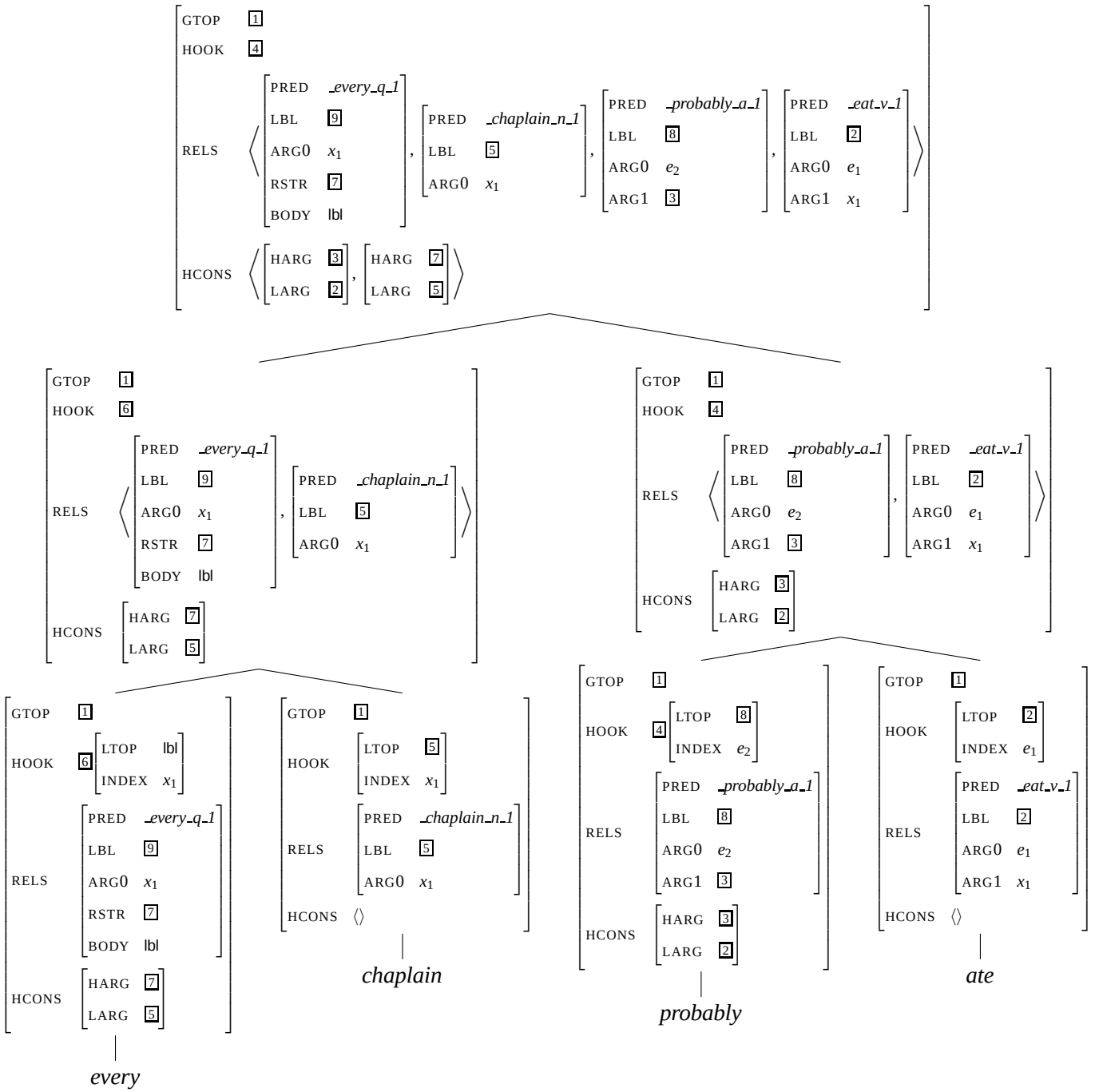



Figure 61: MRS Composition for “Every chaplain probably ate” in Feature Structures

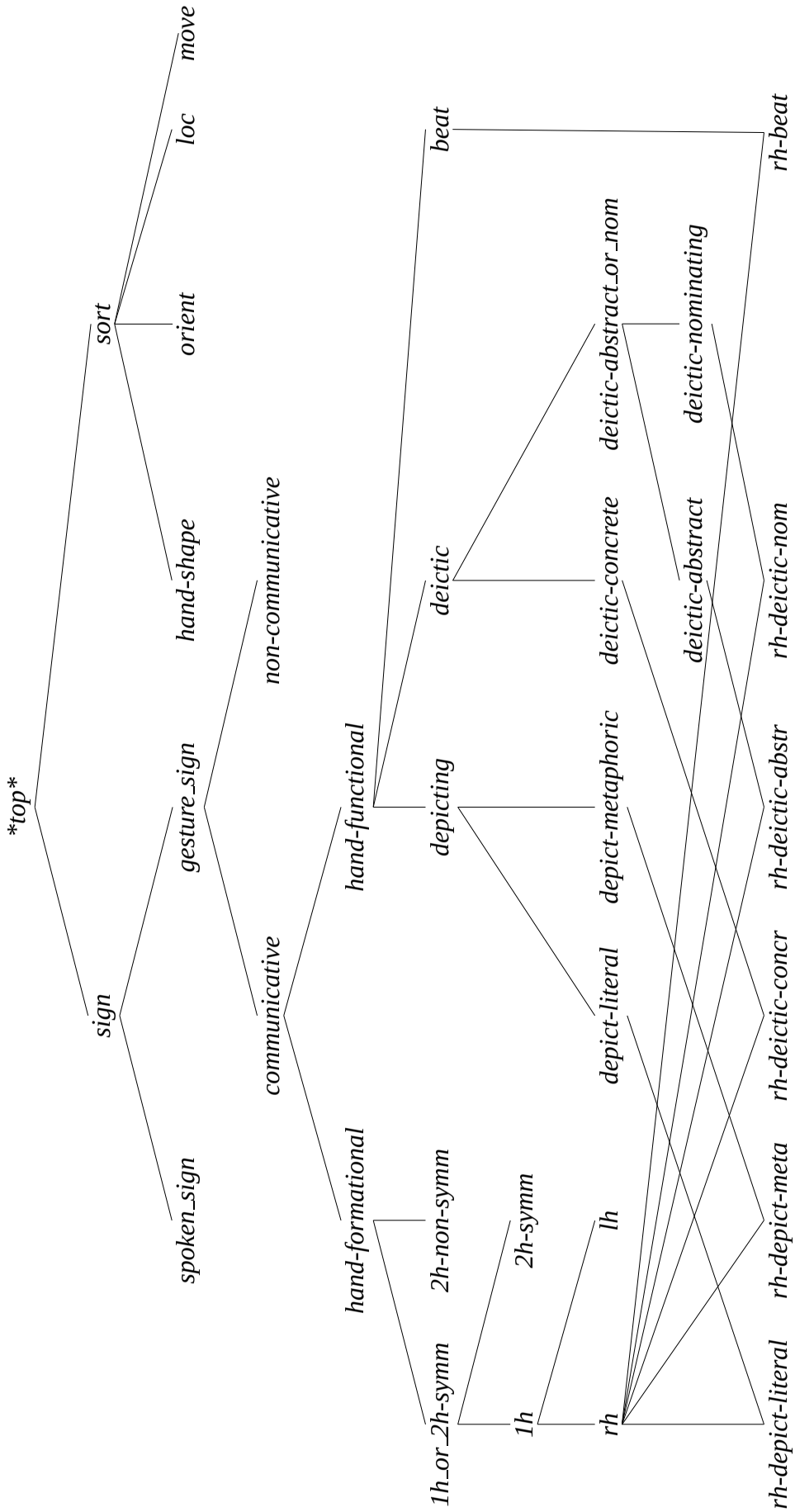



Figure 62: Fragment of the Gesture Type Hierarchy

tural modality: *spoken_sign* and *gesture_sign* are subtypes of *sign*. This distinction is necessary as it allows us to encode the features appropriate for gesture signs without interpolating them with those of spoken signs. For instance, while a feature such as INFLECTED is appropriate for spoken signs, a feature such as HAND-SHAPE is relevant for gestures. A further advantage of this hierarchy is that it can easily scale up to other visual and/or haptic modalities of communication.

The gesture sign is divided into *communicative* and *non-communicative* signals so as to distinguish between movements that are communicatively intended and thus contribute to the final interpretation of the utterance from movements that have no communicative effects. This binary classification was also incorporated in the annotation schema, discussed in Section 4.1. This thesis studies only the communicatively intended gestures, and we therefore do not enrich the hierarchy with any subtypes of non-communicative signs. For the sake of completeness, we still take care to encode a supertype for non-communicative gesture signs, and hence to allow for extending the type hierarchy with the appropriate subtypes by researchers studying these bodily actions (Lausberg and Sloetjes [2009] among others). Gestures of type *communicative* are further subdivided based on criteria of form (*hand-formational*) and of function (*hand-functional*). The former introduces the form features of the hand(s) depending on the articulator exploited by the speaker—one hand or two symmetrical hands (type *1h_or_2h-symm*), or two non-symmetrical hands (type *2h-non-symm*). The distinction between one-handed gestures or two-handed symmetrical gestures on one hand, and two-handed non-symmetrical gestures on the other allows us to encode the different feature-value pairs introduced by the gesture: for instance, while both R-HAND-SHAPE and L-HAND-SHAPE are appropriate for a two-handed non-symmetrical gesture, we need only HAND-SHAPE for a gesture performed by one hand or by two symmetrical hands. In case of bi-handed symmetrical gestures, there is no need to specify separate values for each hand since the values would be identical.

Following the notational scheme proposed by Bressem [2008], we use the "four-feature scheme" to capture gesture form. Within this scheme, gesture is described in terms of the four parameters: hand shape, orientation, location and movement. Given that the palm orientation and the finger orientation within a single movement are not necessarily identical (e.g., recall utterance (5.23) and its TFS representation in Figure 42), we shall differentiate between palm orientation and finger orientation. For one-handed or two-handed symmetrical gestures inherited from the type *1h_or_2h-symm*, we make appropriate the features HAND-SHAPE, PALM-ORIENT(ation), FINGER-ORIENT(ation),

$$\begin{bmatrix} \textit{1h\_or\_2h-symm} & \\ \text{HAND-SHAPE} & \langle \text{hand-shape} \rangle \\ \text{PALM-ORIENT} & \langle \text{orient} \rangle \\ \text{FINGER-ORIENT} & \langle \text{orient} \rangle \\ \text{HAND-LOCATION} & \langle \text{loc} \rangle \\ \text{HAND-MOVEMENT} & \langle \text{move} \rangle \end{bmatrix}$$

Figure 63: Form Features Appropriate to One-Handed and Bi-Handed Symmetrical Gesture

$$\begin{bmatrix} \textit{2h-non-symm} & \\ \text{R-HAND-SHAPE} & \langle \text{hand-shape} \rangle \\ \text{L-HAND-SHAPE} & \langle \text{hand-shape} \rangle \\ \text{R-PALM-ORIENT} & \langle \text{orient} \rangle \\ \text{L-PALM-ORIENT} & \langle \text{orient} \rangle \\ \text{R-FINGER-ORIENT} & \langle \text{orient} \rangle \\ \text{L-FINGER-ORIENT} & \langle \text{orient} \rangle \\ \text{R-HAND-LOCATION} & \langle \text{loc} \rangle \\ \text{L-HAND-LOCATION} & \langle \text{loc} \rangle \\ \text{R-HAND-MOVEMENT} & \langle \text{move} \rangle \\ \text{L-HAND-MOVEMENT} & \langle \text{move} \rangle \end{bmatrix}$$

Figure 64: Form Features Appropriate to Bi-Handed Non-Symmetrical Gesture

HAND-LOCATION and HAND-MOVEMENT, as shown in Figure 63. In contrast, for bi-handed non-symmetrical gestures of type *2h-non-symm* we introduce separate features for the right hand (R) and for the left hand (L), as illustrated in Figure 64. Notice that the value for each feature is specified as a list so as to capture the dynamic change of the shape of the hand. As it is standard in ERG, open-ended lists are represented by angle brackets. For instance,

$$\begin{bmatrix} \text{HAND-SHAPE} & \langle \text{fist, flat-hand} \rangle \end{bmatrix}$$

denotes a hand shape that changed from a fist to open flat or from open flat to a fist. Recall that we do not capture the timing of the dynamic change: the list values are related through conjunction and so the order of their appearance has no effects (see Section 5.3). Furthermore, type *1h* is inherited by gestures exploiting the right hand (*rh*) or the left hand (*lh*).

The functional criteria (type *hand-functional*) capture the function of the gesturing hand. It could be: (a) to depict an object (type *depicting*) where the depicting can be *depict-literal* or *depict-metaphoric*; (b) to point at a landmark (type *deictic*), where the pointing can be *deictic-concrete*, *deictic-abstract* or *deictic-nominating*; or (c) to emphasise some segments of the speech by beating along its rhythm (type *beat*). We

introduce the type *deictic-abstract_or_nom* to account for the different behaviour of concrete deixis vs. abstract/nominating deixis. We will further use this type when implementing the grammar rules in a grammar engineering platform (see Definition 5.3.3 and Chapter 7). Finally, the terminal nodes in the gesture hierarchy are a combination of both the formational and functional categories. For the sake of space, we have illustrated only the gestures performed by the right-hand: a literally depicting gesture performed by the right hand (type *rh-depict-literal*), a metaphorically depicting gesture performed by the right hand (type *rh-depict-meta*), an abstract deictic gesture performed by the right hand (type *rh-deictic-abstr*), a concrete deictic gesture performed by the right hand (type *rh-deictic-concr*), a nominating deictic gesture performed by the right hand (type *rh-deictic-nom*), and a beating gesture performed by the right hand (type *rh-beat*). Due to spatial constraints, the gesture type hierarchy here does not demonstrate the gestural multidimensionality. This, however, is also possible: for instance, the types *depict-literal* and *depict-metaphoric* could be also inherited by a type *depict-literal-meta* which then could be inherited by, say, *rh-depict-literal-meta* if performed by the right hand.

The values of the features appropriate to gesture are encoded within the type *sort* in the type hierarchy. The subtype hierarchies for the types *hand-shape*, *orient*(ation), *loc*(ation) and *move*(ment) are largely based on the notating scheme of Bressem [2008]. This has several advantages. First, it avoids conventionalised descriptions such as "finger ring" or "claw". Second, this scheme offers the right balance between abstraction and granularity: for instance, while the hand shape can be assigned 20 different types in an ASL-based notation (see McNeill [1992, pp. 87–88]), Bressem [2008] captures the same information in a more robust way by means of 3 basic types and 2 additional parameters. Finally, this scheme is suitable for a hierarchical organisation which could be particularly useful for coding with some supertype a hand movement that is not well pronounced and easily discernible.

We start off with Figure 65 which illustrates the possible values for *hand-shape*. The possible values for the hand shape are *fist*—the hand is closed in fist while executing the stroke; *flat-hand*—the hand is open flat; *finger*—the stroke is performed by a single finger (subtype *single-finger*) or by a combination of two or more fingers (subtype *finger-combination*). Gestures performed by the finger(s) are further specified in terms of the fingers that were engaged in the hand movement and also their form. For this reason, the type *finger* introduces the features FINGER-DIGIT and FINGER-FORM. The value of the former is of type *finger-digit* and it determines the finger(s)

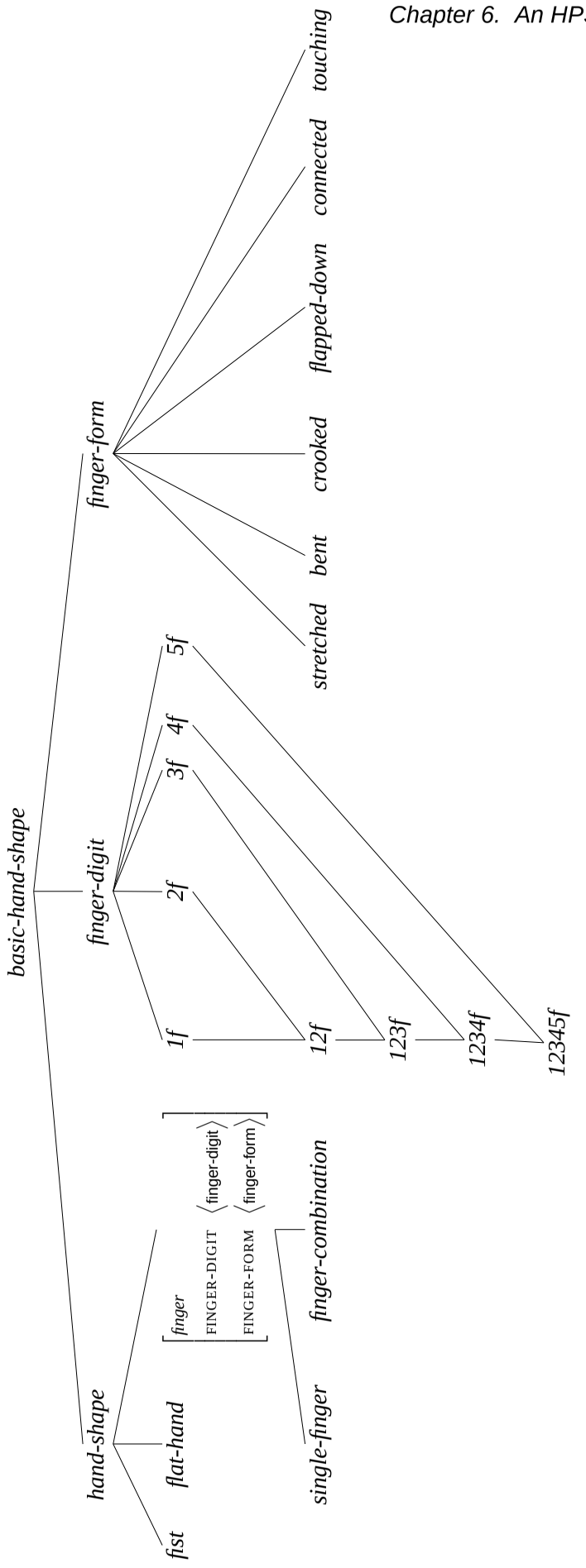



Figure 65: Fragment of the Sort Hierarchy of *hand-shape* based on Bressem [2008]

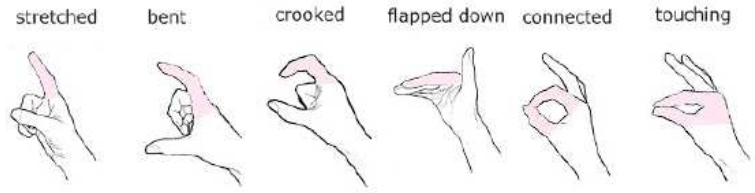



Figure 66: Finger Form Values [Bressem, 2008]

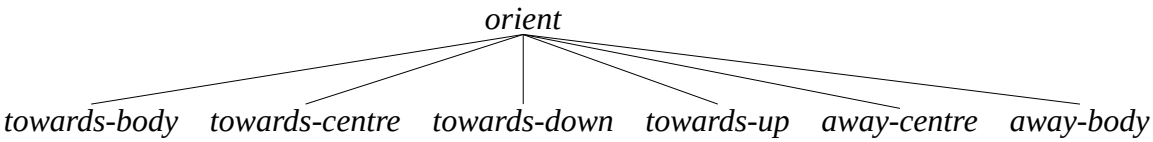



Figure 67: Sort Hierarchy of *orient* based on McNeill [2005]

used in the hand movement. The specific values correspond to each finger where *1f* corresponds to the thumb and *5f* to the pinkie finger. For the sake of space, we have illustrated only those hand shapes that involve the thumb and adjacent fingers including the thumb (ranging from *1f* through *12345f*). In analogy, finger combinations that do not include the thumb and/or adjacent fingers would be defined as *13f, 14f, 15f, 23f, 24f, 25f, 34f, 35f, 45f, 124f, 125f, 234f, 235f, 345f* and *2345f*. Finally, FINGER-FORM is of type *finger-form* whose possible values have been illustrated in Figure 66. Similarly as before, these values are defined as a list to account for a change in the form of the finger.

The values for the hand type *orient*(ation) are based on the typology of McNeill [2005] (see Figure 67). These properties are both appropriate for the orientation of the palm (provided through the attribute PALM-ORIENT) and the orientation of the fingers (provided by the attribute FINGER-ORIENT). The specific properties consider the orientation in relation to the torso; for instance, a feature value pair such as

$$\left[\text{PALM-ORIENT} \quad \left\langle \text{towards-body} \right\rangle\right]$$

describes a hand where the palm is vertically open held in parallel to the torso.

The type hierarchy *loc* for the location of the hand is based on the gesture space proposed by McNeill [1992] (see Figure 68) where:

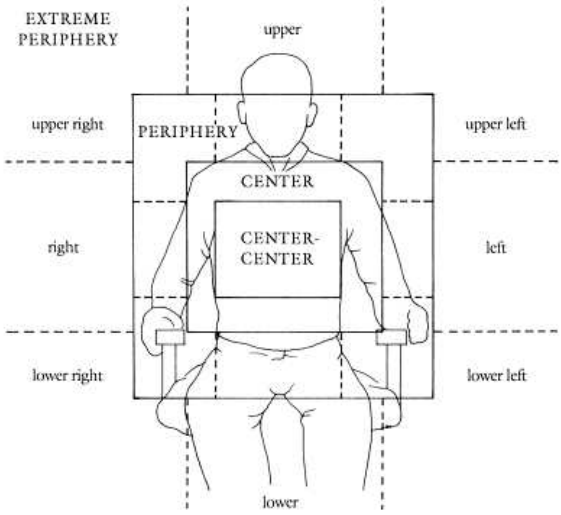



Figure 68: Gesture Space [McNeill, 1992]

> "...the gesture space can be divided into sectors using a system of concentric squares" McNeill [1992, p. 89]

The basic types in the type hierarchy, displayed in Figure 69, are $\vec{c}$, *centre*, *x-dim* and *y-dim*. Recall from Section 2.2.1.2 that following Lascarides and Stone [2009b], deictic gestures are represented in terms of the quantitative value $\vec{c}$ which designates the tip of the index finger and which combines with the rest of the deixis feature-values to identify the region pointed out by the deictic hand. The location *centre* includes the area along the extension of the speaker's torso. Then *x-dim* and *y-dim* refer to the areas along the x and y dimensions where the x-dimension distinguishes between the types *left* and *right*, and the y dimension distinguishes between *lower* and *upper*. Each of these categories is further subdivided into types that identify the peripheries close to the centre (these are *left-periphery, right-periphery, lower-periphery* and *upper-periphery*) and the peripheries further from the centre (these are *extreme-left, extreme-right, extreme-low* and *extreme-up*). The terminal nodes here are a combination along the x and y dimensions. Due to space limitations, we have illustrated only the values built from the combination of type *left-periphery* and each subtype of the y-dimension. The resulting values are *left-low, left-extreme-low, left-up* and *left-extreme-up*. The rest of the terminal nodes are built analogously.

The final subhierarchy *move* is based on Bressem [2008]. To capture the complexity of the movement, Bressem [2008] analyses the movement in terms of the three

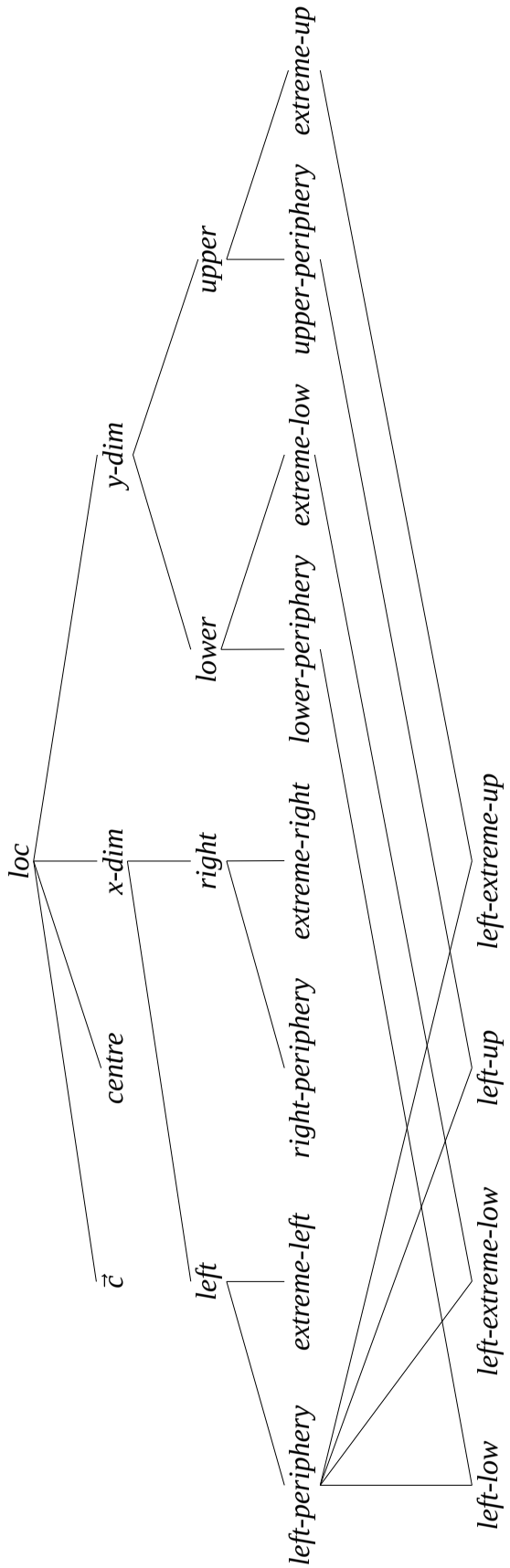



Figure 69: Fragment of the Sort Hierarchy of *loc*

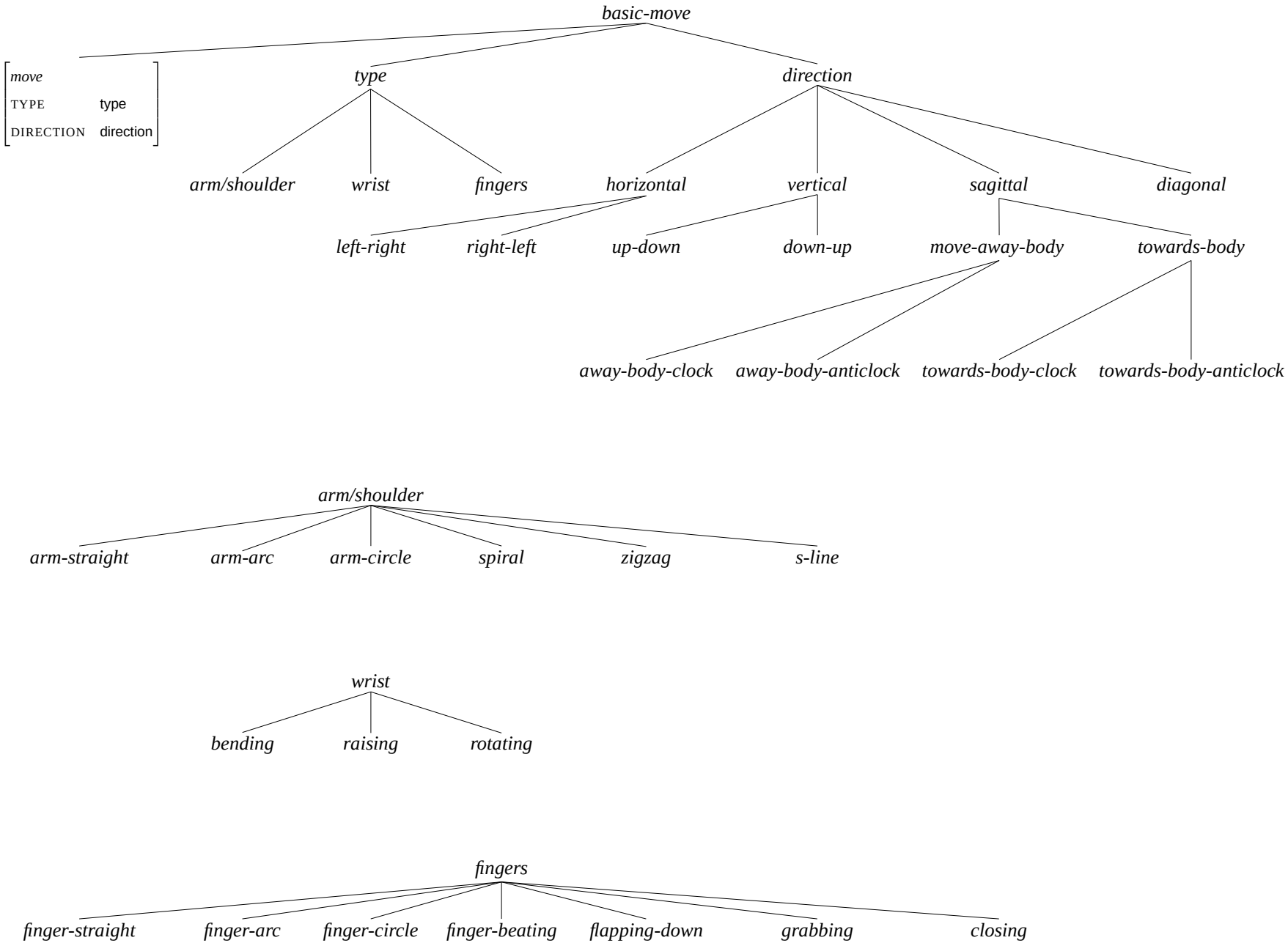



Figure 70: Sort Hierarchy of *move* based on Bressem [2008]

parameters type, direction and character. The movement type designates the line executed by the dominant articulator: the arm and/or shoulder, the wrist or the finger(s). The parameter direction is used to designate the trajectory employed by the hand along the x and y dimension and the sagittal plane. The last parameter character is a rather subjective judgement as to how rapid/slow or accentuated the movement was. Due to its infrequency, we do not encode the movement character. The hierarchical organisation of the movement is displayed in Figure 70. The type *move* is complex: it introduces a feature structure with TYPE and DIRECTION attributes. The movement type is subdivided depending on which was the predominant articulator executing the gesture: *arm/shoulder*, *wrist* or *fingers*. The possible type values for the movement performed by the arm/shoulder include *arm-straight, arm-arc, arm-circle, spiral, zigzag* or *s-line*. Then the movement types performed by the wrist could be *bending, raising* or *rotating*. Finally, the movement executed by the finger(s) could be straight (type *finger-straight*), resembling an arc (type *finger-arc*), circular (type *finger-circle*), beating in the air (type *finger-beating*), flapping down (type *flapping-down*), grabbing (type *grabbing*) or closing (type *closing*).

The direction of the movement is split on the horizontal axis (type *horizontal*), the vertical axis (type *vertical*), the sagittal plane (type *sagittal*) or the diagonal (type *diagonal*). The direction of a horizontal movement could be from the left periphery to the right periphery (type *left-right*) or from the right periphery to the left periphery (type *right-left*). A vertically performed gesture could be either in an upward direction (type *down-up*) or in a downward direction (type *up-down*). Then on the sagittal plane the gesture can be executed towards the body or away from the body in an either clockwise or anticlockwise direction. To capture these variations, we have introduced the subtypes *away-body-clock, away-body-anticlock, towards-body-clock* and *towards-body-anticlock*.

## 6.4 Formalisation of Well-formedness Constraints

Based on the empirical evidence that the performance of the gesture stroke must overlap with a nuclear/pre-nuclear speech elements or with a syntactic (or prosodic) constituent, we proposed definitions of the speech-and-gesture alignment in Section 5.3. Consistent with our goal of implementing a constraint-based multimodal grammar, we now formalise these rules into an HPSG-style notation.

### 6.4.1 HPSG-based Analysis of Prosodic Word and Gesture Alignment

Our HPSG-based analysis begins with the straightforward case of attaching gesture to a single word. The construction rule constrains the word to be nuclear or pre-nuclear accented, and also that the temporal performance of the gesture overlaps the temporal performance of the word. Considering the different semantic contribution of depicting gestures and of deictic gestures, we proceed by formalising them in distinct construction rules.

#### 6.4.1.1 Construction Rule for Aligning a Prosodic Word and Depicting Gesture

The feature structure representation of the rule aligning depicting gesture to a single prosodically prominent word (per Definition 5.3.1) is illustrated in Figure 71. We shall now describe each aspect of this feature structure in turn.

This constraint accounts for a sign of type *word* derived via unification of a single prosodic word of type *word* and a gesture of type *depicting* (or any of its subtypes). Along the lines of Johnston [1998a; 1998b] where explicit temporal constraints are defined within the rule schemata, we use the feature TIME to encode the constraint coming from the relative timing of the spoken and gestural modalities: there must be a temporal overlap between the performance of the gesture stroke (G) and the spoken word (S), that is, $end(G) > start(S)$ and $end(S) > start(G)$. Otherwise, the multimodal signal would be ill-formed. Also, recall from Table 19 that the temporal overlap does not mean only strict identity but it also includes cases where the gesture starts and/or ends at midpoint of the spoken word. We also record the overall duration of the multimodal word which is handled via the functional constraints earlier(7, 10) and later(8, 11). As discussed in Section 5.3.1, recording the start and the end of the multimodal element is necessary for its further integration with other elements, they being unimodal (speech or gesture) or multimodal.

For the gesture daughter (G-DTR), we record its temporal performance, its syntactic and semantic contribution. The syntactic information, encoded within the SYNSEM | CAT attribute, specifies the values of the gesture form-features defined in Figure 63 and in Figure 64. For the sake of space, we use

$$\begin{bmatrix} \text{G-FEATURE} & \text{value} \\ \dots & \end{bmatrix}$$

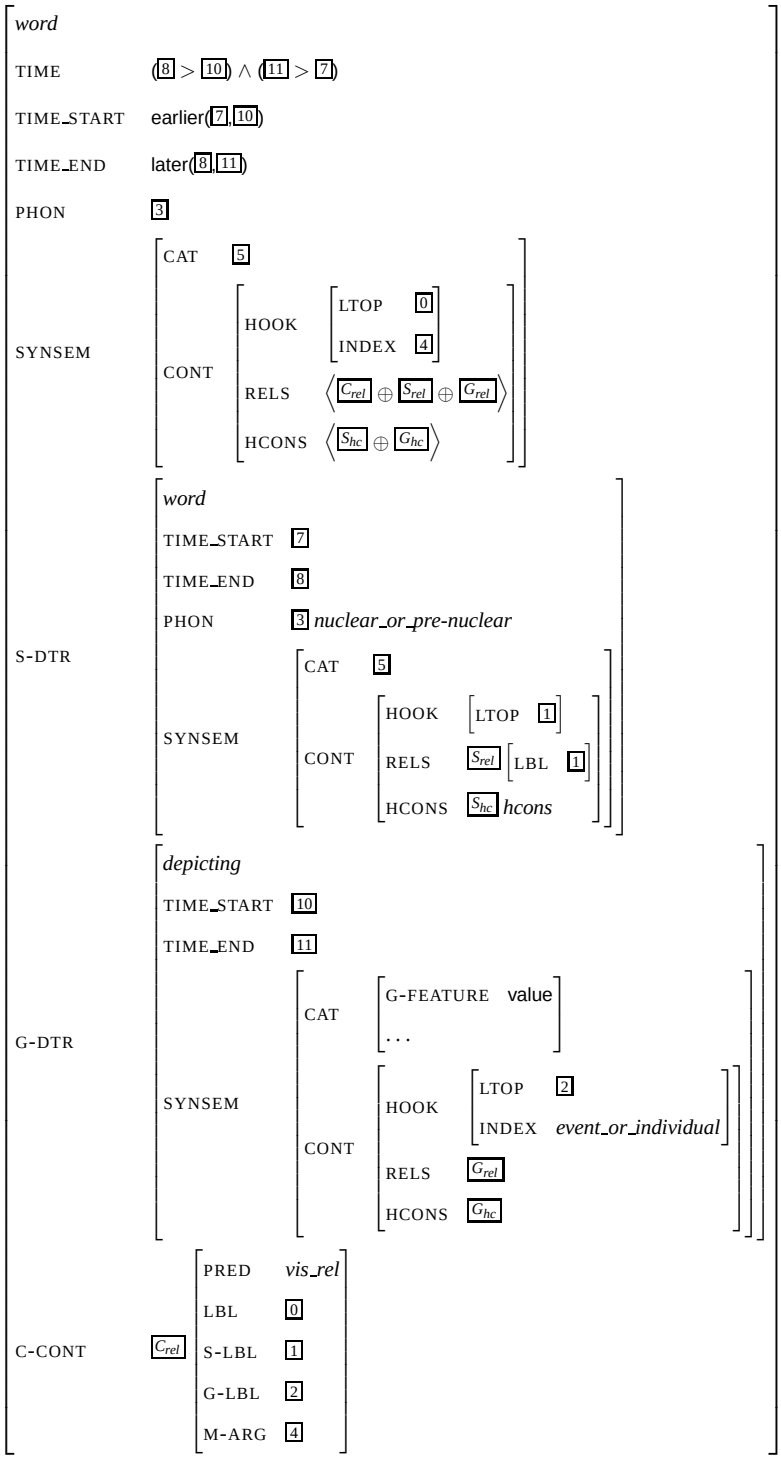



Figure 71: HPSG-based formalisation of the Situated Prosodic Word Constraint aligning depicting gesture and a spoken word

as an abstract notation for the distinct feature-value pairs whose exact number depends on whether the gesture is one-handed/bi-handed symmetrical or bi-handed non-

symmetrical.

For the speech daughter (S-DTR), it is equally important to record its timing, syntax-semantics information and also its prosody. The alignment between a depicting gesture and a spoken word constrains the latter to be a prosodically marked word whose PHON value is *nuclear_or_pre-nuclear*: that is, we restrict the combination of a prosodically unmarked word of type *unmarked* or a non-nuclear marked word of type *non-nuclear* and a gesture. Note that we do not constrain the head to a particular syntactic category: for instance, the gesture can align with a verb such as "mixes" as in "mixes mud", a noun such as "king" as in "king of Scotland", a preposition such as "through" as in "through the drainpipe" or an adjective such as "close" as in "close to the station" as long as the PHON and the TIME conditions are met. The underspecified semantic component of the speech daughter is defined in the familiar fashion in terms of its HOOK, RELS and HCONS features. The rule schema remains as unspecific as possible with respect to its RELS.

The construction rule contributes the underspecified semantic relation *vis_rel* between the speech content and the gesture content. Following the principles of the ERG grammar, any semantics contributed by a rule is encoded within the C-CONT attribute in parallel with SYNSEM. The relation *vis_rel* holds between the topmost label LTOP of the speech-daughter and the topmost label LTOP of the gesture daughter. This is formalised by identifying S-LBL of *vis_rel* with LTOP of the speech content (1) and G-LBL of the relation with LTOP of the gesture content (2). Based on Lascarides and Stone [2009b], *vis_rel* is an underspecified predicate in the MRS that abstracts over the possible rhetorical relations between gesture and speech (e.g., Narration, Depiction or Elaboration, but not Contrast). The resolution of this underspecified relation to a particular value happens outwith the grammar via discourse update, i.e., the process of constructing specific logical forms using the semantic values produced by the grammar and extra-linguistic information such as world knowledge and the mental state of the participants. Finally, as previously discussed in Section 5.3, the construction rule introduces an M-ARG attribute which serves as a pointer to the integrated multimodal signal.

The derivation of the mother node follows the algebra for semantic composition in constraint-based grammars [Copestake, Lascarides, and Flickinger, 2001]. It is strictly compositional: we unify the 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 se-

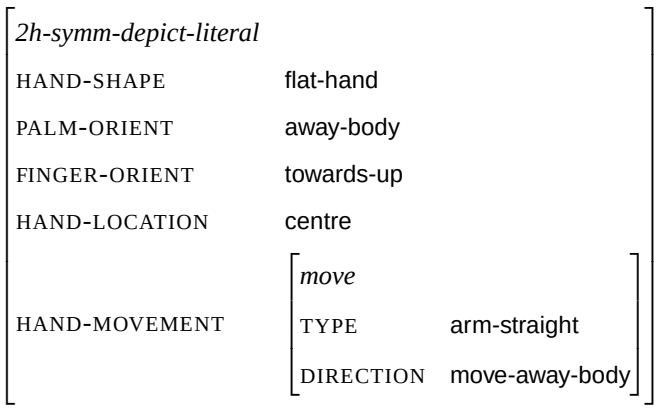



Figure 72: TFS representation of the depicting gesture in (6.2)

mantic representation involves appending the RELS of the speech daughter to the RELS of the gesture daughter, which in turn are appended to the RELS contributed by the construction rule (expressed via the append operator $\oplus$). The HCONS of the mother are accumulated in a similar way from the HCONS of the speech daughter and the HCONS of the gesture daughter. Finally, the HOOK value is contributed by *vis_rel*: the LTOP is identified with its LBL, and the INDEX is identified with the M-ARG.

With this constraint in hand, we can derive a possible speech-gesture analysis for utterance (2.4), repeated in (6.2).

(6.2) The [$_{PN}$bottom] worked [$_{N}$fine] ...
*Both hands are rested on the knees. The speaker lifts them in the frontal space with palms almost facing forward, fingers extended and moves them rapidly to the left and right periphery.*

We begin by representing the form of the depicting gesture into a feature structure, as illustrated in Figure 72. We type the gesture as *2h-symm-depict-literal* to account for both the formational and functional criteria: this is a two-handed symmetrical gesture that literally depicts the bottom cupboards referred to in speech. The feature values are based on the type hierarchy in Figure 62.

Mapping gesture form to meaning (see Figure 73) follows the principles outlined in Section 5.2.2. Each feature value pair maps directly to a labelled semantic relation with an underspecified main argument. Then the modal operator $[\mathcal{G}]$ is added to the whole formula in an outscoping relation with the labels of the semantic relations mapped from the gesture form features. As already discussed, the LTOP of the gesture is identified

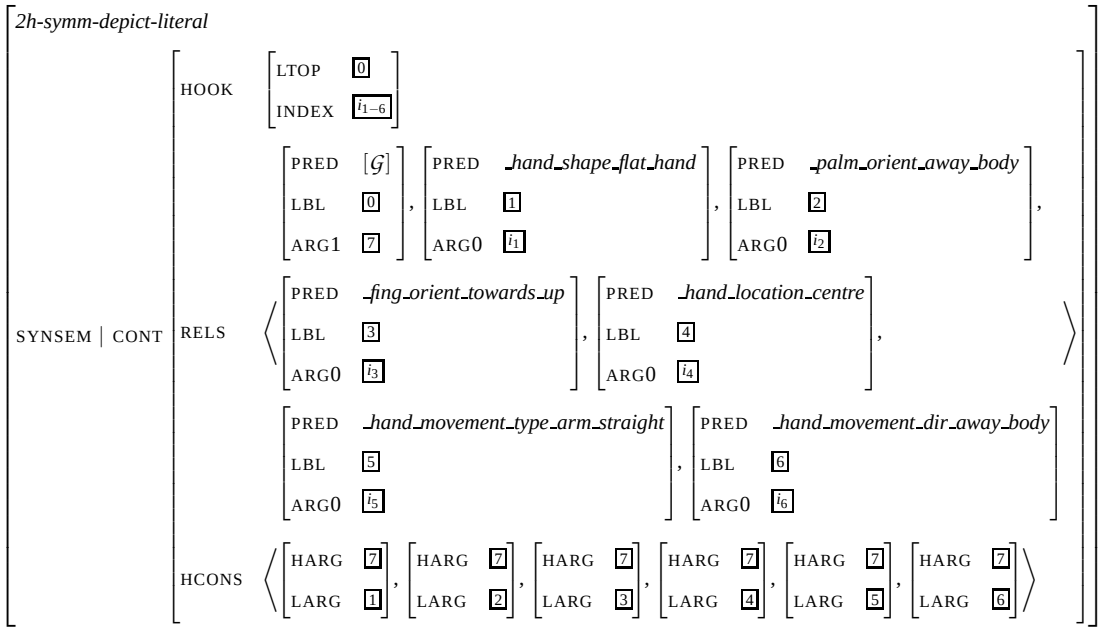



Figure 73: TFS-style MRS semantics mapped from the gesture form in Figure 72

with the LBL of [$\mathcal{G}$], and the INDEX is specified in terms of the abstract label $\boxed{i_{1-6}}$ that in context will resolve to one of the ARG0 of the EPs mapped from form.

The application of the Situated Prosodic Word Constraint to the multimodal action in (6.2) licenses a gesture attachment to the noun "bottom" (see the syntactic derivation and the semantic composition in Figure 74): it is a spoken word marked by a pre-nuclear accent of type *pre-nuclear* and the extension of its temporal performance overlaps the extension of the temporal performance of the depicting gesture. The form-meaning mapping of the gesture is consistent with the principles outlined above, so we forgo any further details. Due to space limitations, Figure 74 illustrates a single elementary predication mapped from gesture form: *_hand_shape_flat_hand*. The rest of the predications behave in the same way, as already shown in Figure 73. The SYNSEM | CAT information for the speech daughter demonstrates a syntactically unsaturated phrase expecting a specifier SPR. Its semantics contains the predication *_bottom_n_*1 contributed by the lexical entry, whose LBL and ARG0 are respectively identified with the HOOK | LTOP and HOOK | INDEX of the phrase. We also introduce the prosodic type *pre-nuclear* to the entry, and its TIME_START and TIME_END values.

The alignment proceeds by first recording the temporal values of the multimodal word by comparing the individual TIME_START and TIME_END values of the daughters. Then the PHON and SYNSEM | CAT values of the speech head daughter are prop-

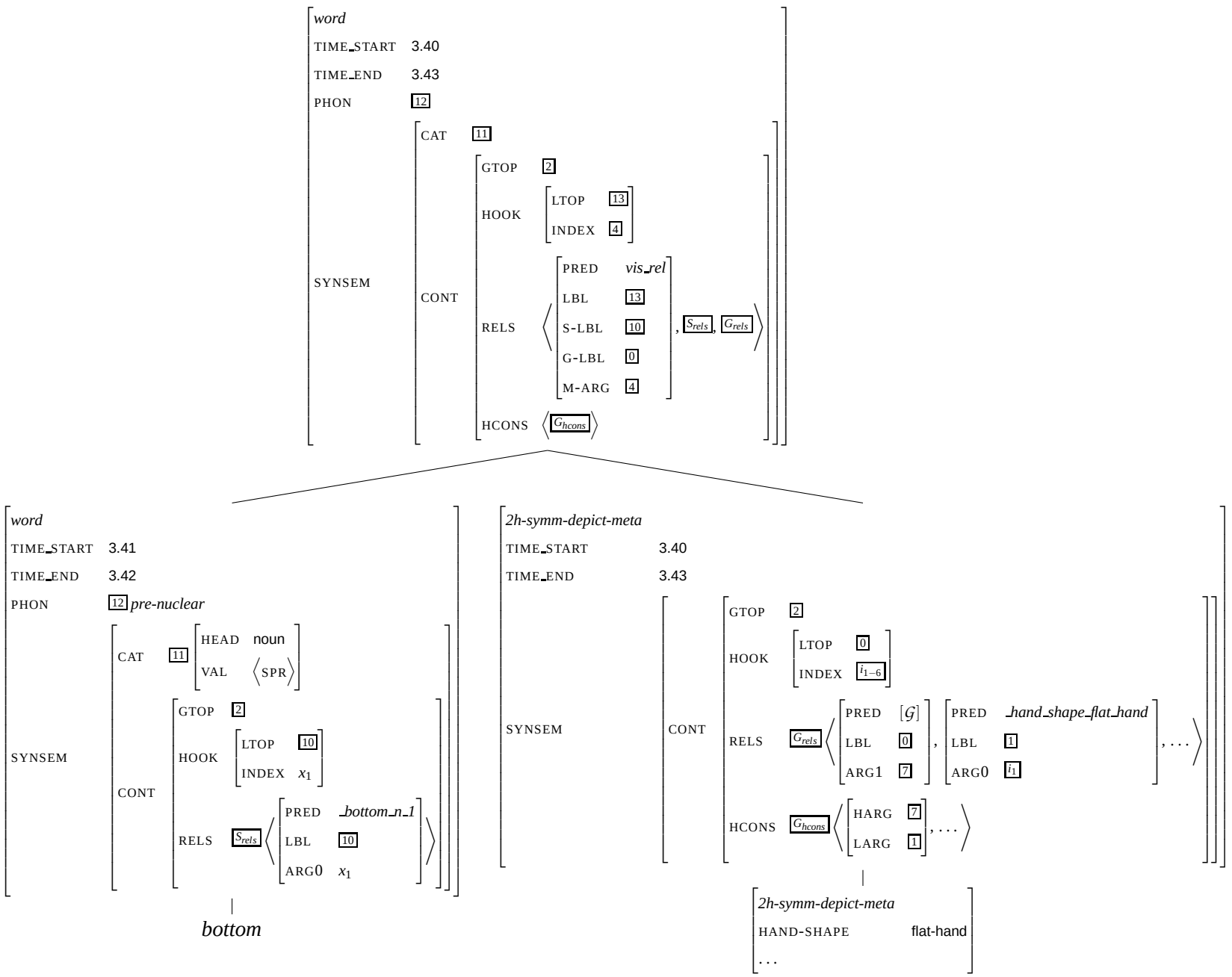



Figure 74: HPSG-based Syntactic Derivation and Semantic Composition for Depicting Gesture + “bottom”

agated up to the mother node. The semantic composition involves appending the underspecified relations of the gesture daughter to those of the speech daughter, which are appended in turn to the semantics contributed by the construction rule itself. The syntactic attachment involves establishing a *vis_rel* between the LTOP of the depicting gesture (co-indexed with LBL of the gesture main relation — the modal operator $[\mathcal{G}]$) and the LTOP of the speech daughter. In this case, the underspecified relation *vis_rel* can resolve in context to a literal depiction of the bottom cupboards.

Any further composition with the specifier "the" would proceed in the standard way: for instance, the ARG0 of the semantic predicate *_the_q* corresponding to the determiner "the" would equate with the semantic index M-ARG of the multimodal word (for comparison, the standard composition of unimodal elements would equate the ARG0 of the quantifier with ARG0 of *_bottom_n_1*).

#### 6.4.1.2 Construction Rule for Aligning a Prosodic Word and Deictic Gesture

Figure 75 illustrates the TFS formalisation of the alignment of a deictic gesture and a spoken word. Like depicting gesture, deictic gesture must temporally overlap the word that it aligns with. There are no further restrictions on the deictic type and so this rule can apply to any subtype inherited from *deictic*: e.g., abstract deixis, concrete deixis and nomination deixis performed by the left hand, right hand or both hands. The SYNSEM values of the deictic daughter (G-DTR) are encoded as detailed in Section 5.2.2.4: the CAT feature contains the list of deixis' appropriate attributes from Figure 63 and Figure 64, here glossed over as

$$\begin{bmatrix} \text{G-FEATURE} & \text{value} \\ \dots & \end{bmatrix}$$

and the CONT component is specified in the standard way in terms of HOOK, RELS and HCONS.

The speech daughter S-DTR is encoded as introduced above in terms of its temporal values TIME_START and TIME_END, its prosodic information PHON and also its syntax/semantics information SYNSEM. The PHON value is constrained to *nuclear_or_pre-nuclear*. We forgo any details about the syntactic category of the speech daughter since it does not constrain the integration.

Recall from Section 5.3 that the full inventory of the possible semantic relations between speech and deixis is captured via an underspecified semantic relation *deictic_rel* [Lascarides and Stone, 2009b]. The construction rule therefore introduces in C-CONT

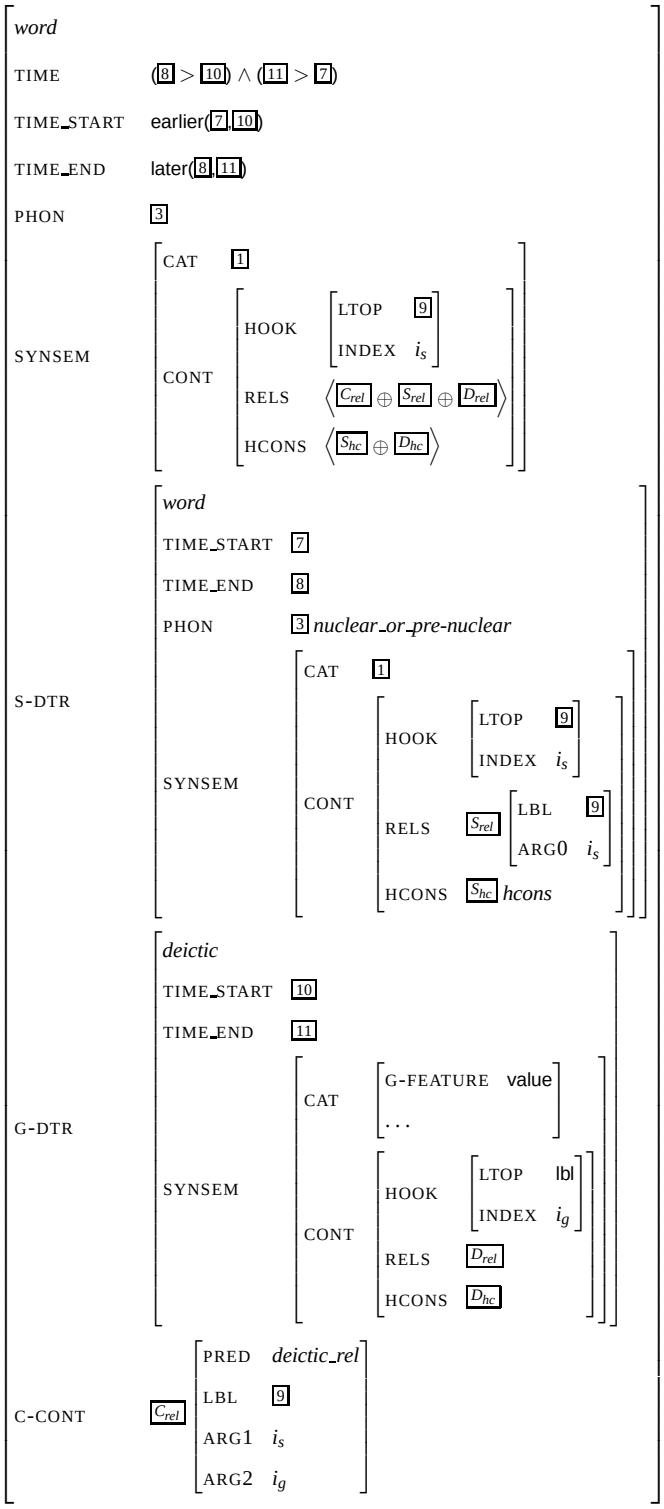



Figure 75: HPSG-based formalisation of the Situated Prosodic Word Constraint aligning deictic gesture and a spoken word

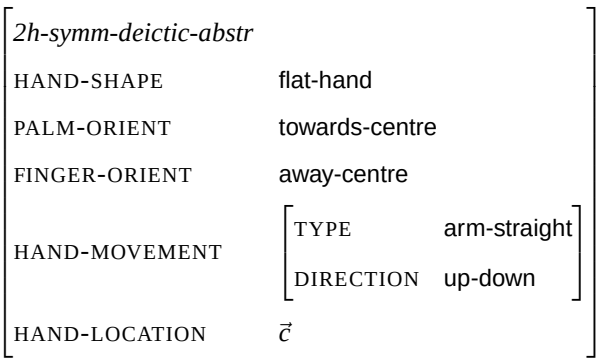



Figure 76: Refinement of the TFS representation of the deictic gesture in (6.3)

*deictic_rel* between the semantic index $i_g$ of the deictic gesture and the semantic index $i_s$ of the speech that captures the fact that the speech and deixis are aligned and thus coherently connected. The treatment of this relation is similar to that of appositives in ERG of the sort "the person, the one that I am pointing to" in that it shares the same label as the speech head daughter since it further restricts the individual/event introduced in speech. In so doing, any quantifier outscoping the head would also outscope this relation. The semantic composition of the mother node is strictly monotonic: it involves appending the relations of the speech daughter to the relations of the deictic daughter, which are then appended to the relation contributed by the rule (notated with $\oplus$). The mother's HCONS are also the result of appending the daughters' HCONS. Since the PHON feature is appropriate to the speech daughter, the PHON value of the mother is co-indexed with the one of the speech daughter. The SYNSEM | CAT value is also co-indexed with SYNSEM | CAT of the speech daughter. In so doing, we preserve the information about the valence and head requirements of the speech daughter.

After having put forth the theoretical apparatus, we can now turn to the hands-on application of this rule to a multimodal action of speech and deixis. As an example, we will be using utterance (5.11), repeated below.

(6.3) There's like a [$_{NN}$little] [$_{N}$hallway]
*Hands are loosely open, vertical, parallel to each other. The speaker moves them downwards.*

We have already illustrated the feature structure representation of this deictic gesture in Figure 39. Given the gesture type hierarchy proposed in Section 6.3, we need to refine this form representation accordingly (see Figure 76). We first amend the gesture

type to *2h-symm-deictic-abstr* to reflect both its formational and functional aspects: this is bi-handed gesture that places a landmark for a virtual hallway in the frontal space; the shape of both hands is symmetrical. We also refine the hand movement with its type and direction as follows: the movement is performed with the arms and the line of the motion is straight. This yields the information

$$\left[\text{TYPE} \quad \text{arm-straight}\right]$$

Then the direction of the movement is along the vertical axis from the upper to the lower centre, which we render as

$$\left[\text{DIRECTION} \quad \text{up-down}\right]$$

We can now map this gesture form to an underspecified semantics in a TFS notation, as shown in Figure 77. Following the principles from Section 5.2, each feature-value pair maps to an elementary predication that "modifies" the spatial area identified by the pointing hand. This is formalised by treating each feature-value pair as a predicate of an intersective modifier whose ARG1 corresponds to ARG0 of the deictic main relation *sp_ref*. Since this is a scopal phrase, the HOOK is identified with ARG0 and LBL values of the scope-bearing element, the quantifier *deictic_q*. As discussed in Section 5.2.2, the LTOP of the phrase is always distinct from the LBL of the scope-bearing element.

The application of the Situated Prosodic Word Constraint to utterance (6.3) would license an attachment to the noun "hallway" since it bears a nuclear accent and also since its temporal performance overlaps the temporal performance of the deictic gesture. Since the processing of the temporal, syntactic and phonological information does not differ from that of depicting gestures, Figure 78 illustrates only the semantic composition of the multimodal utterance. Note that in composition, the underspecified index $i_0$ introduced by the deictic gesture resolves to an individual $x_0$. We also establish an underspecified relation *deictic_rel* between the semantic index $x_1$ of the speech head daughter and the resolved semantic index of the deixis daughter $x_0$. Further, the composition of the situated utterance with the intersective modifier "little", and subsequently with the quantifier "a" proceeds in the standard way where the label of the modifier is shared with the one of the head noun, and hence also with the label of the deictic relation, and it also appears within the restriction of the quantifier.

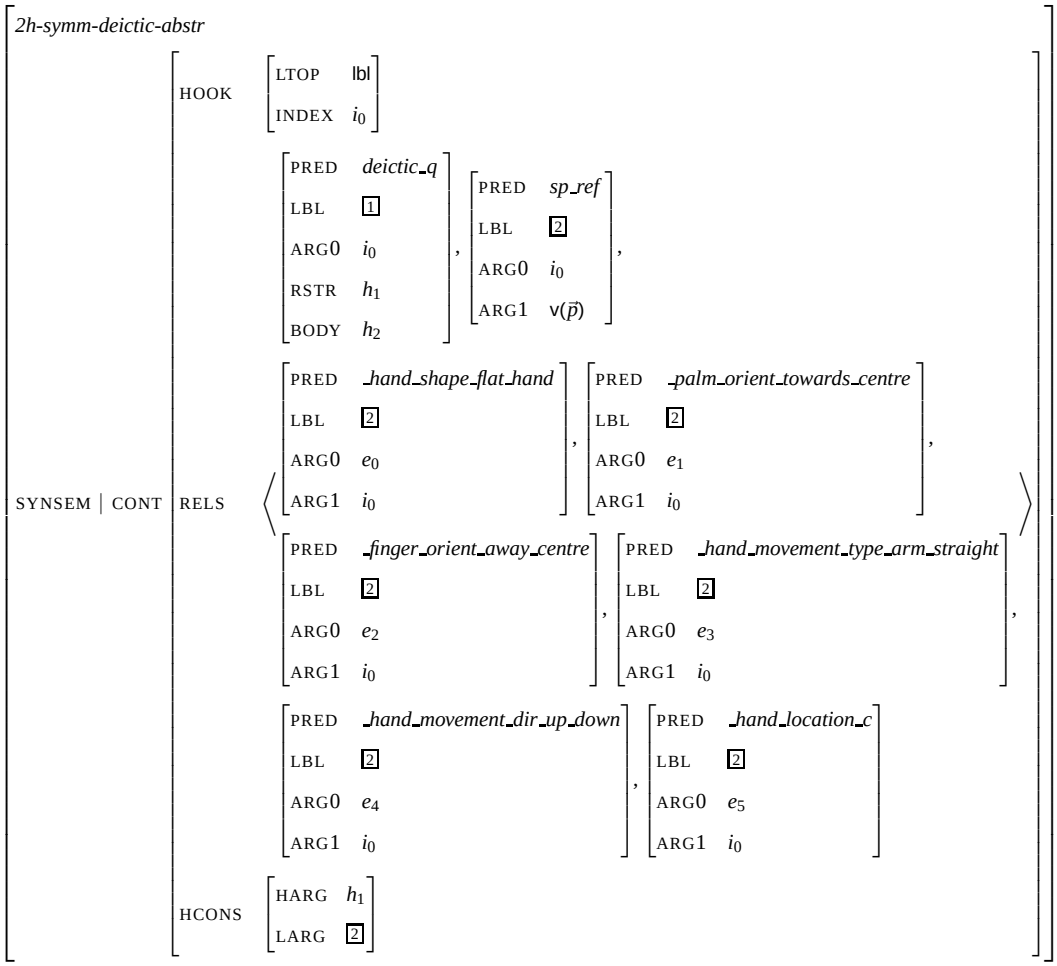



Figure 77: TFS-style MRS semantics mapped from the gesture form in Figure 76

#### 6.4.1.3 Construction Rule for Aligning a Prosodic Word and Deictic Gesture with Defeasible Constraint

In Section 4.3, we provided examples that demonstrated that a temporal and/or prosodic relaxation between the deictic gesture and the semantically related speech element is possible with deictic gestures of type *deictic-concrete* that identify individuals salient in the communicative event. This was accounted for by the rule in Definition 5.3.3, which we now formalise into an HPSG-based construction rule (see Figure 79). This rule does not constrain the temporal relation between the spoken word and the gesture stroke — we allow for precedence and for sequence relations between the spoken and the gestural modality (the overlap relation is also possible, and it was accounted for by the rule in Figure 75). Note that this rule overgenerates the possible analyses with respect to the temporal conditions: we cannot formally express the temporal overlap

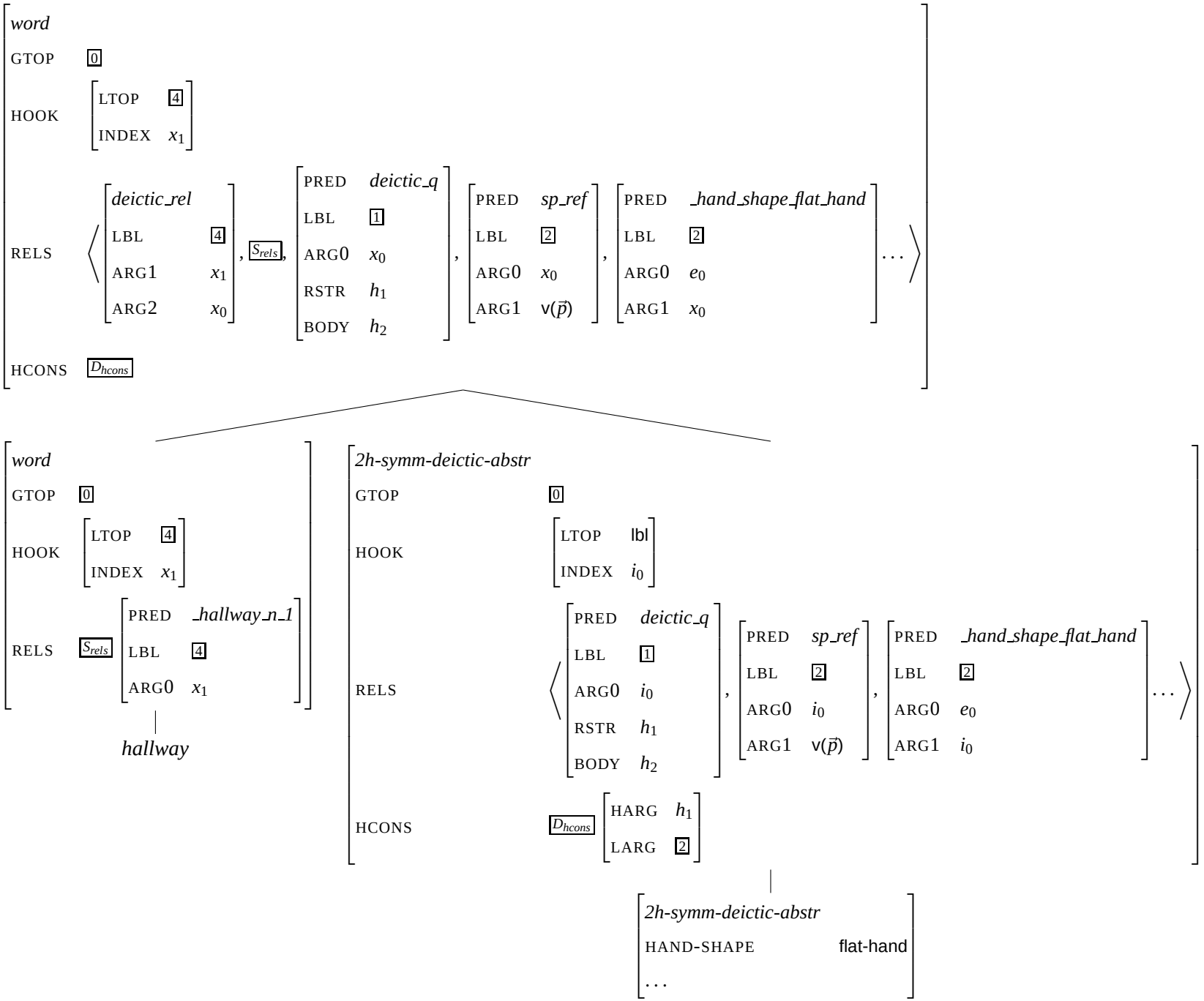



Figure 78: HPSG-based Semantic Composition for Deictic Gesture + “hallway”

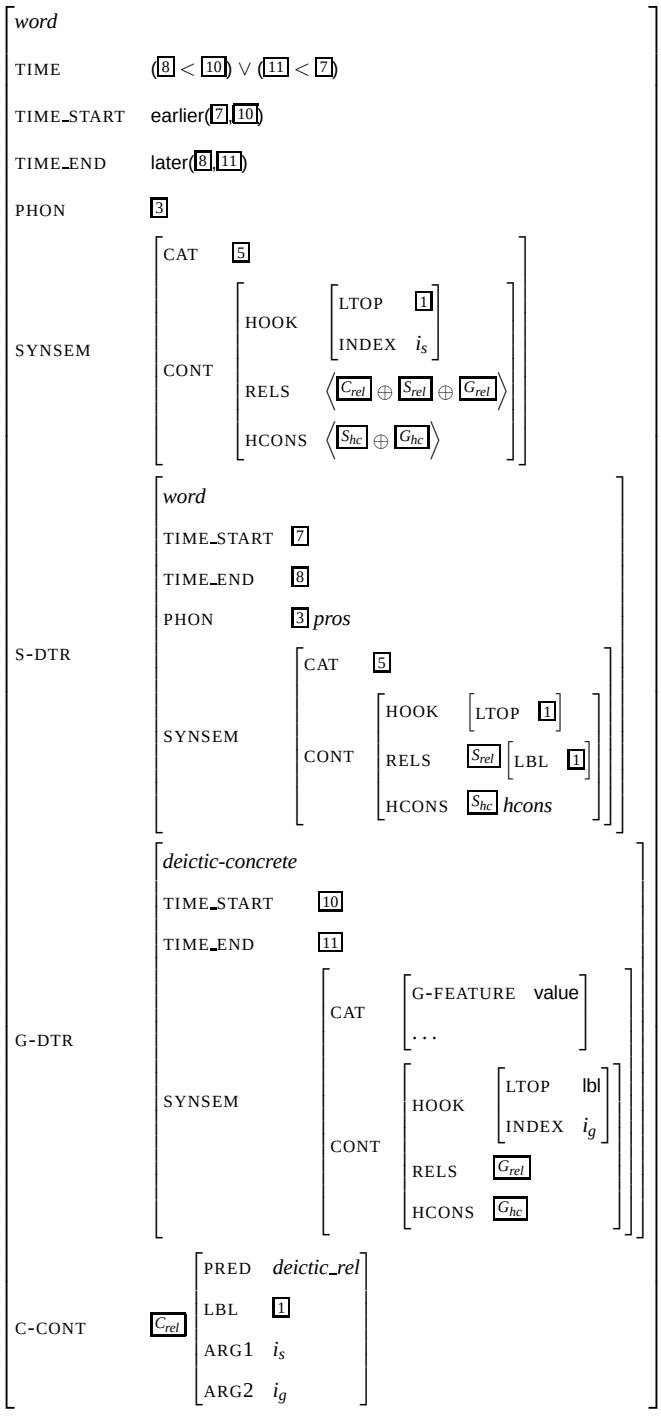



Figure 79: HPSG-based formalisation of the defeasible constraint aligning a spoken word and a concrete deictic gesture

between the gesture and (some portion of) the utterance (as per Definition 5.3.3) since at the point of attachment we need information about the timing of the mother, which

is not accessible yet. Further, the spoken word is not restricted to a particular prosodic type and we therefore use the basic prosodic label *pros* to allow for the alignment between concrete deixis and prosodically marked or unmarked words. Importantly, this rule constrains the alignment to gestures typed as *deictic-concrete* or any of its subtypes. We forgo any further details about the formalisation of this rule, since the rest remains the same as for deictic gestures (see Figure 75 and the corresponding discussion).

With this rule in hand, we can now provide an analysis for utterance (2.3), repeated below.

(6.4) 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.*

The rule from Figure 79 licenses a deixis attachment to the prosodically unmarked "she". Such syntactic attachment would support an interpretation at the semantics/pragmatics level where the referent introduced by the deictic gesture is in an Identity relation with the denotation of the spoken element (recall 5.8). This is the most intuitive interpretation of this multimodal action: namely, the referent identified by the pointing hand is identical to the referent introduced in speech. The lack of this rule, however, would block such an interpretation.

Also notice that this rule would not license an attachment of the deictic gesture from utterance (2.12), repeated in (6.5), to the prosodically unmarked "I".

(6.5) 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.*

Although the forms of the multimodal actions in (6.4) and (6.5) are to an extent comparable in that the temporal performance of the deictic gesture overlaps the temporal performance of the pre-nuclear/nuclear prominent word and not the prosodically unmarked pronoun, the deictic gesture in (6.5) is of type *deictic-abstract*, and so it can attach only the prosodically prominent element it overlaps with or any of its higher projections. This is important since no plausible interpretation can be derived from a syntactic configuration where the deixis in (6.5) aligns with the pronoun "I".

### 6.4.2 HPSG-based Analysis of Spoken Phrase and Gesture Alignment

As discussed in Section 5.3.2, we claim that the form-meaning mapping of an utterance of speech and hand gesture is one-to-many rather than one-to-one. In the grammar, we account for these multiple mappings via construction rules that license multiple speech-and-gesture attachments where each attachment supports potentially distinct possible interpretation(s) in context. While in the previous section we provided a formal analysis of the alignment of a gesture and a single word, in this section we provide a formal analysis of the alignment of a gesture and a constituent. Since the semantic contribution of the gestures remains the same as outlined in Section 6.4.1, we provide a single analysis for depicting and deictic gestures.

The HPSG-style formalisation of the rule from Section 5.3.2 is displayed in Figure 80. We shall now describe the specificities of this rule omitting thus any details that are identical to the Situated Prosodic Word Constraint (recall Figure 71 and Figure 75 and the associated discussions). The alignment of both modalities is constrained by their relative timings: the temporal performance of the speech phrase provided by its TIME_START and TIME_END values should overlap with the temporal performance of the gesture provided by its TIME_START and TIME_END values. The temporal overlap relation is essential as it enables attaching gesture to an entire constituent whose components may happen outside the performance of the gesture (recall (5.27)). Prosody also constrains the alignment: the PHON value of the speech daughter is restricted to type *mtr(*$\tau$*)*—i.e., a metrical tree of any depth [Klein, 2000b]. Recall from Section 6.2.1 that the domain union relation ($\bigcirc$) is used to interpolate the prosodically prominent element—the DTE—into the non-empty list of domain objects. Further, within the SYNSEM | CAT | VAL attribute we make use of the disjunction operator so as to remain as neutral as possible about the number of saturated arguments when the speech-and-gesture alignment takes place. This constraint allows one to attach a gesture to a headed phrase whose COMPS requirements have been fullfilled or to a headed phrase whose both SPR and COMPS requirements have been fullfilled.

It is also necessary to point out the distinct status of *vis_rel* contributed by the Situated Prosodic Word Constraint and the Situated Spoken Phrase Constraint: whereas the alignment of a depicting gesture and a single word contributes as little information as possible for resolving this underspecified relation, the alignment of the head of the phrase with its arguments contributes to its minimal specification and hence the

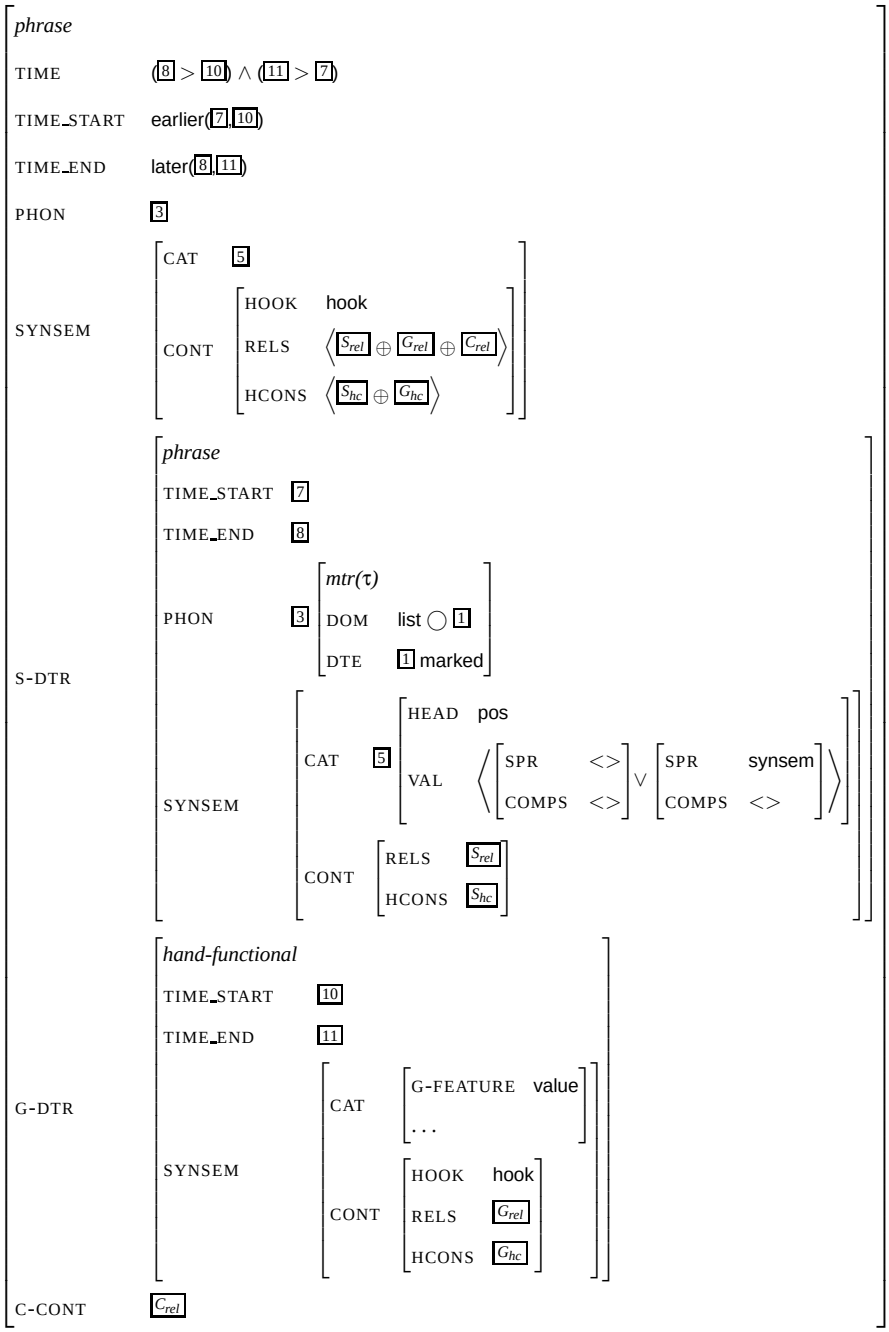



Figure 80: HPSG-based formalisation of the Situated Spoken Phrase Constraint aligning gesture and a constituent structure

choices of resolving this relation are more constrained. This of course is possible with the additional assumption that the interpretation of the speech-and-gesture relation is dependent on how the gesture signal resolves, as per Lascarides and Stone [2009b].

The application of this rule to utterance (6.2), which features a literally depicting

gesture, would allow for the following attachments: first, the gesture can be integrated with the NP “the bottom”, that is, upon combining the pre-nuclear item with its specifier. Another possible configuration involves attaching the gesture to the entire sentence “the bottom worked fine”, that is, upon combining the head daughter with its arguments. This attachment is licensed by the temporal overlap between the entire phrase, which is a metrical tree, and the depicting gesture. The high attachment to the root node of the tree would support a metaphorical interpretation where the gesture qualifies the speech act of completion of a process (recall Section 2.1).

Likewise, the derivation of the speech phrase and deictic gesture proceeds by attaching the gesture to each projection of the speech temporally overlapping the gesture performance. To illustrate this, consider again utterance (6.5). The configurations licensed by the schema in Figure 80 include the deixis attaching to the VP projection “enter my apartment” and also to the S “I enter my apartment”.

**Discrepancy between prosodic and syntactic constituency.** Until now, we discussed cases where the prosodic grouping was isomorphic to the syntactic grouping. The final challenge concerns those cases where the gesture aligns with a prosodic constituent that is not identical to the syntactic constituent, as previously shown in Figure 52. In this section, we propose a possible analysis of prosodic constituency in HPSG. This analysis, however, is not intended as a major revision of HPSG but rather as an ad hoc solution to an underlying issue.

A possible direction could to be assume an expanded notion of syntactic constituency as in Combinatory Categorial Grammar [Steedman, 2000] where there is no discrepancy between prosody and syntax, and functional application and type-raising are used for combining subjects and verbs.

By contrast, we will propose an analysis of structured phonology without disrupting the traditional notion of constituency in the grammar. To date, we are familiar with two main approaches that have looked at prosodic constituency in HPSG: Klein [2000a; 2000b] and Haji-Abdolhosseini [2003]. In Section 6.2.1, we introduced Klein’s [2000a; 2000b] model. A limitation of this approach is that it is *syntacto-centric*, that is, the derivation of the prosodic tree is driven by the derivation of the syntactic tree. This has been exemplified in Figure 81 where we have used the shortcut notation of Klein [2000a]: square brackets indicate a prosodic phrase of type *mtr(full)* and parentheses indicate a prosodic phrase of type *mtr(lnr)*. We consider that this model is not suitable for aligning a prosodic phrase (and its associated semantics) with

gesture since one arrives at the prosodic phrase of the type “Mary prefers”, “he found it” from Figure 81 only after building the complete derivation tree for the input phrase (or clause). The fact that we have no access to the prosodic phrase per se prevents us from attaching gesture to it.

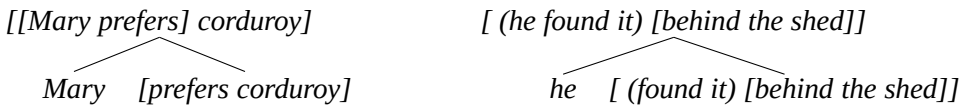



Figure 81: Syntax-driven Prosodic Grouping [Klein, 2000a]

To overcome this limitation, we propose a different architecture where the prosodic phrasing and the syntactic phrasing are constructed independently from each other over the same list of domain objects. Since we need the semantics at both the prosodic level and the syntactic level (assuming a mismatch between the two), we could construct the semantics monotonically at both the prosodic and the syntactic levels. This architecture has been inspired by Haji-Abdolhosseini [2003] who proposed the construction of the syntax/semantics, prosody and information structure over the same list of domain objects. However, while Haji-Abdolhosseini [2003] builds the prosodic, syntactic/semantic and information structure independently, we shall interface the prosodic component with the semantic component on the one hand, and the syntactic with the semantic component on the other, as it has been shown in Figure 82.

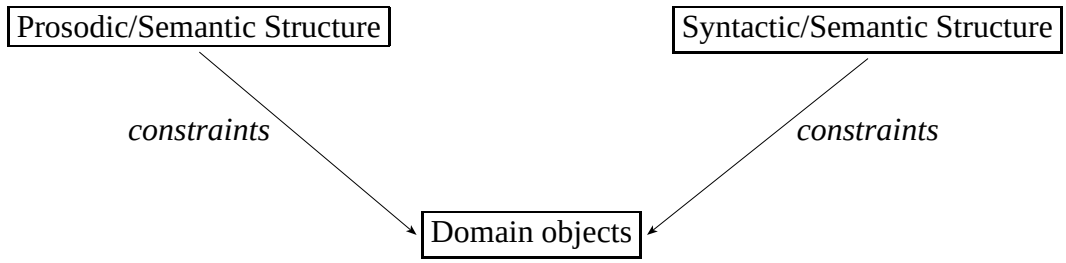



Figure 82: Architecture of the independent prosodic/semantic and syntactic/semantic derivation over the same list of domain objects

In comparison to Klein’s [2000a; 2000b] model, this architecture enables the derivation of the prosodic structure independently from the derivation of the syntactic structure under certain well-formedness constraints. In so doing, we can arrive at a prosodic constituent without having explored the syntactic constituent. Since we also intend to capture the semantic contribution of the gesture in relation to the semantic contribution of the speech phrase, as licensed by the alignment constraints, we construct the semantics of the prosodic phrasing. This separation mechanism would require a differ-

$$\begin{bmatrix} \textit{phrase} & \\ \text{PHON} & \text{pros} \\ \text{SYN} & \text{syn} \\ \text{SEM} & \text{sem} \end{bmatrix} = \begin{bmatrix} \textit{phrase} & \\ \text{PHON} & \text{pros} \\ \text{SEM} & \text{sem} \end{bmatrix} \cup \begin{bmatrix} \textit{phrase} & \\ \text{SYN} & \text{syn} \\ \text{SEM} & \text{sem} \end{bmatrix}$$

Figure 83: Unification of feature structure in the proposed modular architecture

ent approach to linking the syntactic/semantic and the prosodic/semantic components. Along the lines of Gardent [2008], a possible direction would be to assume an interface dimension that unifies the separate components, i.e., the prosodic, syntactic and semantic feature-value pairs as shown in Figure 83. While it could be argued that the semantic information gets duplicated, the semantics of the mother is the product of unifying the semantics of the prosodic phrase and the semantics of the syntactic phrase: since the values are the same, there is no unification failure.

With this machinery, we now propose the HPSG-based Situated Prosodic Phrase Constraint that attaches a gesture to a prosodic phrase, distinct from the syntactic phrase. The formal rule is displayed in Figure 84. Since we have no access to the syntactic information of the speech daughter S-DTR, we record only its prosodic, semantic and temporal information. The derivation of the mother is the result of the following operations over the TIME_START, TIME_END, PHON and SEM features: similarly as before, the alignment is licensed by the temporal overlap of the two modalities; then the prosodic information of the phrase is percolated from the prosodic information of the speech daughter, and finally, the semantics of the mother is accumulated from the semantics of the speech daughter, the semantics of gesture daughter and also the semantics contributed by the construction rule.

The HPSG-style semantic composition of the prosodic phrase “I enter” has been illustrated in Figure 85. Essentially, the semantics of the mother is accumulated from the semantics of the speech daughter and the semantics of the gesture daughter. As already shown in Figure 53, the SLOTS of the speech daughter is not empty to designate a semantically unsaturated phrase anticipating a complement. Also, the construction rule contributes *deictic_rel* between the semantic index $e_{12}$ of the speech and the semantic index $e_6$ of the gesture daughter. The further derivation would proceed by combining this multimodal prosodic phrase with the anticipated complement, the phrase “my apartment”. This combination would resolve the unspecified ARG2 of the

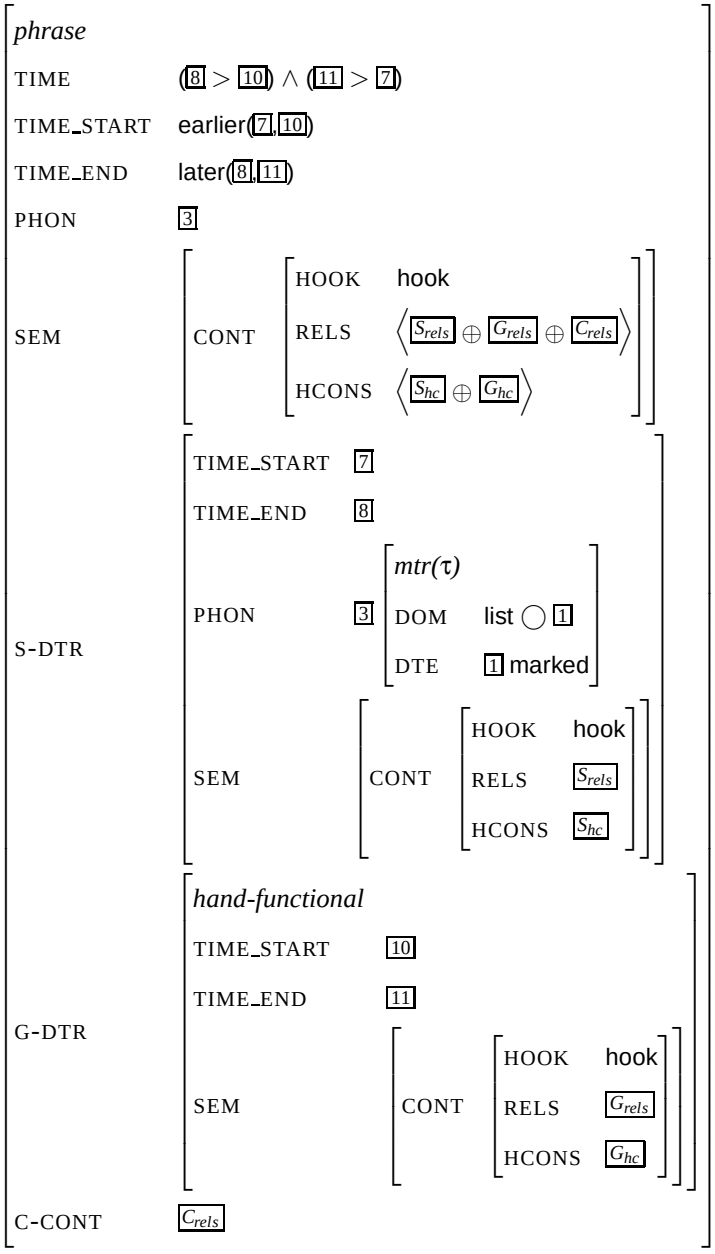



Figure 84: HPSG-based Situated Prosodic Phrase Constraint which accounts for the discrepancy between prosodic constituency and syntactic constituency

predicate *_enter_v_1* to a variable *x* of type individual. The result would be a prosodic tree associated with semantic information. Based on the architecture in Figure 82, the derivation of the syntactic tree and the semantic composition driven from this syntactic tree over the input strings proceeds separately from the prosodic derivation. In other words, we can construct the syntactic phrase shown in Figure 86 and compute its semantics without disrupting the traditional syntactic constituency. The final step of

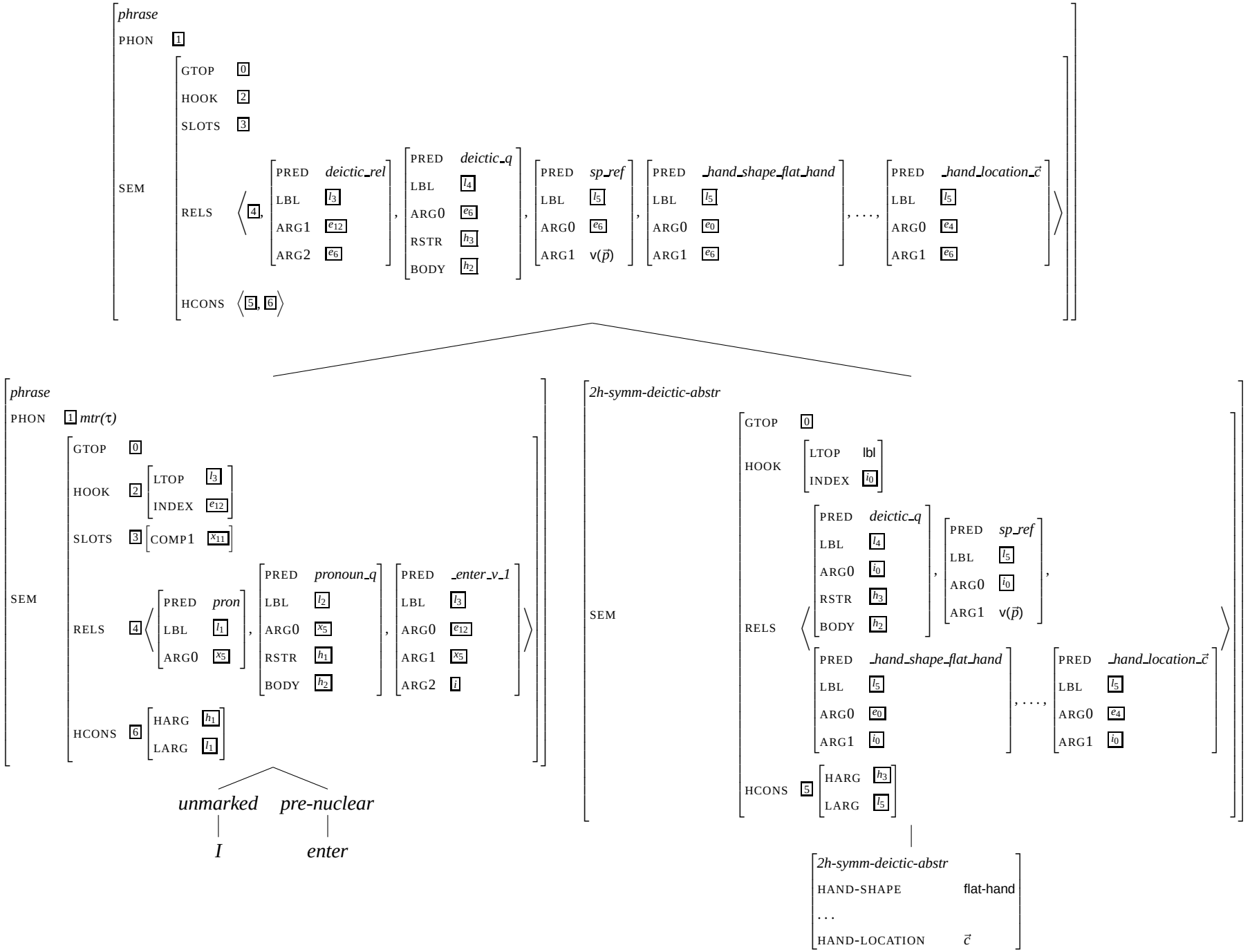



Figure 85: Semantic Composition for Deictic Gesture + the prosodic phrase “I enter”

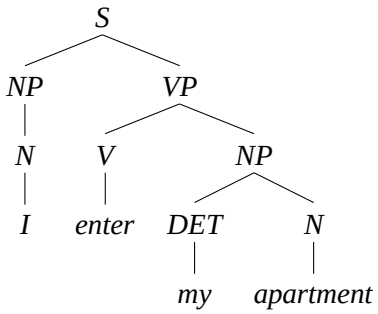



Figure 86: Syntactic Tree for “I enter my apartment”

the derivation concerns the unification of the prosodic/semantic phrase and the syntactic/semantic phrase following the procedure proposed in Figure 83. The compatibility between the SEM values is licensed by their identity.

## 6.5 Summary

This chapter concerned formalising the well-formedness constraints from Section 5.3 as construction rules in the grammar framework of HPSG. We chose the HPSG framework for several reasons, both theoretical (e.g., it can express non-isomorphic syntactic and prosodic constituency, which is vital given that prosody constrains the speech-gesture alignment; it supports semantic underspecification expressed in Minimal Recursion Semantics, which is also the semantic framework we use for the gesture form-meaning mapping) and practical (e.g., it affords an opportunity to implement our work in an existing wide-coverage online grammar).

In this chapter, we gave a brief introduction into how the metrical trees and the underspecified semantic representations of gesture can be converted into typed feature structures so as to make them entirely compatible with the selected grammar framework. We then presented the gesture type hierarchy that was designed with the view of accounting for the range of possible gesture performances. Also, we saw that the gesture sign was specified in parallel with the spoken sign which allowed for encoding features that are only appropriate for gesture signals.

The gesture type hierarchy guided the subsequent formalisation of the construction rules in that the gesture daughter is *typed*. Following the theoretical apparatus laid out in Section 5.3, we presented two major HPSG-based rule schemata: one for deriving a multimodal word, and another for deriving a multimodal phrase. The Sit-

uated Prosodic Word Constraint can be applied to both depicting and deictic gestures provided that the spoken element the gesture attaches to is nuclear/pre-nuclear prominent and also provided that there is temporal overlap between the performance of the gesture and the performance of the prosodic word. This constraint, however, is not sufficient to derive the range of the intended interpretations for concrete deictic gestures. Through real examples, we demonstrated that the temporal overlap and the prosodic prominence condition can be violated with deictic gestures that point at individuals/objects salient in the communicative act. We therefore introduced a rule that took this into account. Further to this, we presented the Situated Spoken Phrase Constraint which allowed for the alignment of a gesture to an entire constituent. Since this constraint does not capture the full range of attachments in that it can be applied only in cases of isomorphism between prosodic and syntactic constituency, we proposed a modular architecture that produces structured prosody and computes its semantics independently from the syntactic/semantic derivation. Based on that architecture, we proposed the Situated Prosodic Phrase Constraint which gives access to the form of the prosodic phrase and its semantics without having to first exploit its syntax (which was the main limitation of the syntax-driven approach of Klein [2000a; 2000b] for the speech-gesture alignment). The Situated Spoken Phrase Constraint, and also the Situated Prosodic Phrase Constraint are essential since they allow for multiple speech–gesture alignments which are not identical to the relative timings of speech and gesture modalities but which necessarily support the final interpretations in context, as determined at the semantics/pragmatics interface.

# Chapter 7

# Implementation of the Multimodal Grammar

Most hand-written online grammars aim to capture cross-linguistic generalisations and/or to test hypotheses across languages. To do that, the grammar engineer addresses mono-lingual or cross-linguistic phenomena of verbal input. The LinGo Grammar Matrix, for instance, provides a core linguistic knowledge that is common to various languages and that can be used as a start-up for the rapid development of grammars of diverse languages [Bender and Oepen, 2002]. In comparison, we intend to demonstrate that the form-meaning mapping of multimodal actions can be captured by using standard methods from linguistic theory, which can be formalised within large-scale grammar engineering platforms suitable for parsing. We shall test our hypotheses regarding multimodal syntax by building a computational model of speech-and-gesture well-formedness, thereby leveraging an existing broad-coverage grammar for English.

This chapter reports on the implementation of the multimodal grammar rules from Chapter 6. This work involves extending the existing wide-coverage LinGO English Resource Grammar (ERG [Flickinger, 2000]) with HPSG-based types and rules that use the form of the linguistic signal, the form of the gesture signal and their relative timing to constrain the meaning of the multimodal action. In Section 1.2, we stated that it is not within our aims to integrate the grammar module into a larger system for recognition and parsing of actual speech and gesture signals. This would require us to build a system that maps the visual input into typed feature structures representing the form of the gestural signal, and also a system for mapping acoustic signals to sequences of words, annotated with their prosodic information. Instead of working with raw visual and audible signals, we analyse TFS representations of multimodal

actions which are essentially what the output of such signal processing systems would look like. We will therefore evaluate our grammar using a manually-crafted test suite that is built in analogy to the traditional phenomenon-based test sets.

This chapter is structured as follows: Section 7.1 provides some background information about the implementation platforms. Section 7.2 details the implementation. It also outlines the major implementation challenges which stem from the divergences between a theoretical HPSG grammar and its computational implementation. Section 7.3 provides details of the evaluation.

## 7.1 Background of the Implementation Platforms

The grammar implementation is set within the Linguistic Knowledge Builder (LKB), which is a grammar and lexicon development environment for typed feature structure grammars such as HPSG [Copestake, 2002]. In LKB, all types, lexical rules and grammar rules are organised within a type hierarchy where subtypes inherit properties (including constraints) from their parents and compatibility between properties is enforced by their unifiability.

LKB is a platform suitable for parsing and generation where the standard input to parse is a sequence of strings.[1] We, however, represent multimodal signals by means of typed feature structures: the speech tokens are augmented with prosodic annotation and the gesture signals are expressed in feature-value pairs which capture the distinct aspects of form (recall Section 5.2). The LKB standard input was thus not suitable for our purposes. To solve this, we used the PET engine [Callmeier, 2000] which reads the same grammar files and produces the same output as LKB, and it also allows for injecting arbitrary XML-based feature structures into the input tokens, as detailed in Adolphs et al. [2008]. In formal terms, the lattice-based input to the PET parser is rendered in the XML-based Feature Structure Chart (FSC) format. This is vital for our purposes as we can inject the input speech tokens with prosodic annotation, and we can also formally represent the feature-value pairs of the gesture tokens.

An illustration of the feature structure for the input speech token can be viewed in Figure 87. The speech token is identified in terms of its +FORM and +PHON values: the former designates that the form of the token is the string “anna” and the latter designates that the speech token is a prosodically marked element of type *nuclear*

[1] Although our grammar will be reversible, in that it can be used by the LKB platform for parsing or generation, we will focus only on parsing here.

```
<fs type="speech_token">
    <f name="+FORM">
        <str>anna</str>
    </f>
    <f name="+PHON">
        <fs type="nuclear">
            <f name="+DOM">
                <str>anna</str>
            </f>
        </fs>
    </f>
</fs>
```

Figure 87: Fragment of a Feature Structure of the Input Speech Token “anna” in FSC format

with the domain of objects (+DOM) containing the string “anna”. The types and their appropriate features are declared in the type hierarchy.

Similarly, in Figure 89 we have shown the FSC-style notation of the feature structure of the input gesture token from Figure 72, repeated in Figure 88.

## 7.2 Implementation of Grammar for Gesture

An overall challenge for implementing a multimodal grammar using the current machinery stems from the fact that the multimodal input is non-linear. The standard input to the current HPSG parsing platforms such as LKB, PET and TRALE[2] is linearly ordered strings (or for PET, strings augmented with feature structures), and so they do not handle signals whose input comes from separate tiers connected through temporal relations. Further to this, the above-mentioned parsing platforms do not support quantitative comparison operations over the time stamps of the input tokens. Filling this gap is of importance, since the temporal performance of the speech signal relative to the temporal performance of the gesture signal is one of the conditions on multimodal alignment (see Chapters 4, 5 and 6).

In more general terms, parsing input stream that is not linear has been the focus

[2] http://www.sfs.uni-tuebingen.de/hpsg/archive/projects/trale/

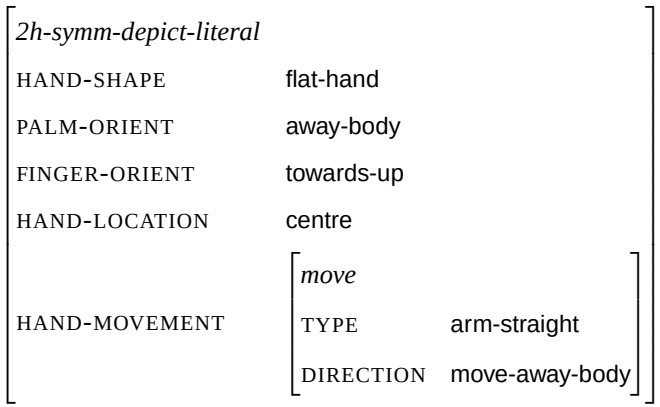



Figure 88: TFS representation of the depicting gesture in (2.4)

```
<fs type="gesture_token">
  <f name="+GESTURE">
    <fs type="2h-symm-depict-literal">
      <f name="HAND-SHAPE"><fs type="flat-hand"/></f>
      <f name="PALM-ORIENT"><fs type="away-body"/></f>
      <f name="FINGER-ORIENT"><fs type="towards-up"/></f>
      <f name="HAND-LOCATION"><fs type="centre"/></f>
      <f name="HAND-MOVEMENT"><fs type="move">
              <f name="TYPE"><fs type="arm-straight"/></f>
              <f name="DIRECTION"><fs type="move-away-body"/></f>
      </f>
     </fs>
  </f>
</fs>
```

Figure 89: Corresponding Feature Structure of the Input Gesture Token in FSC format

of work on spatial parsing for visual programming languages and for two-dimensional graphical interfaces [Lakin, 1987; Wittenburg, Weitzman, and Talley, 1991; Helm, Marruitt, and Odersky, 1991]. Further, the work of Johnston [1998b] discusses a generalisation of CKY chart parsing that supports parsing of inputs that are not linearly ordered.

The grammar rules for speech-and-gesture alignment proposed in Section 5.3 and their subsequent HPSG-style formalisation in Section 6.4 were based on the prosodic properties of the speech signal gesture temporally overlaps with: in general terms, we align gesture with a word that carries the nuclear and/or pre-nuclear accent, or a larger prosodic phrase containing that accent. We further saw that the construction rules contribute underspecified semantic information which ultimately affects how the multimodal action is interpreted in context. In the implementation, we made use of existing features in ERG to account for the relative timing between the speech signal and the gesture signal, and also to express temporal constraints. Then the constraint coming from the prosodic properties of the speech signal was handled via lexical rules. We also used lexical rules to encode the gestural semantics and the semantics contributed by the rule.

The overall grammar development proceeded in two separate stages: first, a pre-processing step which was carried out using the PET chart-mapping technology to declare temporal constraints; and second, parsing the pre-processed input using a grammar that captures linguistic constraints on the speech and gesture attachment and records the semantics of the multimodal action. In Section 7.2.1, we detail how we represented temporal overlap between the input speech and gesture tokens; in Section 7.2.2 we discuss the first, pre-processing, stage and then in Section 7.2.3 we provide details for the grammar rules. The chart-mapping rules and the grammar rules that we defined are included in Appendix C.

### 7.2.1 Representing Temporal Overlap in the Input FSCs

The raw input to our grammar is an Anvil annotation file in an XML format containing the speech and gesture information where each annotation element is stamped with the beginning and end values of its temporal performance (recall discussion in Section 4.1). To account for the overlapping temporal relation between the speech signal and the gesture signal, we made use of the +FROM and +TO features. These features are standardly used by ERG chart-mapping rules to define the span of the input items

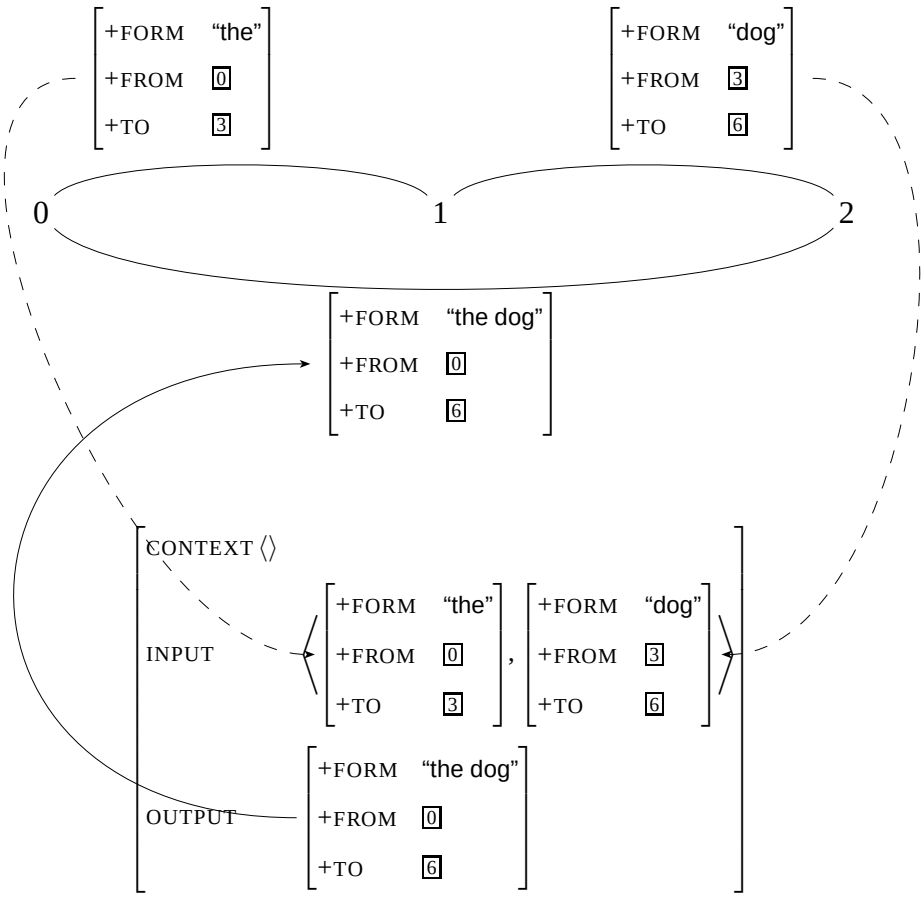



Figure 90: Example of a chart-mapping rule in PET that operates on unimodal input and output [Adolphs, 2009]

so that they could be instantiated as output tokens in the chart. For instance, the rule in Figure 90 operates by matching all input tokens from the chart with the rule arguments, then it removes the input from the chart and outputs an instantiated item [Adolphs, 2009]. We postpone the discussion of the CONTEXT, INPUT and OUTPUT features to the next section.

We accounted for temporal overlap by pre-processing the multimodal XML input so that overlapping TIME_START and TIME_END values of speech and gesture tokens were "translated" into identical +FROM and +TO values. In practice, this meant that overlapping speech and gesture tokens occupied the same place in the chart. This technique handled the non-linear multimodal input, and it also enabled the execution of comparative quantitative operations over the input tokens. For an illustration, consider Figure 91 which features an overlap between the temporal performance of the gesture token of type *rh-depict-literal* and the temporal performance of the speech to-

```
<fs type="speech_token">
  <f name="+FORM">
    <str>anna</str>
  </f>
  <f name="+FROM">
    <str>0</str>
  </f>
  <f name="+TO">
    <str>4</str>
  </f>
</fs>
```

```
<fs type="speech_token">
  <f name="+FORM">
    <str>ate</str>
  </f>
    <f name="+FROM">
      <str>5</str>
    </f>
    <f name="+TO">
      <str>11</str>
    </f>
</fs>
```

```
<fs type="gesture_token">
  <f name="+GESTURE">
    <fs type="rh-depict-literal"></fs>
  </f>
  <f name="+FROM">
    <str>0</str>
  </f>
  <f name="+TO">
    <str>11</str>
  </f>
</fs>
```

Figure 91: Handling temporal overlap between the input speech FSC and the input gesture FSC via identical +FROM and +TO values

kens “anna” and “ate”. Notice that the temporal relation is an *overlap*, which is an abstraction over the full range of fine-grained temporal relations between speech (S) and gesture (G) such as (*precedence(start(S), start(G))* $\wedge$ *identity (end(S), end(G))*), as detailed in Table 19, page 145.

### 7.2.2 Pre-processing via Chart-Mapping Rules

The first step of the multimodal grammar implementation involved pre-processing the speech-and-gesture input with the view of linking the gesture token with the temporally overlapping speech token(s). This operation was enabled via the powerful pre-

processing mechanism of *chart-mapping* which involves re-writing a token chart item into a chart item of arbitrarily specified information [Adolphs et al., 2008]. The chart-mapping rules are of the form:

$$\begin{bmatrix} \text{CONTEXT} & \langle\rangle & \\ \text{INPUT} & \rightarrow & \text{OUTPUT} \\ \text{POSITION} & \langle\rangle & \end{bmatrix}$$

where the CONTEXT, INPUT, OUTPUT and POSITION specify a possibly empty list of attribute-value matrices. The CONTEXT attribute constrains the application of the rule without consuming its element(s); the INPUT component contains the element(s) to be re-written and OUTPUT is the re-written component. The re-writing process is exhaustive in that the element on the input is removed from the chart. To avoid re-writing elements onto elements spanning exactly from the first to the last vertex marker, the POSITION attribute allows for selecting the chart position of input element(s) relative to the output and/or context element(s). To date, PET supports the following positional constraints:

- Identity between two items: formally expressed as `I1 @ I2`
- Strict precedence and strict sequence between two items: `I1 < I2` and `I1 > I2`
- Precedence and sequence between two items: `I1 << I2` and `I1 >> I2`. Note that this relation subsumes `I1 < I2` and `I1 > I2` allowing for some redundancy how particular relations can be expressed.

#### 7.2.2.1 Attaching Gesture Token to the Temporally Overlapping Speech Token

We handled the temporal overlap by means of chart-mapping rules that linked the gesture element with the speech element based on their relative timing expressed through the +FROM and +TO values. For that purpose, we first defined `gesture-unary-rule` (see Figure 92) as a chart-mapping rule that re-writes an input speech token in the context of a gesture token into a combined multimodal token of type *speech+gesture_token*. This rule copies the +GESTURE value of the gesture token and the +PHON value of the speech token onto the output, "gesture-marked", item. The +PHON attribute contains the prosodic type of the speech item and +GESTURE attribute is a feature-structure representation of the type shown in Figure 89. Informally, this rule transforms a speech token from the input chart into a multimodal token for which both the spoken (e.g.,

```
gesture-unary-rule := chart_mapping_rule &
      [ +CONTEXT < gesture_token &
                   [ +GESTURE #gesture ] >,
        +INPUT < speech_token &
                   [ +PHON #phon ] >,
        +OUTPUT < speech+gesture_token &
                   [ +GESTURE #gesture,
                     +PHON #phon ] > ].
```

Figure 92: Fragment of the Definition of `gesture-unary-rule`

+PHON) and the gestural (e.g., +GESTURE) features are appropriate. Since the input speech token is specified within the INPUT attribute, this rule removes it from the chart to re-write it to a combined speech-and-gesture token.

Further, recall from the empirical findings in Chapter 4 that depicting gestures, abstract deixis and nomination deixis require temporal overlap between the gesture stroke and the speech element, whereas concrete deictic gestures allow a higher degree of freedom between the temporal performance of the deictic stroke relative to the temporal performance of the speech item. The rule `gesture-unary-rule` is therefore inherited by further rules that consider the different temporal properties. The rule `gesture-unary-rule-1` (see Figure 93) re-writes a speech token into a combined speech-and-gesture token if the span of the gesture item in the chart is identical to the span of the speech item: that is, they temporally overlap. This is expressed via co-indexation between the +FROM and +TO values of the gesture token and the +FROM and +TO values of the speech token. The output item is a gesture-marked speech token occupying the same position in the chart as the input token, which is expressed via

```
[ +POSITION "O1@I1" ]
```

This rule is applied to gestures of type *hand-formational* where *hand-formational* is a supertype of all gestures, including depicting, abstract, nomination and concrete deixis (recall the gesture type hierarchy in Figure 62).

In contrast, `gesture-unary-rule-2` (see Figure 94) is applied only to gestures of type *deictic-concrete* so as to account for the permitted precedence and sequence relations between the speech token and the concrete deictic gesture token. Unlike the co-indexed +FROM and +TO values in `gesture-unary-rule-1`, the rule `gesture-unary-rule-2`

```
gesture-unary-rule-1 := gesture-unary-rule &
      [ +CONTEXT < [ +GESTURE hand-formational,
                     +FROM #from,
                     +TO #to ] >,
        +INPUT < [ +FROM #from,
                   +TO #to ] >,
        +POSITION "O1@I1" ].
```

Figure 93: Definition of `gesture-unary-rule-1`

```
gesture-unary-rule-2 := gesture-unary-rule &
  [ +CONTEXT < [ +GESTURE deictic-concrete ] >,
    +POSITION "O1@I1" ].
```

Figure 94: Definition of `gesture-unary-rule-2`

remains neutral about the positional (and hence temporal) relation between the gesture token and the speech token, permitting thus the attachment of the gesture token of type *deictic-concrete* to each speech token from the input chart. 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. While the operations of strict precedence or strict sequence might suffice for attaching the concrete deixis in (7.1) to "she", this alignment would be omitted if the gesture was performed a few milliseconds later, along with "spinach", as shown in (7.2). Both performances seem plausible and there is thus no reason to block the attachment of a concrete deictic gesture and a prosodically unmarked element that does not strictly precede or follow the gesture performance.

(7.1) She $\underline{[_N \text{ate}]}$ spinach
*Speaker extends his index finger in the direction of the person designated by "she".*

(7.2) She ate $\underline{[_N \text{spinach}]}$
*Same gesture as in (7.1)*

Likewise, this rule would allow for attaching the concrete deictic gesture from (4.8),

page 114 to “she” despite the fact that the performance of the deixis is not adjacent to that of the pronoun in speech.

To sum up, the application of the chart-mapping rule `gesture-unary-rule` or the two rules inherited from it results in a speech token augmented with gesture feature-values typed as *speech+gesture_token*. The re-writing mechanism copies the combined speech-and-gesture token into the original position in the chart, and so it occupies the exact same cell as the input speech token. In so doing, any operations over the output token would proceed in the same way as for the speech token.

#### 7.2.2.2 Attaching Gesture Token to Multiple Temporally Overlapping Speech Tokens

The chart-mapping rules described above enabled the alignment of a single speech token and a gesture token, that is, a gesture that was performed along a single speech element. To account for the majority of cases where the gesture overlaps more than one speech token, we introduced further chart-mapping rules that distribute the gestural information onto multiple speech tokens within the gestural span. Since the platform does not support underspecification of the arity of the input tokens, we specified separate rules for each arity of multiple speech tokens. The rule `gesture-split-2-rule` in Figure 95, for instance, re-writes the gestural information of the input gesture token into two tokens whose +FROM and +TO values are co-indexed with that of the input speech tokens. The output tokens occupy the same positions in the chart as the speech tokens (expressed through the positional constraints `O1@C1, O2@C2`).

Upon that, the `gesture-unary-rule` is applied so as to instantiate a multimodal *speech+gesture_token* for each speech token. The final result is multiple gesture-marked speech tokens whose span is identical to the span of the gesture.

### 7.2.3 Lexical Rules

In the grammar, we added to the existing ERG word rules prosodic and gestural information by unifying the +PHON and +GESTURE features of the input tokens with the PROSODY and GESTURE values of signs, as shown in Figure 96.[3] By convention, this information is stored as attributes of ORTH. This rule guarantees that the prosodic and

[3] To avoid clash with the already existing feature PHON in ERG, we introduced the attribute PROSODY to encode the prosodic properties of the speech signal. This feature has the same status as the feature PHON used in the HPSG-based analysis in Chapter 6.

```
gesture-split-2-rule := chart_mapping_rule &
  [ +CONTEXT < speech_token & [ +FROM #from, +TO #mid1 ],
               speech_token & [ +FROM #mid2, +TO #to ] >,
    +INPUT < gesture_token &
               [ +FORM #form, +FROM #from,
                 +TO #to, +GESTURE #gesture ] >,
    +OUTPUT < gesture_token &
               [ +FORM #form, +FROM #from,
                 +TO #mid1, +GESTURE #gesture ],
                 gesture_token &
               [ +FORM #form, +FROM #mid2,
                 +TO #to, +GESTURE #gesture ] >,
    +POSITION "O1@C1, O2@C2" ].
```

Figure 95: Fragment of the Definition of `gesture-split-2-rule`

```
basic_word :+
  [ ORTH [ GESTURE #gesture,
           PROSODY #prosody ],
    TOKENS.+LIST < [ +GESTURE #gesture,
                     +PHON #prosody ], ... > ].
```

Figure 96: Extension of the Definition of `basic_word`

gestural information of the output gesture-marked tokens from the pre-processing step will be copied onto signs.

Now that the GESTURE and PROSODY attributes are made appropriate for signs, we first need to add constraints to the alignment based on the prosodic properties of the speech sign, and second, to incorporate the semantic contribution of the gesture. We therefore defined the rule `gesture_lexrule` (see Figure 97) as a supertype of the lexical rules specific to depicting gestures and to deictic gestures. The rule `gesture_lexrule` propagates the SYNSEM and ORTH features up to the mother sign. Note, however, that we do not propagate the gestural feature GESTURE to block any further recursive instantiation of this rule. This is rendered by the type *no-gesture*

```
gesture_lexrule := phrase_or_lexrule &
        [ ORTH [ FORM #form, FROM #from, TO #to,
                 GESTURE no-gesture,
                 PROSODY #prosody ],
          SYNSEM synsem,
          ARGS < [ ORTH [ FORM #form, FROM #from, TO #to,
                          GESTURE hand-formational,
                          PROSODY #prosody ],
                   SYNSEM synsem ] > ].
```

Figure 97: Fragment of the Definition of `gesture_lexrule`

which fails unification with the type *hand-formational*.

The rule `gesture_lexrule` is further inherited by rules specific to deictic gestures and to depicting gestures. This higher degree of specificity allows us to encode the distinct behaviour of the gesture types at the levels of prosody and semantics. Recall from Section 5.3.3 that concrete deictic gestures allow for a temporal and/or prosodic relaxation in that they can be semantically related with speech elements that are not prosodically prominent and/or whose temporal performance is outside the temporal performance of the deixis. Further, we observed in Section 5.2.2 that there is a difference in how depicting gestures map to meaning vs. how deictic gestures map to meaning. Whereas the form-meaning mapping of depicting gestures is based on iconicity—that is, the form resembles its meaning, the form-meaning mapping of deixis is constrained by indexicality—that is, the form of the pointing medium identifies the spatial location of the referent.

We define the rule `depicting_gesture_lexrule` (see Figure 98) as a subtype of `gesture_lexrule` to constrain the gesture-marked word to a prosodically prominent word of type *nuclear_or_pre-nuclear* thereby preventing a prosodically unmarked or a non-nuclear gesture-marked word to undergo this rule. In other words, this rule captures our finding that depicting gestures co-occur with (pre-)nuclear prominence in speech, and so this rule would not produce an analysis for the ill-formed signal in (1.4), repeated in (7.3).

(7.3) * Your [$_N$mother] called.

*The speaker puts his hand to the ear to imitate holding a receiver.*

```
depicting_gesture_lexrule := gesture_lexrule &
        [ ARGS < [ SYNSEM.LOCAL.CONT.RELS.LIST.FIRST.LBL #dltop,
                   ORTH [ GESTURE depicting,
                          PROSODY nuclear_or_pre-nuclear ]] >,
          C-CONT [ HOOK [ LTOP #ltop,
                          INDEX #index ],
                   RELS <! vis-rel &
                        [ PRED vis_rel,
                          LBL #ltop,
                          M-ARG #index,
                          S-LBL #arg1,
                          G-LBL #arg2 ],
                        [ PRED g_mod_rel,
                          LBL #glbl,
                          ARG1 #harg ],
                        [ LBL #larg1 ], [ LBL #larg2 ], [ LBL #larg3 ],
                        [ LBL #larg4 ], [ LBL #larg5 ], [ LBL #larg6 ] !>,
          HCONS <! geq & [ HARG #arg1, LARG #dltop ],
                   qeq & [ HARG #arg2, LARG #glbl ],
                   qeq & [ HARG #harg, LARG #larg1 ],
                   qeq & [ HARG #harg, LARG #larg2 ],
                   qeq & [ HARG #harg, LARG #larg3 ],
                   qeq & [ HARG #harg, LARG #larg4 ],
                   qeq & [ HARG #harg, LARG #larg5 ],
                   qeq & [ HARG #harg, LARG #larg6 ]  !> ] ].
```

Figure 98: Definition of depicting_gesture_lexrule

The semantic contribution of the lexical rule is declared within the C-CONT attribute. The implementation of the semantics component for depicting gestures follows the principles detailed in Section 5.2. The lexical rule introduces a *vis-rel* (see the definition in Figure 99) whose treatment is similar to the treatment of *subord-or-conj-relation*. With subordinating conjunctions such as “because”, “while” and with coordinating conjunctions such as “and”, “but”, the English Resource Grammar uses a relation of type *subord-or-conj-relation* [Flickinger, Bender, and Oepen, 2003]. This relation introduces two arguments which, for subordinating conjunctions, are identified with the LTOP of the main clause and with the LTOP of the matrix clauses, and for coordinating conjunctions, with the LTOPs of the main clauses. In a similar way, *vis-rel*

```
vis-relation := relation &
[ M-ARG index,
  S-LBL handle,
  G-LBL handle ].
```

Figure 99: Definition of `vis-rel`

takes two handles as arguments which correspond to the main labels of the speech component and the gesture component: this is obtained by enforcing equality in HCONS between the value of S-LBL (that is, `#arg1` is equated with `#dltop`) and value of the gesture-marked word (that is, `#arg2` is equated with `#glbl`). This equation is a slight modification from the theory proposed in Section 5.2, in that S-LBL is not identified with the LTOP of the speech component. This is for the sake of quantifiers such as "this", whose scope floats among other scope-bearing elements and hence their LTOP is not equal to their LBL. Then the LTOP of gesture corresponds to the label of the $[\mathcal{G}]$ operator (rendered as `g_mod_rel`). Note also that the two arguments S-LBL and G-LBL are in an "outscoping" `geq` (greater than or equal to) constraint to express underspecification of the range of spoken phrases aligned with gesture. In this way, 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. For instance, plugging the gesture to "bottom" in (2.4), repeated in (7.4) means that the relation is not restricted to the EP contributed by "bottom" but it can also be over the EPs of a higher projection.

(7.4) The $[\underline{_{PN}\text{bottom}]}$ worked $[_{N}\text{fine}]$ ...
*Both hands are rested on the knees. The speaker lifts them in the frontal space with palms almost facing forward, fingers extended and moves them rapidly to the left and right periphery.*

Further, recall from our theory in Section 5.3 that *vis-rel* introduces an M-ARG which will be the argument of the gesture-marked word, and so it will serve as an argument to any external predicate. Unlike conjunctions in ERG, the value of the PRED attribute—*vis-rel*—is not provided by the lexical entry but by the lexical rule.

The gesture semantics is a bag of elementary predications (see Section 5.2), all of which are outscoped by the gestural modality $[\mathcal{G}]$ (represented as `g_mod_rel`). The

```
basic_deixis_lexrule := gesture_lexrule &
 [ ARGS < [ SYNSEM.LOCAL.CONT [ HOOK.LTOP #ltop ],
            ORTH.GESTURE deictic ]  >,
   C-CONT [ HOOK [ LTOP #ltop ] ] ].
```

Figure 100: Definition of basic_deixis_lexrule

rule depicting_gesture_lexrule therefore introduces in RELS labels (here #larg1 ...#larg6) for each gestural EP. The instantiation of the particular EPs comes from the gestural lexical entry as they are mapped from the gestural feature-value pairs. To account for the fact that they are outscoped by the gestural modality, we enforced hole conditions (ERG represents them as qeq) between ARG1 of the gestural modality and the label of each predication. As already discussed (see Section 5.2.2), hole conditions are used as markers of where a labelled formula can be plugged into a hole.

To specify the distinct semantic contribution of deixis, gesture_lexrule is also inherited by the rule basic_deixis_lexrule (see Figure 100) and its two subtypes: deixis_lexrule_6_rel (see Figure 101) and deixis_lexrule_7_rel. Since the number of RELS cannot be left underspecified, the differentiation between the two subtypes captures gestures contributing six semantic relations and gestures contributing seven semantic relations.[4] The behaviour of both rules deixis_lexrule_6_rel and deixis_lexrule_7_rel is analogous and we shall concentrate on the former. Similarly as before, the rule introduces the gesture semantics to gesture-marked signs. The implementation of this rule follows the main principles outlined in Section 5.2: it introduces a deictic_rel, illustrated in Figure 102, which is defined as a three-argument relation where the main argument is of type *event* and the speech and gesture arguments (S-ARG and G-ARG, respectively) are of type *semarg*, that is, attribute value matrices.

To account for the fact that this relation holds between the speech component and the gesture component, the value of S-ARG of the deictic relation is made token identical with the value #sarg of the main argument of the speech component, and the value

[4]Recall from the gesture type hierarchy (Section 6.3) that a gesture performed with the finger—i.e., the hand shape is of type *finger*–would contribute two feature value-pairs: FINGER-DIGIT and FINGER-FORM. This is not the case with gestures whose hand shape is of type *fist* and *flat-hand*. Given that the form-meaning mapping reads the elementary predications directly off the feature structure of the gesture form, we need to distinguish the exact number of predications yielded by the gesture. The same distinction holds for depicting gestures as well.

```
deixis_lexrule_6_rel := basic_deixis_lexrule &
 [ ARGS < [ SYNSEM.LOCAL.CONT [ RELS.LIST.FIRST.ARG0 #sarg ] ] >,
   C-CONT [ HOOK [ LTOP #ltop ],
            RELS <! deictic_relation &
                  [ PRED deictic_rel,
                    LBL #ltop,
                    S-ARG #sarg,
                    G-ARG #garg ],
                  [ PRED deictic_q, ARG0 #garg, RSTR #rstr ],
                  [ PRED sp_ref,
                    LBL #sp-lbl,
                    ARG0 event_or_index & #garg,
                    ARG1 v_p_space ],
                  [ LBL #sp-lbl, ARG0 event, ARG1 #garg ],
                  [ LBL #sp-lbl, ARG0 event, ARG1 #garg ],
                  [ LBL #sp-lbl, ARG0 event, ARG1 #garg ],
                  [ LBL #sp-lbl, ARG0 event, ARG1 #garg ],
                  [ LBL #sp-lbl, ARG0 event, ARG1 #garg ],
                  [ LBL #sp-lbl, ARG0 event, ARG1 #garg ] !>,
            HCONS <! qeq & [ HARG #rstr, LARG #sp-lbl ] !> ] ].
```

Figure 101: Definition of deixis_lexrule_6_rel

```
deictic_relation := arg0_relation &
  [ PRED deictic_rel,
    ARG0 event,
    S-ARG semarg,
    G-ARG semarg ].
```

Figure 102: Definition of deictic_relation

of G-ARG is made token identical with the value #garg of the main argument of the gesture semantics. Also, in ERG the index of scope-bearing elements such as negation and scopal adverbs is not identified with their main argument, ARG0, and we have thus made ARG0 (and not the semantic index) of the speech component token identical with the S-ARG of the deictic relation. This slight deviation from the theoretical principles

```
abstract_deixis_lexrule := gesture_lexrule &
 [ ARGS < [ ORTH [ GESTURE deictic-abstr_or_nom,
                   PROSODY nucl_or_pre-nucl ] ] > ]..
```

Figure 103: Definition of `abstract_deixis_lexrule`

laid out in Section 5.2—i.e., S-ARG of the deictic relation is identical with the HOOK | INDEX of the speech daughter (and not its main argument, ARG0)—is for the sake of a relation between scopal elements and deictic gesture. The rest of the implementation of the deictic semantics is consistent with our theory: we have already mentioned that deictic gestures provide the spatial reference `sp_ref` of an event or an individual (encoded as type `event_or_index`) in the denoted physical space. We use the type `v_p_space` as a notation for $v(\vec{p})$ where $v$ is the function that maps the physical space identified by the gesture to the actually denoted space. This distinction is essential as it allows us to account for abstract and nomination deixis where the physical space identified by the hand is not identical to the denotation of the gesture itself (recall earlier discussion about this matter in Section 5.2). Again for consistency with the ERG, every individual is bound by a quantifier and hence `deictic_q` binds the reference introduced by the gesture. This is accounted for by the constraints in HCONS which stipulate equality between the restrictor of the quantifier `#rstr` and the label of the spatial reference `#sp-lbl`.

Finally, to capture the different interaction between abstract (and nomination) deixis and speech signals on one hand, and between concrete deixis and speech signals on the other, we defined two rules, subtypes of `gesture_lexrule`: `abstract_deixis_lexrule` (see Figure 103) is applied to gestures of type `deictic-abstr_or_nom` and it constrains the PROSODY of the gesture-marked sign to a prosodically marked word of type `nucl_or_pre-nuclear`. In contrast, `concrete_deixis_lexrule` which is applied only to gestures of type `deictic-concrete` (see Figure 104) is less restrictive in terms of the prosodic marking of the gesture-marked word in that it allows for both prosodically marked and prosodically unmarked words to combine with the gesture semantics. This is formalised using the prosodic type `pros`, the root node in our prosodic hierarchy (see Figure 58). This formalises our finding that only gestures identifying objects or individuals salient in the communicative situation are not constrained by the prosodic prominence of the speech element.

```
concrete_deixis_lexrule := gesture_lexrule &
  [ ARGS < [ ORTH [ GESTURE deictic-concrete,
                    PROSODY pros ] ] > ].
```

Figure 104: Definition of `concrete_deixis_lexrule`

Since in ERG the semantic composition is guided by the syntactic tree, we were not able to account for gestures attaching to prosodic phrases that are not isomorphic to syntactic constituents. For that purpose, we would have to construct grammar rules for the prosodic grouping equipped with semantics and then to provide interface rules for the unification of the syntax/semantics component and of the prosody/semantics component—an endeavour that lies outwith the scope of the current implementation.

## 7.3 Evaluation

We evaluated the grammar by measuring its coverage over a hand-crafted test suite.[5] This is the standard method for evaluating manually created precision grammars. In Section 7.3.1, we provide an overview of the construction of the test suite. In Section 7.3.2, we discuss the grammar performance produced by the grammar profiling system [incr tsdb()].

### 7.3.1 Test Suite Design

In contrast to test corpora, which include a sample of naturally occurring data, test suites are systematically constructed and well structured with the view of addressing the full range of phenomena of interest, thereby aiming for exhaustivity. The construction of our test suite is analogous to the one of traditional phenomenon-based test suites used for testing all DELPH-IN grammars:[6] it is manually crafted to ensure coverage of well-formed and ill-formed data, but inspired by an examination of natural data [Oepen, Netter, and Klein, 1997]. The key properties (and advantages over test corpora) of test suites are *systematicity, control over test data* and *progressivity*, as detailed by Oepen, Netter, and Klein [1997]. For the purposes of constructing a

[5] The complete test suite will be released upon the completion of this dissertation. It is not included in the current thesis due to space limitations.

[6] `http://moin.delph-in.net/`

multimodal test set, we modified these properties as follows:

- **Systematicity.** This involves systematic testing of a particular phenomenon over well-formed examples and their ill-formed counterparts. The ill-formed test items were derived by applying the following operations:

  1. Prosodic permutation: varying the prosodic markedness of the input strings, for instance, from test suite item (7.5) we derive (7.6) as a further item (it is marked as ill-formed to reflect intuitions of native English speakers and our empirical findings from Chapter 4):

     (7.5) [$_N$Anna] ate
     *Depicting gesture along with "Anna"*

     (7.6) * Anna [$_N$ate]
     *Depicting gesture along with "Anna"*

  2. Gesture variation: testing the integration of a speech item with the distinct gestural dimensions, for instance we derive (7.8) and (7.9) from (7.7):

     (7.7) Anna [$_N$ate]
     *Concrete deixis along with "Anna"*

     (7.8) * Anna [$_N$ate]
     *Depicting gesture along with "Anna"*

     (7.9) * Anna [$_N$ate]
     *Abstract deixis along with "Anna"*

  3. Temporal permutation: moving the gestural performance over the distinct speech items, for instance:

     (7.10) [$_N$Anna] ate
     *Depicting gesture along with "Anna"*

     (7.11) * [$_N$Anna] ate
     *Depicting gesture along with "ate"*

- **Control over test data.** The design of a test suite allows for isolating and testing particular phenomena without being redundant. For instance, we tested the interaction of a prosodically unmarked verb head with a temporally co-occurring gesture only in the context of intransitive constructions and not transitive ones. Having a control over the vocabulary is yet another advantage of test suites. We used a small vocabulary containing non-ambiguous words. We also used three

gesture lexical items (see Section C.3) as representatives of the different gesture types: depicting, abstract deictic and concrete deictic. The inclusion of more gesture items of the same dimension has certainly no effects on the grammar performance.

- **Progressivity.** The gradual increase in the complexity of the test items is essential for the isolation of a particular phenomenon. In our test suite, each test item (or rather a subset of test items) addressed a single phenomenon, thus allowing us to systematically investigate the behaviour of a particular phenomenon in interaction with the prosodic, gestural and temporal permutations, discussed above. Following our earlier observations about the interaction of gesture with the syntax of the utterance (Section 1.3.2.2), we addressed the following core linguistic phenomena:
  - intransitivity: “Anna ate”
  - transitivity: “Anna ate spinach”
  - complex NPs of a determiner and a noun: “This student ate”
  - complex NPs of a determiner, modifier and a noun: “This crazy student ate”
  - coordination of pre-head modifiers: “This crazy but intelligent student ate”
  - negation: “Anna didn’t eat”
  - S adverbial modification: “Presumably Anna ate”
  - VP adverbial modification: “Anna probably ate”
  - VP coordination: “He attacked and beat the enemy”

The basic test set illustrating the full range of grammatical phenomenon contained 9 well-formed unimodal test items. Upon applying the operations of prosodic, gestural and temporal permutations, the test suite amounted to 471 multimodal test items, of which 339 well-formed and 132 ill-formed. This test suite can be used to evaluate *any* multimodal grammar, not only an HPSG-based one.

Our motivation for not using naturally occurring data as a test corpus is based on the tradition for constructing test suites for a wide range of natural language applications: a test corpus does not provide a representative sample of the range of phenomena of interest, it does not contain systematically produced permutations, and it does not aim for exhaustiveness [Oepen, Netter, and Klein, 1997].

### 7.3.2 Grammar Performance Testing

The performance of the grammar was tested within the [incr tsdb()] grammar profiling system since it enables fully automated batch processing of test suites, and also since it creates a competence and performance profile of the grammar coverage [Oepen, 2001]. The tool fully supports FSC-based input as long as each test item is strictly kept in a single line. We initially produced the test items as plain text files which were then automatically converted into test suite files suitable for parsing with [incr tsdb()]. Following the formatting instructions of Emily Bender,[7] we annotated each test example with the following pieces of information:

- Source: designates where the example comes from. Our test items were marked with "author", that is, they were constructed by the author.
- Vetted: indicates where the grammaticality judgement of the test item comes from. It has the following options: "t" vetted by a native speaker, "f" not vetted, and "s" vetted from a grammar book. Our test items were marked with "f".
- Judgement: indicates the grammaticality judgement with the two options being "g" grammatical and "u" ungrammatical. As stated above, our test suite includes a mixture of grammatical and ungrammatical examples.
- Phenomena: lists the phenomena represented by a test item. We provided the linguistic phenomenon (for instance, `_intransitivity, _coord`), the prosodic permutation (for instance, `_subject_np_accented` indicates that the subject NP was accented and `_verb_head_accented` indicates an accented verb head), the gesture dimension (where the options were depicting `_gest_depict`, abstract deixis `_gest_deictic_abstract` or concrete deixis `_gest_deictic_concrete`) and the temporal performance of the gesture in relation to the syntactic category of the speech element (for instance, `_gest_subj_np` indicates that the gesture happened along the subject NP and `_gest_verb_head` indicates a gesture performance along with the verb head).

The coverage profile over the test suite is summarised in Table 20. To enable the grammar writer to closely inspect the coverage, [incr tsdb()] allows for setting up an aggregation criterion such as phenomenon, grammaticality, number of words or input length. The left-most column in Table 20 uses the aggregation criterion of length of

[7] Presented here `http://courses.washington.edu/ling567/testsuites.html`

| ‘gesture/12-08-29/pet’ Coverage Profile | | | | | | |
|---|---|---|---|---|---|---|
| **Aggregate** | **total items** ♯ | **positive items** ♯ | **word string** ϕ | **lexical items** ϕ | **total results** ♯ | **overall coverage** % |
| $100 \leq$ *i-length* $< 105$ | 42 | 42 | 101.00 | 26.29 | 42 | 100.0 |
| $95 \leq$ *i-length* $< 100$ | 84 | 50 | 97.00 | 26.60 | 50 | 100.0 |
| $75 \leq$ *i-length* $< 80$ | 78 | 54 | 76.93 | 12.00 | 54 | 100.0 |
| $65 \leq$ *i-length* $< 70$ | 83 | 83 | 68.00 | 9.42 | 83 | 100.0 |
| $60 \leq$ *i-length* $< 65$ | 166 | 96 | 64.00 | 9.42 | 96 | 100.0 |
| $55 \leq$ *i-length* $< 60$ | 6 | 6 | 57.00 | 7.00 | 6 | 100.0 |
| $50 \leq$ *i-length* $< 55$ | 12 | 8 | 53.00 | 7.00 | 8 | 100.0 |
| **Total** | **471** | **339** | **76.11** | **14.35** | **339** | **100.0** |

(generated by [incr tsdb()] at 29-aug-2012 (10:43 h))

Table 20: Coverage Profile of Test Items generated by [incr tsdb()]

input items, that is, the number of words within a test item. Note, however, that here the aggregation is not applied to strings but to complex feature structures and hence the numbers are not representative of the actual length of the input test item. The next column—**total items**—shows the number of well-formed and ill-formed test items per aggregate, the total of which amounts to 471 test items. The next column to the right—**positive items**—shows the number of well-formed items per aggregate where the total amounts to 339 items. The **word string** column computes the average length of the test items per aggregate (these values should be viewed symbolically since they are based on feature structures); the **lexical items** column shows the average number of lexical items (again, these values are computed from feature structures). The **total results** column displays the number of results found per aggregate. Essentially, our grammar is able to parse successfully all positive items and hence the **overall coverage** in the right-most column is 100%. We manually inspected the derived analyses to make sure that they are consistent with our theory. While the grammar successfully parses all well-formed examples, the inclusion of `gesture-unary-rule-2` results in overgeneration (recall from Section 7.2.2 that `gest-unary-rule-2` applies to concrete deictic gestures and it enables the gesture to attach to any adjacent token without further constraints). For instance, in (7.2), `gesture-unary-rule-2` produces a parse tree where the gesture attaches to both “she” and “ate”: while the former produces a plausible

| 'gesture/12-08-29/pet' Overgeneration Profile | | | | | | |
|---|---|---|---|---|---|---|
| **Aggregate** | **total items** ♯ | **negative items** ♯ | **word string** ϕ | **lexical items** ϕ | **total results** ♯ | **overall coverage** % |
| $100 \leq$ *i-length* $< 105$ | 42 | 0 | 0.00 | 0.00 | 0 | 0.0 |
| $95 \leq$ *i-length* $< 100$ | 84 | 34 | 97.00 | 25.82 | 0 | 0.0 |
| $75 \leq$ *i-length* $< 80$ | 78 | 24 | 75.00 | 12.00 | 0 | 0.0 |
| $65 \leq$ *i-length* $< 70$ | 83 | 0 | 0.00 | 0.00 | 0 | 0.0 |
| $60 \leq$ *i-length* $< 65$ | 166 | 70 | 64.00 | 9.43 | 0 | 0.0 |
| $55 \leq$ *i-length* $< 60$ | 6 | 0 | 0.00 | 0.00 | 0 | 0.0 |
| $50 \leq$ *i-length* $< 55$ | 12 | 4 | 53.00 | 7.00 | 0 | 0.0 |
| **Total** | **471** | **132** | **74.17** | **14.05** | **0** | **0.0** |

(generated by [incr tsdb()] at 29-aug-2012 (10:45 h))

Table 21: Overgeneration Profile of Test Items generated by [incr tsdb()]

semantic interpretation, the latter is semantically infelicitous.

The profiling system also produces an Overgeneration Profile, displayed in Figure 21. This profile is generated in analogy to the Coverage Profile, and so the various columns represent the same elements per aggregate. The main difference is that the Overgeneration statistics uses the negative data as a base set.

Finally, [incr tsdb()] creates a performance profile, displayed in Table 22. In contrast to the coverage and overgeneration profiles which are indicative of the grammar behaviour, the performance profile rather illustrates system efficiency. The specific information is again organised per aggregation criterion — here this is the length of the input test items. The second column from left-to-right—**items**—shows the number of items per aggregate. Then the **etasks** column displays the average number of carried out parsings; the **filter** column is the percentage of filtered out parsings; the **edges** column shows the average number of constructed edges in the chart; **first** is a measurement of the time (in seconds) to produce the first parse; **total** displays the total time in seconds spent to produce the total number of parses; and finally **space** illustrates in kilobites the memory used for parsing [Oepen, 2001]. Notice that the increase in length of the input test items is accompanied by an increase in the resources allocated.

| **'gesture/12-08-29/pet'** Performance Profile | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Aggregate** | **items** ♯ | **etasks** ϕ | **filter** % | **edges** ϕ | **first** ϕ (s) | **total** ϕ (s) | **space** ϕ (kb) |
| $100 \leq$ *i-length* $< 105$ | 42 | 831 | 97.7 | 217 | 0.04 | 0.04 | 8884 |
| $95 \leq$ *i-length* $< 100$ | 84 | 672 | 97.6 | 162 | 0.03 | 0.03 | 8309 |
| $75 \leq$ *i-length* $< 80$ | 78 | 271 | 96.9 | 67 | 0.01 | 0.01 | 6565 |
| $65 \leq$ *i-length* $< 70$ | 83 | 232 | 97.2 | 67 | 0.01 | 0.01 | 6428 |
| $60 \leq$ *i-length* $< 65$ | 166 | 187 | 96.9 | 48 | 0.01 | 0.01 | 6226 |
| $55 \leq$ *i-length* $< 60$ | 6 | 142 | 96.8 | 49 | 0.01 | 0.01 | 6164 |
| $50 \leq$ *i-length* $< 55$ | 12 | 126 | 96.3 | 38 | 0.01 | 0.01 | 6042 |
| **Total** | **471** | **350** | **97.4** | **90** | **0.02** | **0.02** | **6921** |

(generated by [incr tsdb()] at 29-aug-2012 (10:46 h))

Table 22: Performance Profile of Test Items generated by [incr tsdb()]

## 7.4 Summary

This chapter discussed the implementation of the multimodal grammar in the PET parsing platform. Whereas wide-coverage of the multimodal phenomena was not envisaged, this chapter provided a proof-of-concept that gesture can be expressed by the same formal techniques and tools applied to language. Needless to say, it was necessary to adapt those techniques and tools to the specific nature of multimodality without modifying the parsing platform. For instance, most parsing platforms support only input from strings with the system of Johnston [1998b] being a notable exception. We solved this issue by using the PET chart-mapping machinery which allowed us to augment each string with arbitrarily specified feature structures. Further to this, the main challenge for the multimodal grammar engineering stems from the fact that the multimodal input is not linear: namely, it comes from distinct channels that mutually interact through temporal relations. To represent temporal overlap, we used identity between the +FROM and +TO values of the input speech token and the +FROM and +TO values of the input gesture token. We used this pre-processing step to record the candidates for the speech-gesture alignment. Further, we used lexical rules to constrain the choices of the alignment using the prosodic properties of the speech element, to equip the multimodal elements with the gestural semantics, and also to define an underspecified relation between the aligned speech element and gesture element. The

final step of the implementation involved testing the grammar coverage and performance against a test suite, manually crafted in the tradition of the phenomenon-based test suites. We reached 100% coverage for all well-formed examples and 0% for their ill-formed counterparts. The value of this test suite is that it can be used for testing other grammars equipped with speech and gesture.

This multimodal grammar engineering effort contributes to the current approaches to multimodal studies in being the first one to attempt implementation of a grammar of gesture in a domain-independent way. Essentially, it builds on an existing wide coverage grammar and it provides the background for further development of wide-coverage multimodal grammars.

# Chapter 8

# Conclusions

The final chapter concludes our study of the alignment of speech and co-speech hand gesture in a constraint-based grammar. We end with a summary of the main claims and challenges addressed in this thesis. We then proceed with a discussion of this thesis' original contributions and we finish with an outline of the possible directions for future work.

## 8.1 Summary

This thesis advanced a new theory that analysed the form-meaning mapping of multimodal actions using well-established linguistic methods such as constraint-based syntactic derivation and semantic composition. Our main claim, set out in Chapter 1, is that the mapping of multimodal form to meaning is captured within a grammar that produces abstract meaning representations that *underspecify* what the multimodal action can mean in context. Our grammar rules constrain the speech-gesture alignment as informed by the *form* of the linguistic utterance. Given current models of the semantics/pragmatics interface—specifically that discourse structure blocks the availability of referents (see Sections 1.3.1 and 5.1)—our construction rules comply with those models by means of constraints on the alignment, thereby capturing the plausible pragmatic interpretations in context. Otherwise stated, the alignment of speech and gesture (which we formalise via attachments in the syntax tree) constrains the gesture interpretation, and hence the relation between the speech content and the gesture content in a specific context.

Capturing multimodal meaning via underspecification semantics (see Section 5.2) is based on the highly ambiguous gesture form which is not sufficient to yield a com-

plete interpretation even in the final context of use. In Chapter 2, we showed that the ambiguities project onto several levels:

1. A single gesture often corresponds to **multiple interpretations** in context, where each interpretation supports a different relation between the speech content and the gesture content. We illustrated the distinct logical forms mapped from the form of the multimodal action using example (2.8), repeated in (8.1), where the speaker performed a circular movement with her right hand over the left palm.

   (8.1) So [$_{H^*}$he mix]es [$_{X^*}$mud] ...
   *Speaker's left hand is rested on the knee with palm open supine. The right hand is held loose with fingers facing downwards over the left hand. The speaker performs consecutively four rotation movements with her right hand over the left palm.*

   Interpreting the gesture as denoting the mud and some salient property of it—namely, that it was going round—would support an Elaboration relation between the denotation of "mud" and the gesture action (i.e., the gesture provides more context regarding the mud). Another possibility would be to interpret the gesture as enacting the event of mixing mud from the speaker's viewpoint, which would support a Depiction relation. Other interpretations are also possible (recall (5.2), (5.3) and (5.4)).

2. The distinct gesture interpretations are derived from distinct speech-gesture alignments which we formalise via **attachment ambiguities** in the syntax tree: namely, there is ambiguity with respect to the tree node gesture attaches to, which in turn blocks the availability of referents that serve to derive the plausible interpretation for the multimodal action. For instance, interpreting the gesture in (8.1) as denoting the mud is supported by a gesture attachment to the NP "mud" which blocks the availability of the denotations of the verb "mixes" and the NP "he" for interpreting this gesture. Conversely, an attachment to the S "he mixes mud" gives access to the NP's denotation and the verb's denotation, and hence they feature when reasoning about what the gesture means in context.

3. The **gesture form**, i.e., the physical shape of the hands while performing the motion, is often ambiguous which enables a single gesture to resolve to different predications in context which are not always of unique arity.

4. We talked about **"syntactic" ambiguity** to reflect the fact that outside of context the hand motion can be often ambiguous with respect to its dimension—depicting vs. deictic—which ultimately affects the form-meaning mapping, and hence the relation between the speech content and the gesture content.

The grammar produces meaning representations that support the plausible interpretations of multimodal actions, while constraining the choices of alignment governed by the *form* of the linguistic phrase. In Chapter 1, we motivated aligning speech and gesture in the grammar on the grounds that *form* constrains the alignment, which in turn has effects on the pragmatic interpretation of gesture. Given the current models of the semantics/pragmatics interface, the pragmatic interpretations of multimodal actions are inferred from their compositional semantics, commonsense reasoning and world knowledge, but, crucially, not from their form directly (in other words, the compositional semantic representation is rich enough to support commonsense reasoning with the context to yield the pragmatic interpretation). The grammar, therefore, captures constraints on the choices of the possible speech and gesture alignments. This was empirically validated in Chapter 4 where we provided empirical evidence that the performance of gesture interacts with the prosodic prominence in speech. More specifically, the gesture strokes are performed along with the nuclear accents in speech in the default case of broad-focussed utterances. An early pre-nuclear rise which is signalled by a higher acoustic pitch is also a reliable factor that could predict the gesture performance. While this prosodic constraint holds for depicting gestures and for abstract and nomination deixis,[1] we established that a certain degree of freedom is possible with concrete deictic gestures that identify the spatial coordinates of referents salient in the communicative action. In particular, a concrete deixis can be semantically related with a speech element that is *not* prosodically prominent. Likewise, a concrete deictic gesture can be performed a few milliseconds before or after uttering the semantically related speech element (it being prominent or non-prominent). We assume that this prosodic and/or temporal mismatch is compensated for by salience in the communicative event: namely, the physical co-location enables the interlocutor to anchor the gesture referent, and so the mismatch does not violate perception. The same compensatory mechanism does not seem to hold for depicting and abstract/nomination deictic gestures that anchor their referents in the speech and not in the physical space.

Given this empirical evidence, Chapter 5 proposed grammar rules that account for

[1] We had one instance of a nominating gesture, but they generally behave like abstract pointing in that the referent is not at the spatial coordinates identified by the pointing signal.

multimodal well-formedness. The proposed rules are sensitive to the type of ambiguities discussed above, thereby producing abstract representations supporting the contextual interpretations. We claimed that the temporal performance of speech relative to the temporal performance of gesture alone is too restrictive with respect to the possible alignments: for instance, while the gesture referring to throwing ground rice in (1.8) was performed along with the verb "throw", we demonstrated that the gesture can also attach to the VP "throw ground rice", and hence semantically relate with its denotation. The construction rules therefore license multiple attachments in the syntax tree, where the constraints on the attachments come from the prosodic prominence of the temporally overlapping speech signal. In so doing, the rules not only comply with the empirical evidence that gesture interacts with prosodic prominence in speech, but also they produce abstract meaning representations without undergenerating or overgenerating the gesture interpretations in context. The grammar framework is thus entirely compatible with the current models of discourse structure at the semantics/pragmatics interface, and so resolving the underspecified logical formulae happens by using standard mechanisms for pragmatic reasoning.

Driven by our aim to extend an existing wide-coverage linguistic grammar with construction rules for speech-gesture alignment, we formalised those rules in the HPSG framework (see Chapter 6). Since the gesture performance is constrained by the prosodic structure, an added challenge was to account for the divergences between syntax tree and prosodic tree. A limitation of the existing HPSG-based analysis of prosodic constituency is that it is syntactocentric [Klein, 2000a; Klein, 2000b]: namely, the prosodic derivation is driven by the syntactic derivation, and so the prosodic structure is accessed only through the syntactic structure. Since this was insufficient for our purposes, we proposed a different architecture where the syntax tree (equipped with its semantics) is built independently from the prosodic tree (equipped with its semantics) over the same list of input elements, and there is a rule that unifies the syntax/semantics component and the prosody/semantics component. The advantage of this model is that we can attach a gesture to a prosodic tree that does not necessarily correspond to a syntax tree, and still produce a meaning representation of the multimodal action. By using this framework, we demonstrated that an isomorphism between the syntax/semantics component and prosody is not a necessary condition for the formalisation of construction rules for speech-gesture alignment.

Chapter 7 concluded our study with an implementation of the theoretical rules in an existing online grammar for English—the English Resource Grammar. Using the

current grammar engineering and parsing platform, the main challenge for the grammar engineer was the non-linear input, i.e., the speech and gesture modalities overlap temporally, and the alignment is constrained by the relative timestamps of the input tokens. We handled this via a pre-processing step where we mapped the overlapping timestamps of the speech and gesture tokens into identical token edges. In so doing, we were able to perform quantitative operations over them such as overlap, precedence and sequence (where precedence and sequence were applied to concrete deixis). We further used lexical rules to spell out the constraints coming from prosody, and also to equip the conjoined multimodal entities with the gesture semantics and with the underspecified relations between the speech content and the gesture content (*vis_rel* for depicting gestures or *deictic_rel* for deictic gestures). In the tradition of evaluating hand-crafted linguistic grammars, we evaluated the grammar coverage by using a manually constructed multimodal test suite. We systematically tested syntactic phenomena (intransitivity, transitivity, complex NPs, coordination, negation and modification) over 471 well-formed and ill-formed examples (72% well-formed) where the ill-formed items were derived by means of the following operations: prosodic permutation (varying the prosodic markedness over the speech elements); gesture variation (testing distinct gesture types) and temporal permutation (moving the gestural performance over the distinct speech items). We reached 100% coverage of all well-formed test items and 0% coverage of the ill-formed examples.

## 8.2 Contribution

While the literature offers some formal analyses of multimodal syntax (see Section 3.5), there does not yet exist a formal model of speech-gesture alignment that is domain-independent, that is predictive about multimodal (un)grammaticalities and that is uniform with the current models of the semantics/pragmatics interface. This thesis' original contribution is to fill this gap: we addressed the form-meaning mapping of multimodal actions within a domain-independent grammar that spells out constraints on the speech-gesture alignment, as inspired by examination of natural data, and that outputs meaning representations that are entirely compatible with pragmatic reasoning. Moreover, our formal analysis of multimodal actions does not restrict gesture interpretation to a single possibility (as it is standardly done in the literature), but it instead models the full range of gesture ambiguities by exploiting mechanisms for underspecifying meaning. The crucial output of this analysis are underspecified logical formulae that support

the plausible interpretations of gesture in context, and also the distinct ways gesture can be semantically related with speech. It is standardly assumed in the gesture literature that speech and gesture denote the same referent (see Chapter 3). Using examples, we demonstrated that identity between the speech content and the gesture content is only a subset of the ways speech and gesture relate. By underspecifying gesture meaning as mapped from form, and by underspecifying the semantic relation between speech and gesture (expressed via the underspecified relations *vis_rel* and *deictic_rel*), our model can scale up to the various ways gesture can be interpreted, and hence to the various relations that can be inferred between the speech content and the gesture content. In so doing, our grammar theory is the first one to actually provide the methodology for abstracting over the range of multimodal meanings in context (including those that we did not encounter in our corpora).

Our model accounts for the full spectrum of gesture interpretations as revealed by the ambiguous gesture form, while constrained by the *form* of the speech signal. In so doing, we contributed a novel multimodal grammar theory that is constrained not only by quantitative criteria—the timing of gesture relative to the timing of speech—but also qualitative criteria—the prosodic properties of the speech signal. While the relative timings enforce one-to-one mappings, the inclusion of the linguistic properties allows us to analyse the one-to-many mappings and thus to account for the various ways (as licensed by form) gesture can attach to the speech signal, and hence be semantically related with it.

By drawing on well-established methods from linguistics, this work achieved a grammar model that is entirely uniform with the principles for constraint-based semantic composition for purely linguistic grammars. The grammar output is abstract meaning representations that fully comply with the existing coherence-based models for inferring pragmatic interpretations. We practically demonstrated this by augmenting an existing wide-coverage online grammar with an implementation of the theoretical grammar rules.

## 8.3 Future Directions

While the interdisciplinary nature of this thesis opens up various possible directions for future work, we will enumerate just a few.

In this thesis, we analysed gestures in terms of symbolic representations: namely, typed feature structures. A natural direction would be the recognition of the visual sig-

nal and its mapping to feature structure representations. This task is challenging as it requires fine-grained information about the hand shape and the direction of the movement. Such information can be obtained by highly sensitive motion capture sensors or by sophisticated models of visual processing. The other challenge for the gesture recognition involves the distinction between communicative and non-communicative hand signals, and also the recognition of gesture boundaries. We anticipate that the co-temporal speech (such as its acoustic properties and informativeness), along with the kinetic properties of the hand motion are reliable parameters for distinguishing communicative signals, and also for recognising the beginning and ending of gestures.

This work could also proceed with a study of the interaction between prosody and gesture (stroke) for signalling information structure. Given the empirical evidence that the performance of gesture is reliant on the prosodic prominence in speech, and also given that prosody is the standard marker of information status in English, we hypothesise that the gesture performance plays a role in signalling salience in the discourse model. An added value to this work would be to study the information structure–gesture correspondence in languages that use devices, other than prosody, for expressing topic/focus phenomena, e.g., discourse particles in Japanese and Korean, word order in Italian and Spanish, or a mixture of syntactic and prosodic devices such as Bulgarian and Croatian. Some of the questions that could be addressed include: are the distinct alignments from the grammar a matter of focus projection; how does gesture interact with the two dimensions of information structure related phenomena (focus/background and theme/rheme); do focus markers across languages (pitch accent, phonological phrasing, word order, particles, etc) interact with the gesture stroke to signal the contrasting elements in the discourse?

Another direction for future research would be to adapt the existing grammar engineering platforms to make them entirely suitable for implementing and parsing *multimodal* grammars. While we used a pre-processing step to account for the temporal overlap, a multimodal grammar engineering platform should be able to handle the non-linear multimodal input, thereby performing quantitative comparison operations over the time stamps of the input tokens. Given our approach of using the same formal language for speech and co-speech gesture, we demonstrated that the difference between speech and gesture is just a matter of degree: while the content of units of speech can be unambiguously recovered using pragmatic reasoning, gestures often remain ambiguous even within the final context of use. Instead of aiming at one preferred analysis, the multimodal grammar engineering platform should support higher level of ambiguous

attachments and corresponding meaning representations.

# Appendix A

# Instructions for Gesture Annotation in Anvil

## A.1 Introduction

These instructions have been prepared as a guide for annotating spontaneous, improvised hand gestures performed in speech accompaniment. The annotators' task is to mark the exact point in time where each gesture starts and ends, what type of gesture it is (to be explained shortly), and also what it means in its specific context-of-use. In order to make a judgement about the gesture meaning, it is essential to listen to the speech.

### A.1.1 Gestures: Binary Classification

In general, hand movements are classified into *communicative* or *non-communicative* signals. Whereas the former contribute content to the discourse (e.g., pointing to some particular fruits while uttering the NP "these apples" in "I really liked these apples"[1] unambiguously determines which apples the speaker refers to), the latter are meaningless bodily movements such as nervous ticks or movements satisfying bodily needs (e.g., rubbing the eyes, scratching one's nose, adjusting one's hair) that do not play a communicative role. Annotators will be asked to mark initially **all** hand movements as either communicative or non-communicative and upon that to provide more detailed information about the communicative movements, which are our matter of interest.

[1] In the examples that follow, we use underlining to identify the spoken phrase which is uttered while performing the gesture

### A.1.2 Gesture Categories

We recognise four main categories of communicative gestures:

1. **Iconic gestures.** They depict what is conveyed in the accompanying speech by creating a visual image of its feature(s) in the space just in front of the speaker (i.e., the virtual space). For instance, a round hand movement while uttering "He mixes mud" visually depicts the act of mixing something; the extension of the left and right hand to the left and to the right periphery respectively while saying "I caught a really big fish creates a visual image of the size of the fish.

2. **Metaphoric gestures.** They also depict what is conveyed in the accompanying speech but the depicted image represents an abstract concept of the accompanying speech. For instance, holding one hand up with an open palm and fingers relaxed while saying "I have many more ideas" can be interpreted as the metaphor for the container containing the ideas.

3. **Deictic gestures.** The speaker points to something abstract or concrete in the space around him. Depending on the object of reference, deictic gestures fall into several subgroups as follows:

   - *concrete* deixis: for instance, the speaker points with an extended index finger or an extended arm to a physically present individual while uttering "I saw this man".
   - *abstract* deixis: the speaker points to a virtually created object in the frontal space. For instance, a speaker says "And there was this guy" and he uses his index finger to refer to some individual while, in fact, there is no one physically present.
   - *nomination* deixis: the speaker extends an index finger to give prominence to a word or phrase. For instance, a speaker says "You'll notice that's also part of Laxton junior school" and along with "Laxton" he raises his voice and extends his index finger to make it more prominent and discernable among the other words [Kendon, 2004].

4. **Beats.** The hand(s) beats along with the rhythm of the speech to highlight important bits of the speech content. For instance, the speaker waves his hand up and down along with the stressed word while saying "Go ahead" [Pelachaud et al., 1995].

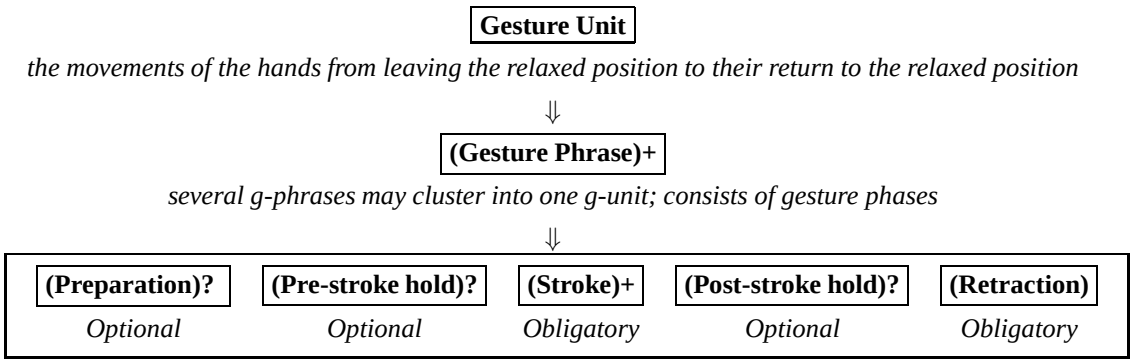



Figure 105: Structural Organisation of Gesture

Note that these gesture categories are not mutually exclusive. It is often the case that within a single gesture we can identify features of more than one category. For instance, if a speaker says "Just go round the corner" while extending his right hand forward and flexing his palm to the left direction, this gesture incorporates some deictic features (indicating a point in space) and iconic features (the flexing movement depicts a corner). Annotators should be particularly attentive to the gestural mutual inclusivity and should record it accordingly.

### A.1.3 Gesture's Anatomy

For each gesture, annotators will be asked to mark:

- the point in time where the gesture begins and ends
- the gesture phases (to be discussed shortly)
- the gesture meaning

The annotation is based on the gesture anatomy of McNeill [2005] and Kendon [2004]. The entire gesture excursion (see Figure 105)—the *gesture unit*—comprises the period of time between successive motions of the limbs. A gesture begins with the departure of the hands from a relaxed position and ends with their return to a rest. A gesture unit can contain one or more *gesture phrases (g-phrases)* which include the interval from the beginning of a gesture to its most expressive part. The gesture phrase is what one intuitively recognises as a 'gesture'. The g-phrase itself consists of one or more *gesture phases*. Depending on how the gesture unfolds in time, the gesture phrase may contain:

- *preparation* (optional): the physical effort necessary for the hands to move from rest and perform the stroke.
- *pre-stroke hold* (optional): a cessation before the most expressive part. The pre-stroke hold serves as a trigger for the lexical item(s) produced during the stroke. The pre-stroke hold indicates the place where the speech-gesture synchrony is about to begin and it also helps to re-establish the temporal cohesion.
- *stroke* (obligatory): the most prominent content-bearing element; it also involves the greatest kinetic effort.
- *post-stroke hold* (optional): the fingers and/or hand(s) sustain their expressing position. The post-stroke hold lets the speaker maintain the idea conveyed by the stroke.
- *recovery/retraction* (obligatory): return to a resting position. The recovery is not part of the g-phrase in Kendon [2004].

The goal of the annotation task is to identify the meaning of a gesture in context, and so *it is important to pay particular attention to the content-bearing components of the gesture: the gesture stroke and the gesture post-stroke hold*, the combination of which is also known as a nucleus [Kendon, 2004] and an expressive phase [Kita, van Gijn, and van der Hulst, 1998]. The rest of the phases (preparation, pre-stroke hold, retraction) are the physical effort necessary for the limbs to reach or relax from the expressive focus. They do not, however, contribute content to the multimodal action.

## A.2 Annotation Tool and Process

### A.2.1 Anvil

The annotation is performed on a 63-second video fragment provided by Loehr [2004]. Gestures will be annotated using the Anvil labelling tool [Kipp, 2001] which is freely available (I can also provide a copy). The annotators will be asked to install the software; instructions for how to do this are included in the README file of the software package. The user interface (see Figure 106) can be described as follows:

- the upper middle window displays the main video

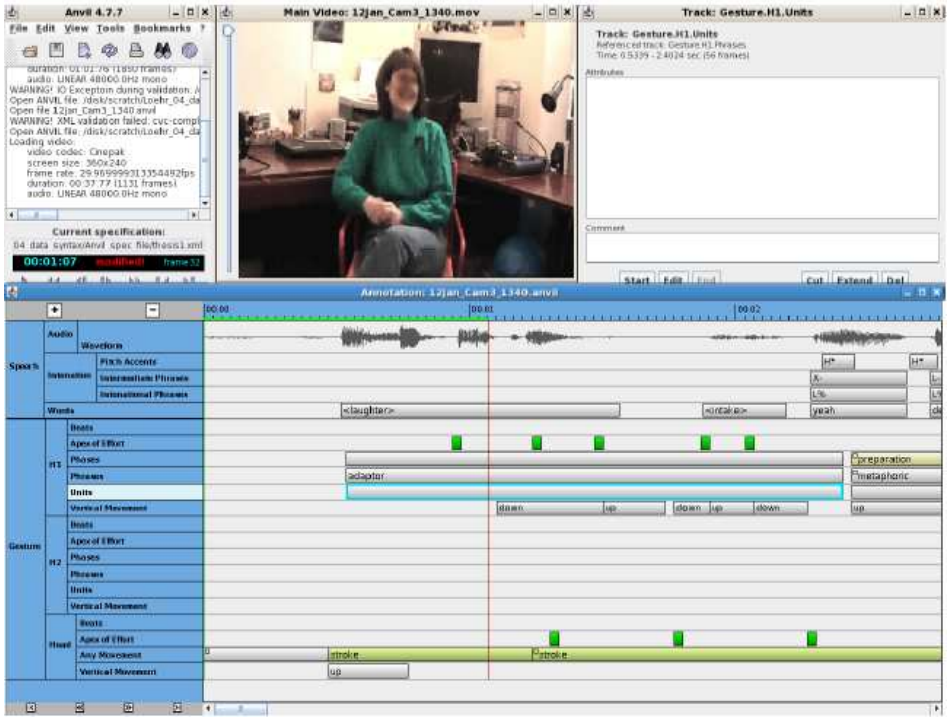



Figure 106: Anvil Interface

- the upper left window displays the program's specification and can be used for navigating through the video
- the upper right window details the currently selected track (its attributes and comments)
- the lower window is the so called *annotation board* — this is where the coding is performed. The left hand side displays the user-defined tracks and groups, and the right hand side is the vertical playback line synchronised with the video.

### A.2.2 Procedure

Prior to annotation, the annotators should make sure they have available:

i. the Anvil annotation tool: this can be either downloaded for free from `http://www.anvil-software.de/download.html` (an e-mail of request to the author, Michael Kipp, is required) or I can give you a copy

ii. video file (to be annotated): *28Sep_Cam1_0000.mov*

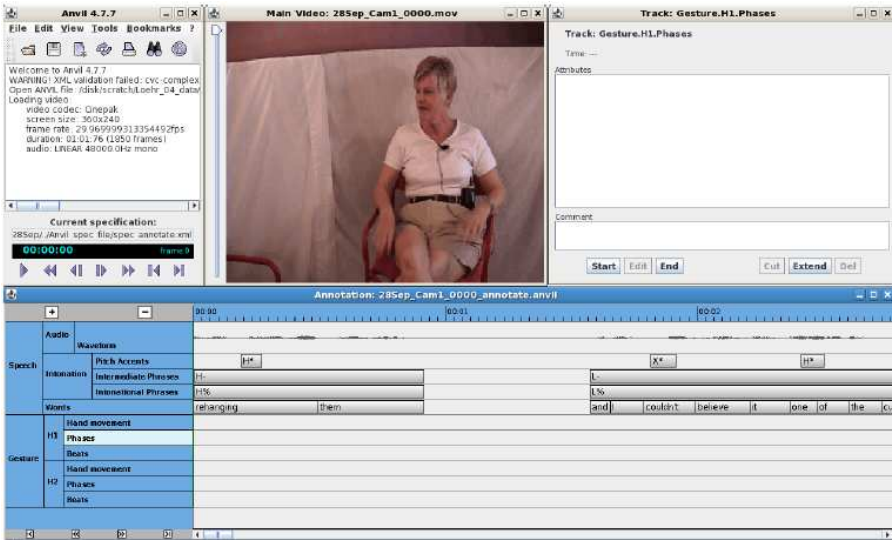



Figure 107: Anvil Interface Prior Gesture Annotation

iii. specification file: *spec_annotate.xml*

iv. an anvil xml file: *28Sep_Cam1_0000_annotate.anvil*. This file already contains speech transcription and prosodic information, and your task is to augment it with information about gestures.

1. PASS 1: Startup Anvil following the instructions in the Anvil documentation corresponding to your OS. Then click **File** > **Open** in Anvil's menu bar and load the video *28Sep_Cam1_0000.mov* file, the specification file and the anvil xml file by browsing to the containing folder. Anvil will ask you for some optional information such as your name and some comment. After loading all files, your workspace should look as shown in Figure 107. Watch the complete video segment from the beginning to the very end *without* interruption. This helps get a general idea of the topic and also of the speaker's way of interacting. At this stage, focus on the video (you can enlarge the video window) and ignore the annotation board.

2. PASS 2: The speech and gesture information is provided in the form of tracks as follows:

   (a) The large upper track labelled Speech contains the audio waveform, the prosodic annotation and the transcribed speech. They are not subject to

editing but please if you notice some annotation along the Speech track that you highly disagree with, make a note and we will discuss it further.

(b) The lower track group labelled Gesture (which is empty now) will contain the gesture annotation. This group is split into two subgroups. The subgroup, marked as H1, refers to the speaker's dominant gesturing hand, i.e., the hand that plays the dominant role in executing the gesture signal. Whenever both hands are involved in a two-handed gesture, this is annotated within the H1 track. The subgroup, marked as H2, refers to the hand that plays a lesser role in signalling the gesture. Within each subgroup, we have specified the following subtracks:

i. Hand Movement: the track that contains the qualitative information of the gesture, namely:
   * whether the movement is communicative or non-communicative
   * what category the gesture is (iconic, metaphoric, emblem, deictic, or any combination of them)
   * the meaning of gesture

ii. Phases: see the definition above.

iii. Beats: see the definition above. Note that even if both hands are engaged in, say, an iconic gesture, the hands can still beat along the speech rhythm.

3. PASS 3: In the first step, the annotators will be asked to identify the beginning and end of *all* hand movements no matter whether at a first glance they appear significant or not. For this purpose, click on the Hand Movement track to activate it, and then place the green record line at the beginning of the movement, the red playback line at the desired end and then press END on the track window as displayed in Figure 108, or right-click on END from the drop-down list. The window shown in Figure 109 automatically pops up. Please note that information about the specific items is hidden behind the bubble on the RHS of the window. At this stage, it is highly recommended that you do not enter any information apart from the binary classification communicative vs. noncommunicative movement. For this reason, select your choice from the drop-down menu of *binary* and leave the boxes unticked and the fields empty for a futher pass after accumulating more information. Click on OK to record the information and return to the annotation board. Continue to the end of the video file.

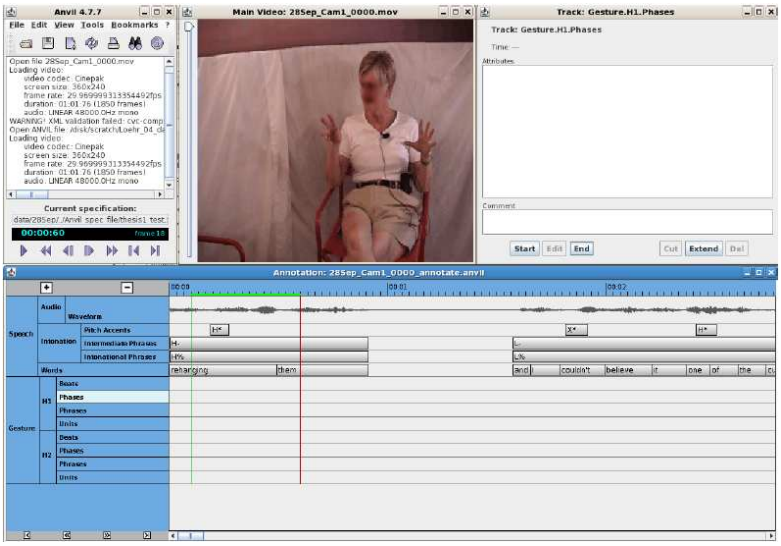



Figure 108: Adding Track Element

4. PASS 4: At this stage, we focus only on the communicative gestures. Take into consideration that the stroke boundaries may be different from the boundaries of the hand movement marked in PASS 3. The annotation shall be performed from the smaller to the larger composing elements. We begin by identifying the gesture phases within a phrase in the following manner:

   (a) Annotate the gesture stroke. For that purpose, you need to compare the meaning conveyed by the hands with the meaning conveyed by the speech (words, phrases and also larger discourse units). Take into account that the stroke is typically of greater effort in terms of forcefulness of movement, tenseness of handshape, etc. It is recommended that you first navigate through the video segment, identify the gesture and only after that you add it to the annotation board.

   Once you have found the span of a stroke, you need to insert a new track element by marking its beginning and its end on the annotation board. To do this, follow the procedure described above: i.e., click somewhere in the Phase track to activate it, place the green record line at the beginning, the red playback line at the desired end and press END on the track window. An edit window pops up (see Figure 110) where you will be prompted to

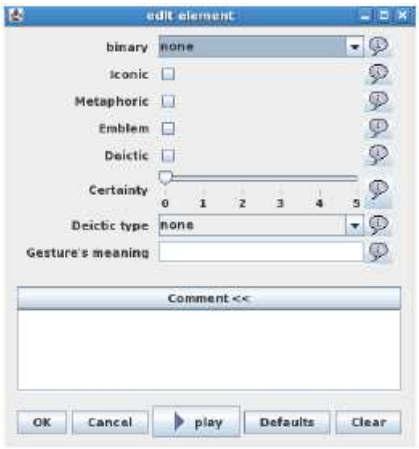



Figure 109: Window for Qualitative Information


Figure 110: Edit Window

specify the gesture phase. Click OK to proceed.

(b) Annotate the other phases: preparation, retraction. These phases might be absent, i.e., two consecutive strokes are quite common. Follow the above-mentioned procedure. *In case of doubt whether to annotate one phase as a preparation for the following gesture or as a retraction of the preceding gesture, Duncan [Under perpetual revision] recommends to decide in favour of the preparation for the following gesture based on the assumption that "a gesture is a forward-looking activity reflective of ideas yet to come in speech".*

(c) Annotate hold phases: pre-stroke hold and post-stroke hold. These are also optional. Take into consideration that the pre-stroke hold usually serves

as an orientation where the speech-gesture synchrony is about to begin, whereas the post-stroke hold demarcates an elaboration, continuation of the idea expressed by the stroke. Note, however, that what appears as a post-stroke hold may not be expressive, and thus have no discernable meaning. For instance, the speaker does not retract the hand to a relaxed position while exhibiting speech disfluencies. In this case, the energy articulated by the hand is no longer perceived, and so it does not convey any meaning. This can be solved by listening to the synchronous speech.

Continue to the end of the video file. Make sure that the sum all gesture phases within one hand movement equals the temporal interval of the hand movement. For instance, if one gesture contains preparation, stroke and post-stroke hold, the beginning of the preparation should be in the exact same time frame as the beginning of the hand movement; the end of the post-stroke hold should be the same as the end of the hand movement.This is easy to see by navigating with the green and red lines.

5. PASS 5: Go back to the beginning of the file and add the missing information about each communicative hand movement. Click on the element to activate it, then either right-click and select EDIT or click on the EDIT button from the track window. The window from Figure 109 appears again. Please enter all the missing information, namely: tick off the relevant box(es) for the gesture category(-ies) (ticking off more than one box indicates that one gesture incorporates features of several gesture categories). If you have selected a deictic gesture, specify its type from the drop-down list. Note that in case of a mutually inclusive gesture, only concrete deixis can occur with another gesture category. This means that an occurrence of an abstract deixis or nomination deixis plus an iconic or metaphoric gesture is not permissible. To identify the gesture meaning, attend to the gesture stroke, post-stroke hold (see the remark above about non-expressive holds), the speech and the overall discourse. The meanings are expected to be of the form: a sentence (for iconic and metaphoric gestures) or "gesture points to X" where X is something abstract or concrete or "gesture highlights important bits of information" or any combination of these.

   As previously mentioned, take into account that the gesture categories are not mutually exclusive. For now we record a gesture of several categories in the COMMENT field. This field should also be used in case of a doubt or uncertainty.


Figure 111: Add Track Element

The more information, the better. Click OK after entering all the relevant details.

6. PASS 6: Annotate beats. They can be performed either in isolation, e.g., flicking the hand up and down, or they can occur within a gesture phrase, e.g., both hands are moved to the front to depict a square and from that position they move up/down to emphasise essential bits of the spoken phrases. Duncan [Under perpetual revision] points out that a gesture labelled as iconic, metaphoric or else, and whose stroke coincides with a prosodic prominence, is also analysed as a beat. Beats are added in the usual way by clicking on the Beat track, marking its beginning and end and clicking on the END button from the track window. A new window pops up (see Figure 111). Click OK to record it. Continue to the end of the video file.

7. PASS 7: Start over again to verify that no gesture has been omitted, that the time interval of a single hand movement is equal to the sum of the temporal intervals of its phases and that everything has been annotated accordingly.

# Appendix B

# NXT Queries

## B.1 Depicting Gestures

1. The following query (without the new lines) searches the number of gestures strokes (excluding the adaptors) in Loehr's [2004] data:

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase):
        $gˆ$p & !($g@dimension="adaptor") & $p@type="stroke"
```

   Result: 95

   Note that in the results, each stroke gets a separately represented parent (in case of two strokes within a single gesture, this counts as two). If we want all the strokes per parent, then we'll need a complex query separated by double colons:

```
($g H1-gesture | H2-gesture)::($p H1-phase | H2-phase):
        $gˆ$p & !($g@dimension="adaptor") & $p@type="stroke"
```

   Result: 94

2. Query searching the number of strokes in Loehr's [2004] data overlapping at least one pitch accent (irrespective of whether low or high):

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($a accent): $gˆ$p & !($g@dimension="adaptor")
        & $p@type="stroke" & $p#$a & $a@tone˜/.+/
```

   Result: 71

3. Query searching the number of strokes in Loehr's [2004] data overlapping at least one accented word (not the accent itself)

```
($g H1-gesture | H2-gesture) ($p H1-phase | H2-phase)::
        ($w word)($a accent): $gˆ$p & !($g@dimension="adaptor")
        & $p@type="stroke" &
        $p#$w & $a>$w & $a@tone˜/.+/
```

Result: 79

4. Query counting the number of strokes overlapping an accented word with a fuzzy match of .275 msec

```
 for obs in L01 L02 L03 L04;
        do java -DNXT_FUZZINESS=0.275
        CountQueryResults -c gesture-prosody-meta.xml
        -o $obs -q '($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($w word)($a accent):
        $gˆ$p & !($g@dimension="adaptor") & $p@type="stroke"
        & $p#$w & $a>$w & $a@tone˜/.+/' ; done
```

Result: 91

5. Query searching strokes uniquely overlapping H* accent:

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($a accent)(forall $a2 accent):
        $gˆ$p & !($g@dimension="adaptor")
        & $p@type="stroke"
        & $a#$p & ($a2#$p -> $a==$a2 | $a2@tone!˜/.+/)
        & $a@tone="H*"
```

Result: 27

6. Query searching strokes uniquely overlapping X* accent:

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($a accent)(forall $a2 accent):
        $gˆ$p & !($g@dimension="adaptor") &
        $p@type="stroke"
        & $a#$p & ($a2#$p -> $a==$a2 | $a2@tone!˜/.+/)
        & $a@tone="X*"
```

Result: 14

7. Query searching strokes uniquely overlapping L+H* accent:

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($a accent)(forall $a2 accent):
        $g^$p & !($g@dimension="adaptor") &
        $p@type="stroke"
        & $a#$p & ($a2#$p -> $a==$a2 | $a2@tone!~/.+/)
        & $a@tone="L+H*"
```

Result: 9

8. Query searching strokes uniquely overlapping !H* accent:

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($a accent)(forall $a2 accent):
        $g^$p & !($g@dimension="adaptor") &
        $p@type="stroke"
        & $a#$p & ($a2#$p -> $a==$a2 | $a2@tone!~/.+/)
        & $a@tone="!H*"
```

Result: 5

9. Query searching strokes uniquely overlapping L* accent:

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($a accent)(forall $a2 accent):
        $g^$p & !($g@dimension="adaptor") &
        $p@type="stroke"
        & $a#$p & ($a2#$p -> $a==$a2 | $a2@tone!~/.+/)
        & $a@tone="L*"
```

Result: 3

10. Query searching strokes uniquely overlapping L+!H* accent:

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($a accent)(forall $a2 accent):
        $g^$p & !($g@dimension="adaptor") &
        $p@type="stroke"
        & $a#$p & ($a2#$p -> $a==$a2 | $a2@tone!~/.+/)
        & $a@tone="L+!H*"
```

Result: 2

11. Query searching for strokes overlapped uniquely by two accents of type H* and L* (note that the order of variable specification does not matter, and so this query will also find strokes overlapping an L*, H* sequence)

```
        ($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
                ($a accent)($a2 accent):!($g@dimension="adaptor")
                & $gˆ$p & $p@type="stroke" & $a@tone="H*" & $a2@tone="L*"
                & $a#$p & $a2#$p & $a!=$a2::
                        (forall $a3 accent):
                        $a3#$p -> ($a3==$a || $a3==$a2)
                        | ($a2@tone!~/.+/ || $a3@tone!~/.+/)
```

Result: 1

12. Query searching for strokes overlapped uniquely by two accents of type H* and H*

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
                ($a accent)($a2 accent):!($g@dimension="adaptor")
                & $gˆ$p & $p@type="stroke" & $a@tone="H*" & $a2@tone="H*"
                & $a#$p & $a2#$p & $a!=$a2::
                        (forall $a3 accent):
                        $a3#$p -> ($a3==$a || $a3==$a2)
                        | ($a2@tone!~/.+/ || $a3@tone!~/.+/)
```

Result: 3

13. Query searching for strokes overlapped uniquely by two accents of type H* and X*

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
                ($a accent)($a2 accent):!($g@dimension="adaptor")
                & $gˆ$p & $p@type="stroke" & $a@tone="H*" & $a2@tone="X*"
                & $a#$p & $a2#$p & $a!=$a2::
                        (forall $a3 accent):
                        $a3#$p -> ($a3==$a || $a3==$a2)
                        | ($a2@tone!~/.+/ || $a3@tone!~/.+/)
```

Result: 2

14. Query searching for strokes overlapped uniquely by two accents of type L+H* and X*

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
                ($a accent)($a2 accent):!($g@dimension="adaptor")
                & $gˆ$p & $p@type="stroke" & $a@tone="L+H*" & $a2@tone="X*"
                & $a#$p & $a2#$p & $a!=$a2::
                        (forall $a3 accent):
```

```
                $a3#$p -> ($a3==$a || $a3==$a2)
                | ($a2@tone!~/.+/ || $a3@tone!~/.+/)
```

Result: 1

15. Query searching for strokes overlapped uniquely by two accents of type L+H* and H*

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($a accent)($a2 accent):!($g@dimension="adaptor")
        & $gˆ$p & $p@type="stroke" & $a@tone="L+H*" & $a2@tone="H*"
        & $a#$p & $a2#$p & $a!=$a2::
                (forall $a3 accent):
                $a3#$p -> ($a3==$a || $a3==$a2)
                | ($a2@tone!~/.+/ || $a3@tone!~/.+/)
```

Result: 2

16. Query searching the strokes overlapped by an accent marked with its type (nuclear, non-nuclear or pre-nuclear)

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($a accent):$gˆ$p & !($g@dimension="adaptor")
        & $p@type="stroke"
        & $a#$p & $a@tone~/.*/
```

Result: 75

17. Query searching the strokes overlapped by a nuclear accent

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($a accent)(forall $a2 accent):
        $gˆ$p & !($g@dimension="adaptor")
        & $p@type="stroke" & $a#$p
        & ($a2#$p -> $a==$a2 | $a2@tone!~/.*/)
        & $a@type="nuclear"
```

Result: 38

18. Query searching the strokes overlapped by a pre-nuclear accent

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($a accent)(forall $a2 accent):
        $gˆ$p & !($g@dimension="adaptor")
```

```
        & $p@type="stroke" & $a#$p
        & ($a2#$p -> $a==$a2 | $a2@tone!~/.*/)
        & $a@type="pre-nuclear"
```

Result: 8

19. Query searching the strokes overlapped by a non-nuclear accent

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($a accent)(forall $a2 accent):
        $g^$p & !($g@dimension="adaptor")
        & $p@type="stroke" & $a#$p
        & ($a2#$p -> $a==$a2 | $a2@tone!~/.*/)
        & $a@type="non-nuclear"
```

Result: 15

20. Query searching the strokes overlapped by a combination of nuclear and non-nuclear accent

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($a accent)($a2 accent):!($g@dimension="adaptor")
        & $g^$p & $p@type="stroke"
        & $a@type="nuclear" & $a2@type="non-nuclear"
        & $a#$p & $a2#$p & $a!=$a2::
        (forall $a3 accent):$a3#$p ->
                ($a3==$a || $a3==$a2)
                | ($a2@type!~/.+/ || $a3@type!~/.+/)
```

Result: 8

21. Query searching the strokes overlapped by two nuclear accents

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($a accent)($a2 accent):!($g@dimension="adaptor")
        & $g^$p & $p@type="stroke"
        & $a@type="nuclear" & $a2@type="nuclear"
        & $a#$p & $a2#$p & $a!=$a2::
        (forall $a3 accent): $a3#$p ->
                ($a3==$a || $a3==$a2)
                | ($a2@type!~/.+/ || $a3@type!~/.+/)
```

Result: 1

22. Query searching the strokes overlapped by two non-nuclear accents

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($a accent)($a2 accent):!($g@dimension="adaptor")
        & $gˆ$p & $p@type="stroke"
        & $a@type="non-nuclear" & $a2@type="non-nuclear"
        & $a#$p & $a2#$p & $a!=$a2::
        (forall $a3 accent): $a3#$p ->
                ($a3==$a || $a3==$a2)
                | ($a2@type!~/.+/ || $a3@type!~/.+/)
```

Result: 1

## B.2 Deictic gestures

1. Query searching the total number of deictic gesture strokes

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase):
        $gˆ$p & !($g@dimension="adaptor")
        & $g@dimension~/.*eictic.*/
        & $p@type="stroke"
```

Result Talkbank: 82

Result AMI: 22

2. Query searching the total strokes overlapped by a pitch accent

```
 ($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)::
        ($a accent):
        $gˆ$p & $g@dimension~/.*deictic.*/
        & $p@type="stroke" & $p#$a
```

Result Talkbank: 74

Result AMI: 18

3. Query searching the total strokes overlapped by a pitch accented word

```
($g H1-gesture | H2-gesture) ($p H1-phase | H2-phase)::
        ($w word)($a accent): $gˆ$p & $g@dimension~/.*deictic.*/
        & $p@type="stroke"
        & $p#$w & $a>$w
```

Result Talkbank: 82

Result AMI: 22

4. Query searching the strokes overlapped by at least one nuclear or pre-nuclear accented word

```
($g H1-gesture | H2-gesture) ($p H1-phase | H2-phase)::
        ($w word)($a accent): $gˆ$p & $g@dimension~/.*deictic.*/
        & $p@type="stroke" &
        $p#$w & $a>$w
        & ($a@type="nuclear" || $a@type="pre-nuclear")
```

Result Talkbank: 82

Result AMI: 21

## B.3 Beats

1. Query searching total number of beats

```
($b beat)
```

Result in all observations: 36

2. Query searching number of beats overlapped by gesture strokes, excluding the adaptors

```
($g H1-gesture | H2-gesture)($p H1-phase | H2-phase)($b beat):
        $gˆ$p & !($g@dimension="adaptor") & $b#$p & $p@type="stroke"
```

Result in all observations: 4

# Appendix C

# Extending ERG with Gesture

## C.1 Chart-Mapping Rules

```
;;tmt.tdl needed

gesture-split-6-rule := chart_mapping_rule &
  [ +CONTEXT < speech_token & [ +FROM #from, +TO #mid1 ],
               speech_token & [ +FROM #mid2, +TO #mid3 ],
               speech_token & [ +FROM #mid4, +TO #mid5 ],
               speech_token & [ +FROM #mid6, +TO #mid7 ],
               speech_token & [ +FROM #mid8, +TO #mid9 ],
               speech_token & [ +FROM #mid10, +TO #to ] >,
    +INPUT < gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #from, +TO #to, +GESTURE #gesture ] >,
    +OUTPUT < gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #from, +TO #mid1, +GESTURE #gesture ],
             gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #mid2, +TO #mid3, +GESTURE #gesture ],
             gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #mid4, +TO #mid5, +GESTURE #gesture ],
```

```
              gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #mid6, +TO #mid7, +GESTURE #gesture ],
              gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #mid8, +TO #mid9, +GESTURE #gesture ],
              gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #mid10, +TO #to, +GESTURE #gesture] >,
    +POSITION "O1@C1, O2@C2, O3@C3, O4@C4, O5@C5, O6@C6, C1<C2, C2<C3, C3<C4, C5<C6" ].

gesture-split-4-rule := chart_mapping_rule &
  [ +CONTEXT < speech_token & [ +FROM #from, +TO #mid1 ],
               speech_token & [ +FROM #mid2, +TO #mid3 ],
               speech_token & [ +FROM #mid4, +TO #mid5 ],
               speech_token & [ +FROM #mid6, +TO #to ] >,
    +INPUT < gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #from, +TO #to, +GESTURE #gesture ] >,
    +OUTPUT < gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #from, +TO #mid1, +GESTURE #gesture ],
              gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #mid2, +TO #mid3, +GESTURE #gesture ],
              gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #mid4, +TO #mid5, +GESTURE #gesture ],
              gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #mid6, +TO #to, +GESTURE #gesture] >,
    +POSITION "O1@C1, O2@C2, O3@C3, O4@C4, C1<C2, C2<C3, C3<C4" ].

gesture-split-3-rule := chart_mapping_rule &
```

```
  [ +CONTEXT < speech_token & [ +FROM #from, +TO #mid1 ],
               speech_token & [ +FROM #mid2, +TO #mid3 ],
               speech_token & [ +FROM #mid4, +TO #to ] >,
    +INPUT < gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #from, +TO #to, +GESTURE #gesture ] >,
    +OUTPUT < gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #from, +TO #mid1, +GESTURE #gesture ],
             gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #mid2, +TO #mid3, +GESTURE #gesture ],
             gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #mid4, +TO #to, +GESTURE #gesture ] >,
    +POSITION "O1@C1, O2@C2, O3@C3, C1<C2, C2<C3" ].

gesture-split-2-rule := chart_mapping_rule &
  [ +CONTEXT < speech_token & [ +FROM #from, +TO #mid1 ],
               speech_token & [ +FROM #mid2, +TO #to ] >,
    +INPUT < gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #from, +TO #to, +GESTURE #gesture ] >,
    +OUTPUT < gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #from, +TO #mid1, +GESTURE #gesture ],
             gesture_token &
               [ +FORM #form, +CLASS #class, +TRAIT #trait,
                 +PRED #pred, +CARG #carg, +ID #id,
                 +FROM #mid2, +TO #to, +GESTURE #gesture ] >,
    +POSITION "O1@C1, O2@C2" ].

 gesture-unary-rule-1 := gesture-unary-rule &
  [ +CONTEXT < [ +GESTURE hand-formational,
                 +FROM #from,
                 +TO #to ] >,
```

```
    +INPUT < [ +FROM #from,
               +TO #to ] >,
    +POSITION "O1@I1" ].

gesture-unary-rule-2 := gesture-unary-rule &
  [ +CONTEXT < [ +GESTURE deictic-concrete ] >,
    +POSITION "O1@I1" ].

ditch_gesture_tmr := chart_mapping_rule &
   [ +INPUT < gesture_token >,
     +OUTPUT < > ].
```

## C.2 Grammar for Speech and Gesture

```
;---------------------------------------------------------------------
;; Additions to ERG types:

;; Add to basic_word to unify TOKENS gesture and prosody information in sign
;; (by convention, stored as attributes of ORTH)
;;
basic_word :+
  [ ORTH [ GESTURE #gesture, PROSODY #prosody ],
    TOKENS.+LIST < [ +GESTURE #gesture, +PHON #prosody ], ... > ].

;; Add to inflectional rule type and punctuation rule type to propagate
;; GESTURE and PROSODY features

lex_rule_infl_affixed :+
  [ ORTH [ GESTURE #gesture, PROSODY #prosody ],
    DTR.ORTH [ GESTURE #gesture, PROSODY #prosody ] ].

basic_punctuation_rule :+
  [ ORTH [ GESTURE #gesture, PROSODY #prosody ],
    ARGS < [ ORTH [ GESTURE #gesture, PROSODY #prosody ] ] > ].

;; Add to unary and binary phrases to force gesture-marked words to undergo
;; a gesture lexical rule which maps the token properties to semantics.

norm_unary_phrase :+
  [ ARGS < [ ORTH.GESTURE no-gesture ] > ].
```

```
binary_rule_left_to_right :+
  [ ARGS < [ ORTH.GESTURE no-gesture ],
           [ ORTH.GESTURE no-gesture ] > ].

binary_rule_right_to_left :+
   [ ARGS < [ ORTH.GESTURE no-gesture ],
            [ ORTH.GESTURE no-gesture ] > ].

;; Add to 'token' to include +PHON, +GESTURE, +START, +END attributes:
;;
token :+
[ +PHON pros,
  +GESTURE basic-gesture,
  +START string,
  +END string ].

;; End additions to existing ERG types
;---------------------------------------------------------------

;; Add three subtypes of token
;;
 speech_token := token &
   [ +GESTURE no-gesture ].

gesture_token := token.

speech+gesture_token := token.

;; Add subtype of the value of ORTH to introduce features GESTURE, PROSODY
;;
orthog+gesture := orthography &
  [ GESTURE basic-gesture,
    PROSODY pros ].

; ------------------------------------------------------------
; Prosodic types, largely based on Klein (2000)
; ------------------------------------------------------------

pros := *avm* &
        [ +DOM *top* ].
```

```
marked := pros.

unmarked := pros.

p-word := marked.

nucl_or_pre-nucl := p-word.

nuclear := nucl_or_pre-nucl.

pre-nuclear := nucl_or_pre-nucl.

non-nuclear := p-word.

mtr-tree := marked.

; ------------------------------------------------------------
; Gesture types
; ------------------------------------------------------------

;; DPF - Added supertype basic-gesture and its subtype no-gesture to make sure
;; the gesture lexical rules only apply to words with gesture marking.

basic-gesture := *avm*.

gesture-non-communicative := basic-gesture.

gesture-communicative := basic-gesture.

no-gesture := basic-gesture.

hand-formational := gesture-communicative.

1h_or_2h-symm := hand-formational &
[ HAND-SHAPE hand-shape,
  PALM-ORIENT orient,
  FINGER-ORIENT orient,
  HAND-LOCATION loc,
  HAND-MOVEMENT move ].

2h-non-symm := hand-formational &
[ R-HAND-SHAPE hand-shape,
```

```
  L-HAND-SHAPE hand-shape,
  R-PALM-ORIENT orient,
  L-PALM-ORIENT orient,
  R-FINGER-ORIENT orient,
  L-FINGER-ORIENT orient,
  R-HAND-LOCATION loc,
  L-HAND-LOCATION loc,
  R-HAND-MOVEMENT move,
  L-HAND-MOVEMENT move ].

1h := 1h_or_2h-symm.
2h-symm := 1h_or_2h-symm.
rh := 1h.
lh := 1h.

hand-functional := gesture-communicative.

depicting := hand-functional.
deictic := hand-functional.
beat := hand-functional.

depict-literal := depicting.
depict-metaphoric := depicting.

deictic-concrete := deictic.
deictic-abstr_or_nom := deictic.
deictic-abstract := deictic-abstr_or_nom.
deictic-nominating := deictic-abstr_or_nom.

depict-literal-meta := depict-literal & depict-metaphoric.
depict-literal-deictic-concr := depict-literal & deictic-concrete.
depict-literal-deictic-abstr := depict-literal & deictic-abstract.
depict-literal-deictic-nom := depict-literal & deictic-nominating.
depict-literal-beat := depict-literal & beat.

depict-meta-deictic-concr := depict-metaphoric & deictic-concrete.
depict-meta-deictic-abstr := depict-metaphoric & deictic-abstract.
depict-meta-deictic-nom := depict-metaphoric & deictic-nominating.
depict-meta-beat := depict-metaphoric & beat.

deictic-concr-beat := deictic-concrete & beat.
```

```
deictic-abstr-beat := deictic-abstract & beat.

deictic-nom-beat := deictic-nominating & beat.

rh-depict-literal := rh & depict-literal.
rh-depict-meta := rh & depict-metaphoric.
rh-deictic-concr := rh & deictic-concrete.
rh-deictic-abstr := rh & deictic-abstract.
rh-deictic-nom := rh & deictic-nominating.
rh-beat := rh & beat.
rh-depict-literal-meta := rh & depict-literal-meta.
rh-depict-literal-deictic-concr := rh & depict-literal-deictic-concr.
rh-depict-literal-deictic-abstr := rh & depict-literal-deictic-abstr.
rh-depict-literal-deictic-nom := rh & depict-literal-deictic-nom.
rh-depict-literal-beat := rh & depict-literal-beat.
rh-depict-meta-deictic-concr := rh & depict-meta-deictic-concr.
rh-depict-meta-deictic-abstr := rh & depict-meta-deictic-abstr.
rh-depict-meta-deictic-nom := rh & depict-meta-deictic-nom.
rh-depict-meta-beat := rh & depict-meta-beat.
rh-deictic-concr-beat := rh & deictic-concr-beat.
rh-deictic-abstr-beat := rh & deictic-abstr-beat.
rh-deictic-nom-beat := rh & deictic-nom-beat.

lh-depict-literal := lh & depict-literal.
lh-depict-meta := lh & depict-metaphoric.
lh-deictic-concr := lh & deictic-concrete.
lh-deictic-abstr := lh & deictic-abstract.
lh-deictic-nom := lh & deictic-nominating.
lh-beat := lh & beat.
lh-depict-literal-meta := lh & depict-literal-meta.
lh-depict-literal-deictic-concr := lh & depict-literal-deictic-concr.
lh-depict-literal-deictic-abstr := lh & depict-literal-deictic-abstr.
lh-depict-literal-deictic-nom := lh & depict-literal-deictic-nom.
lh-depict-literal-beat := lh & depict-literal-beat.
lh-depict-meta-deictic-concr := lh & depict-meta-deictic-concr.
lh-depict-meta-deictic-abstr := lh & depict-meta-deictic-abstr.
lh-depict-meta-deictic-nom := lh & depict-meta-deictic-nom.
lh-depict-meta-beat := lh & depict-meta-beat.
lh-deictic-concr-beat := lh & deictic-concr-beat.
lh-deictic-abstr-beat := lh & deictic-abstr-beat.
lh-deictic-nom-beat := lh & deictic-nom-beat.
```

```
2h-symm-depict-literal := 2h-symm & depict-literal.
2h-symm-depict-meta := 2h-symm & depict-metaphoric.
2h-symm-deictic-concr := 2h-symm & deictic-concrete.
2h-symm-deictic-abstr := 2h-symm & deictic-abstract.
2h-symm-deictic-nom := 2h-symm & deictic-nominating.
2h-symm-beat := 2h-symm & beat.
2h-symm-depict-literal-meta := 2h-symm & depict-literal-meta.
2h-symm-depict-literal-deictic-concr := 2h-symm & depict-literal-deictic-concr.
2h-symm-depict-literal-deictic-abstr := 2h-symm & depict-literal-deictic-abstr.
2h-symm-depict-literal-deictic-nom := 2h-symm & depict-literal-deictic-nom.
2h-symm-depict-literal-beat := 2h-symm & depict-literal-beat.
2h-symm-depict-meta-deictic-concr := 2h-symm & depict-meta-deictic-concr.
2h-symm-depict-meta-deictic-abstr := 2h-symm & depict-meta-deictic-abstr.
2h-symm-depict-meta-deictic-nom := 2h-symm & depict-meta-deictic-nom.
2h-symm-depict-meta-beat := 2h-symm & depict-meta-beat.
2h-symm-deictic-concr-beat := 2h-symm & deictic-concr-beat.
2h-symm-deictic-abstr-beat := 2h-symm & deictic-abstr-beat.
2h-symm-deictic-nom-beat := 2h-symm & deictic-nom-beat.


;; The values of orient are based on McNeill 1992.

orient := *sort*.
towards-body := orient.
towards-centre := orient.
towards-down := orient.
towards-up := orient.
away-centre := orient.
away-body := orient.

;; The values of hand-shape are based on Bressem 2008.

basic-hand-shape := *sort*.
hand-shape := basic-hand-shape.
finger-digit := basic-hand-shape.
finger-form := basic-hand-shape.
fist := hand-shape.
flat-hand := hand-shape.
finger := hand-shape &
          [ FINGER-DIGIT finger-digit,
            FINGER-FORM finger-form ].
single-finger := finger.
```

```
finger-combination := finger.

1f := finger-digit.
2f := finger-digit.
3f := finger-digit.
4f := finger-digit.
5f := finger-digit.

12f := 1f & 2f.
13f := 1f & 3f.
14f := 1f & 4f.
15f := 1f & 5f.

23f := 2f & 3f.
24f := 2f & 4f.
25f := 2f & 5f.

34f := 3f & 4f.
35f := 3f & 5f.

45f := 4f & 5f.

123f := 12f & 3f.
124f := 12f & 4f.
125f := 12f & 5f.

134f := 13f & 4f.
135f := 13f & 5f.

145f := 14f & 5f.

1234f := 123f & 4f.
1235f := 123f & 5f.

1245f := 124f & 5f.

12345f := 1234f & 5f.

stretched := finger-form.
bent := finger-form.
crooked := finger-form.
flapped-down := finger-form.
```

```
connected := finger-form.
touching := finger-form.

;The values for loc are based on the gesture space proposed by McNeill (1992).

loc := *sort*.

;The c_coord designates the tip of the index finger of the pointing hand
;(appropriate for deictic gestures). This is based on Lascardies & Stone
;(2009).

c_coord := loc.
centre := loc.
x-dim := loc.
y-dim := loc.

left := x-dim.
right := y-dim.

lower := y-dim.
upper := y-dim.

left-periphery := left.
extreme-left := left.
right-periphery := right.
extreme-right := right.

lower-periphery := lower.
extreme-low := lower.
upper-periphery := upper.
extreme-up := upper.

left-low := left-periphery & lower-periphery.
left-extreme-low := left-periphery & extreme-low.
left-up := left-periphery & upper-periphery.
left-extreme-up := left-periphery & extreme-up.

extreme-left-low := extreme-left & lower-periphery.
extreme-left-extreme-low := extreme-left & extreme-low.
extreme-left-up := extreme-left & upper-periphery.
extreme-left-extreme-up := extreme-left & extreme-up.
```

```
right-low := right-periphery & lower-periphery.
right-extreme-low := right-periphery & extreme-low.
right-up := right-periphery & upper-periphery.
right-extreme-up := right-periphery & extreme-up.

extreme-right-low := extreme-right & lower-periphery.
extreme-right-extreme-low := extreme-right & extreme-low.
extreme-right-up := extreme-right & upper-periphery.
extreme-right-extreme-up := extreme-right & extreme-up.

; The values for move are based on Bressem (2008).

basic-move := *sort*.

move := basic-move &
        [ TYPE move-type,
          DIRECTION move-direction ].

move-type := basic-move.
move-direction := basic-move.

arm-shoulder := move-type.
wrist := move-type.
fingers := move-type.

arm-straight := arm-shoulder.
arm-arc := arm-shoulder.
arm-circle := arm-shoulder.
spiral := arm-shoulder.
zigzarg := arm-shoulder.
s-line := arm-shoulder.

bending := wrist.
raising := wrist.
rotating := wrist.

finger-straight := fingers.
finger-arc := fingers.
finger-circle := fingers.
finger-beating := fingers.
flapping-down := fingers.
grabbibg := fingers.
```

```
closing := fingers.

horizontal := move-direction.
vertical := move-direction.
sagittal := move-direction.
diagonal := move-direction.

left-right := horizontal.
right-left := horizontal.

up-down := vertical.
down-up := vertical.

move-away-body := sagittal.
move-towards-body := sagittal.
away-body-clock := move-away-body.
away-body-anticlock := move-away-body.
towards-body-clock := move-towards-body.
towards-body-anticlock := move-towards-body.


; KSA: vis-relation is similar to subord-or-conj-relation: it takes two handles
; as arguments which correspond to the LTOPs of both S and G dtrs. The names
; explicitly define the order of S and G dtr in case this has semantic effects.
; Unlike conj-s, the value of the PRED attribute is not provided by the lexical
; entry but the lexical rule. I introduce the attribute M(ultimodal)-ARG which
; serves as the argument of the conjoined speech+gesture phrase, and so it can
; be taken as an argument by any external predicate.

vis-relation := relation &
[ M-ARG index,
  S-LBL handle,
  G-LBL handle ].

deictic_relation := arg0_relation &
  [ PRED deictic_rel,
    ARG0 event,
    S-ARG semarg,
    G-ARG semarg ].

; -----------------------------------------------------------
; predsorts
```

```
; ----------------------------------------------------------

vis_rel := predsort.
g_mod_rel := quant_rel.
deictic_rel := quant_rel.

; KSA: The gesture provides the spatial reference (sp_ref) of an object or
; event (event_or_index) located at the physical space that is denoted by the
; gesture: v_p_space, i.e., a function v maps the physical space identified by
; the gesture to the actually denoted space (v_p_space). This distinction is
; necessary as it allows us to accommodate abstract deictic gestures where the
; space identified by the gesture is not equal to what the gesture actually
; denotes. For consistency with ERG, every individual is bound by a quantifier
; and hence deictic_q binds the referent introduced by gesture. This analysis
; is based on Lascarides & Stone (2009).

sp_ref := predsort.
deictic_q := predsort.
v_p_space := predsort.

; ----------------------------------------------------------
; Lexical rule types
; ----------------------------------------------------------

gesture_lexrule := phrase_or_lexrule &
  [ INFLECTD +,
    ORTH [ FORM #form, FROM #from, TO #to, CLASS #class,
           GESTURE no-gesture, PROSODY #pros ],
    SYNSEM [ LOCAL [ CAT #cat,
                     CONT [ RELS [ LIST #rfirst,
                                   LAST #rlast ],
                            HCONS [ LIST #hcfirst,
                                    LAST #hclast ] ],
                     CONJ #conj,
                     AGR #agr ],
             NONLOC #nonloc,
             LEX #lex,
             MODIFD #modif,
             PHON #phon,
             PUNCT #punct,
             LKEYS #lkeys ],
    ARGS < [ INFLECTD +,
```

```
              ORTH [ FORM #form, FROM #from, TO #to, CLASS #class,
                     GESTURE hand-formational, PROSODY #pros ],
              SYNSEM [ LOCAL [ CAT #cat,
                               CONT [ HOOK.XARG #xarg,
                                      RELS [ LIST #rfirst,
                                             LAST #rmiddle ],
                                      HCONS [ LIST #hcfirst,
                                              LAST #hcmiddle ] ],
                               CONJ #conj,
                               AGR #agr ],
                       NONLOC #nonloc,
                       LEX #lex,
                       MODIFD #modif,
                       PUNCT #punct & [ LPUNCT no_punct,
                                        RPUNCT no_punct ],
                       PHON #phon,
                       LKEYS #lkeys ],
              IDIOM #idiom,
              DIALECT #dialect,
              GENRE #genre,
              KEY-ARG #keyarg ] >,
    C-CONT [ HOOK.XARG #xarg,
             RELS [ LIST #rmiddle,
                    LAST #rlast ],
             HCONS [ LIST #hcmiddle,
                     LAST #hclast ] ],
    IDIOM #idiom,
    DIALECT #dialect,
    GENRE #genre,
    KEY-ARG #keyarg ].

depicting_gesture_lexrule := gesture_lexrule &
  [ ARGS < [ SYNSEM.LOCAL.CONT.RELS.LIST.FIRST.LBL #dltop,
             ORTH [ GESTURE depicting,
                    PROSODY nucl_or_pre-nucl ]] >,
    C-CONT [ HOOK [ LTOP #ltop,
                    INDEX #index ],
             RELS <! vis-relation &
                   [ PRED vis_rel,
                     LBL #ltop,
                     M-ARG #index,
                     S-LBL #arg1,
```

```
                    G-LBL #arg2 ],
                  [ PRED g_mod_rel,
                    LBL #glbl,
                    ARG1 #harg ],
                  [ LBL #larg1 ],
                  [ LBL #larg2 ],
                  [ LBL #larg3 ],
                  [ LBL #larg4 ],
                  [ LBL #larg5 ],
                  [ LBL #larg6 ]!>,
             HCONS <! qeq & [ HARG #arg1, LARG #dltop ],
                     qeq & [ HARG #arg2, LARG #glbl ],
                     qeq & [ HARG #harg, LARG #larg1 ],
                     qeq & [ HARG #harg, LARG #larg2 ],
                     qeq & [ HARG #harg, LARG #larg3 ],
                     qeq & [ HARG #harg, LARG #larg4 ],
                     qeq & [ HARG #harg, LARG #larg5 ],
                     qeq & [ HARG #harg, LARG #larg6 ] !> ] ].

;; Add an "outscopes" handle constraint type (‘greater than or equal to’)
;; to express the relation between vis_rel’s S-LBL and the gesture-marked
;; sign, to support underspecification of the range of the gesture’s
;; interpretation for varying phrases containing the gesture-marked sign.

basic_deixis_lexrule := gesture_lexrule &
 [ ARGS < [ SYNSEM.LOCAL.CONT [ HOOK.LTOP #ltop ],
            ORTH.GESTURE deictic ]  >,
   C-CONT [ HOOK [ LTOP #ltop ] ] ].

;; KSA: The number of the relations in RELS cannot be underspecified. We
;; therefore need separate rules for gestures of 6 relations and for gestures
;; of 7 relations.

deixis_lexrule_6_rel := basic_deixis_lexrule &
 [ ARGS < [ SYNSEM.LOCAL.CONT [ RELS.LIST.FIRST.ARG0 #sarg ] ] >,
   C-CONT [ HOOK [ LTOP #ltop ],
            RELS <! deictic_relation &
                  [ PRED deictic_rel,
                    LBL #ltop,
                    S-ARG #sarg,
                    G-ARG #garg ],
                  [ PRED deictic_q,
```

```
                    ARG0 #garg,
                    RSTR #rstr ],
                  [ PRED sp_ref,
                    LBL #sp-lbl,
                    ARG0 event_or_index & #garg,
                    ARG1 v_p_space ],
                  [ LBL #sp-lbl,
                    ARG0 event,
                    ARG1 #garg ],
                  [ LBL #sp-lbl,
                    ARG0 event,
                    ARG1 #garg ],
                  [ LBL #sp-lbl,
                    ARG0 event,
                    ARG1 #garg ],
                  [ LBL #sp-lbl,
                    ARG0 event,
                    ARG1 #garg ],
                  [ LBL #sp-lbl,
                    ARG0 event,
                    ARG1 #garg ],
                  [ LBL #sp-lbl,
                    ARG0 event,
                    ARG1 #garg ],
                   [ LBL #sp-lbl,
                    ARG0 event,
                    ARG1 #garg ] !>,
            HCONS <! qeq & [ HARG #rstr, LARG #sp-lbl ] !> ] ].

deixis_lexrule_7_rel := basic_deixis_lexrule &
[ ARGS < [ SYNSEM.LOCAL.CONT [ RELS.LIST.FIRST.ARG0 #sarg ] ] >,
   C-CONT [ HOOK [ LTOP #ltop ],
            RELS <! deictic_relation &
                  [ PRED deictic_rel,
                    LBL #ltop,
                    S-ARG #sarg,
                    G-ARG #garg ],
                  [ PRED deictic_q,
                    ARG0 #garg,
                    RSTR #rstr ],
                  [ PRED sp_ref,
                    LBL #sp-lbl,
                    ARG0 event_or_index & #garg,
                    ARG1 v_p_space ],
                  [ LBL #sp-lbl,
```

```
                    ARG0 event,
                    ARG1 #garg ],
                  [ LBL #sp-lbl,
                    ARG0 event,
                    ARG1 #garg ],
                  [ LBL #sp-lbl,
                    ARG0 event,
                    ARG1 #garg ],
                  [ LBL #sp-lbl,
                    ARG0 event,
                    ARG1 #garg ],
                  [ LBL #sp-lbl,
                    ARG0 event,
                    ARG1 #garg ],
                   [ LBL #sp-lbl,
                    ARG0 event,
                    ARG1 #garg ],
                   [ LBL #sp-lbl,
                    ARG0 event,
                    ARG1 #garg ] !>,
            HCONS <! qeq & [ HARG #rstr, LARG #sp-lbl ] !> ] ].

abstract_deixis_lexrule := gesture_lexrule &
 [ ARGS < [ ORTH [ GESTURE deictic-abstr_or_nom,
                   PROSODY nucl_or_pre-nucl ] ] > ].

concrete_deixis_lexrule := gesture_lexrule &
  [ ARGS < [ ORTH [ GESTURE deictic-concrete,
                    PROSODY pros ] ] > ].

concrete_deixis_lexrule_6_rel := concrete_deixis_lexrule & deixis_lexrule_6_rel.
concrete_deixis_lexrule_7_rel := concrete_deixis_lexrule & deixis_lexrule_7_rel.
abstract_deixis_lexrule_6_rel := abstract_deixis_lexrule & deixis_lexrule_6_rel.
abstract_deixis_lexrule_7_rel := abstract_deixis_lexrule & deixis_lexrule_7_rel.

geq := scp_pr & [ HARG.INSTLOC #1, LARG.INSTLOC #1 ].
```

## C.3 Lexical Rules for Gesture Types

```
eating_depict_g := depicting_gesture_lexrule &
```

```
[ ARGS < [ TOKENS.+LIST < [ +GESTURE rh-depict-literal &
                                    [ HAND-SHAPE flat-hand,
                                      PALM-ORIENT towards-body,
                                      FINGER-ORIENT away-centre,
                                      HAND-LOCATION centre,
                                      HAND-MOVEMENT [ TYPE arm-straight,
                                                      DIRECTION away-body-clock ] ] ] > ] >,
  C-CONT.RELS <! relation, relation,
               [ PRED "hand-shape-flat-hand_rel" ],
               [ PRED "palm-orient-towards-body_rel" ],
               [ PRED "finger-orient-away-centre_rel" ],
               [ PRED "hand-location_centre_rel" ],
               [ PRED "hand-movement-type_arm_straight_rel" ],
               [ PRED "hand-movement-direction-away-body-clock_rel" ]!> ].


eating_abstract_g := abstract_deixis_lexrule_6_rel &
 [ ARGS < [ TOKENS.+LIST <[ +GESTURE  rh-deictic-abstr &
                                    [ HAND-SHAPE flat-hand,
                                      PALM-ORIENT away-centre,
                                      FINGER-ORIENT away-body,
                                      HAND-LOCATION c_coord,
                                      HAND-MOVEMENT [ TYPE arm-straight,
                                                      DIRECTION move-away-body ] ] ] > ] >,
   C-CONT.RELS <! relation, relation, relation,
                [ PRED "hand-shape-flat-hand_rel" ],
                [ PRED "palm-orient-away-centre_rel" ],
                [ PRED "finger-orient-away-body_rel" ],
                [ PRED "hand-location-c-coord_rel" ],
                [ PRED "hand-movement-type-arm-straight_rel" ],
                [ PRED "hand-movement-direction-move-away-body_rel" ] !> ].


eating_concrete_g := concrete_deixis_lexrule_7_rel &
[ ARGS < [ TOKENS.+LIST <[ +GESTURE rh-deictic-concr &
                                   [  HAND-SHAPE [ FINGER-DIGIT 1f,
                                                   FINGER-FORM stretched ],
                                      PALM-ORIENT away-centre,
                                      FINGER-ORIENT away-body,
                                      HAND-LOCATION c_coord,
                                      HAND-MOVEMENT [ TYPE finger-straight,
                                                      DIRECTION move-away-body ] ] ] > ] >,
```

```
C-CONT.RELS <! relation, relation, relation,
              [ PRED "hand-shape-finger-digit-1f_rel" ],
              [ PRED "hand-shape-finger-form-streched_rel" ],
              [ PRED "palm-orient-away-centre_rel" ],
              [ PRED "finger-orient-away-body_rel" ],
              [ PRED "hand-location-c-coord_rel" ],
              [ PRED "hand-movement-type-finger-straight_rel" ],
              [ PRED "hand-movement-direction-move-away-body_rel" ] !> ].
```

# References

[Adolphs2009] Adolphs, Peter. 2009. Tutorial: Chart-mapping in PET. Presented at the fifth DELPH-IN Summit, Barcelona. http://www.delph-in.net/2009/cm.pdf.

[Adolphs et al.2008] Adolphs, Peter, 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.

[Alahverdzhieva, Flickinger, and Lascarides2012] Alahverdzhieva, Katya, Dan Flickinger, and Alex Lascarides. 2012. Multimodal grammar implementation. In *Proceedings of Annual Meeting of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies (NAACL-HLT 2012)*, Montreal, Canada.

[Alahverdzhieva and Lascarides2010] Alahverdzhieva, Katya and Alex Lascarides. 2010. Analysing speech and co-speech gesture in constraint-based grammars. In Stefan Müller, editor, *The Proceedings of the 17th International Conference on Head-Driven Phrase Structure Grammar*, pages 6–26, Stanford. CSLI Publications.

[Alahverdzhieva and Lascarides2011a] Alahverdzhieva, Katya and Alex Lascarides. 2011a. An hpsg approach to synchronous speech and deixis. In Stefan Müller, editor, *Proceedings of the 18th International Conference on Head-Driven Phase Structure Grammar (HPSG)*, pages 6–24, Seattle. CSLI Publications.

[Alahverdzhieva and Lascarides2011b] Alahverdzhieva, Katya and Alex Lascarides. 2011b. Integration of speech and deictic gesture in a multimodal grammar. In *Proceedings of Traitement Automatique de Langues Naturelles (TALN 2011)*, Montpellier, France.

[Alahverdzhieva and Lascarides2011c] Alahverdzhieva, Katya and Alex Lascarides. 2011c. Semantic composition of multimodal actions in constraint-based grammars. In *Proceedings of Constraints in Discourse (CID) 2011*, Agay-Roches Rouges, Var, France.

[Alibali, Kita, and Young2000] Alibali, Martha, Sotaro Kita, and Amanda J. Young. 2000. Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes*, 15:593–613(21).

[Alshawi1992] Alshawi, Hiyan. 1992. *The Core Language Engine*. Cambridge: MIT Press.

[Asher and Lascarides2003] Asher, Nicholas and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

[Bangalore and Johnston2000] Bangalore, Srinivas and Michael Johnston. 2000. Integrating multimodal language processing with speech recognition. In *INTERSPEECH*, pages 126–129. ISCA.

[Bangalore and Johnston2009] Bangalore, Srinivas and Michael Johnston. 2009. Robust understanding in multimodal interfaces. *Computational Linguistics*, 35(3):345–397.

[Baumann2006] Baumann, Stefan. 2006. *The intonation of givenness - evidence from German*. Linguistische Arbeiten 508. Niemeyer, Tübingen.

[Bavelas et al.2008] Bavelas, Janet, Jennifer Gerwing, Chantelle Sutton, and Danielle Prevost. 2008. Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58:495–520.

[Bavelas et al.2002] Bavelas, Janet, Christine Kenwood, Trudy Johnson, and B Phillips. 2002. An experimental study of when and how speakers use gestures to communicate. *Gesture*, 2(1):1–17.

[Bavelas and Chovil2000] Bavelas, Janet Beavin and Nicole Chovil. 2000. Visible acts of meaning. An integrated message model of language use in face-to-face dialogue. *Journal of Language and Social Psychology*, (19):163–194.

[Bavelas and Chovil2006] Bavelas, Janet Beavin 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*. Thousand Oaks, CA: Sage, pages 97–115.

[Bavelas et al.1995] Bavelas, Janet Beavin, Nicole Chovil, Linda Coates, and Lori Roe. 1995. Gestures specialized for dialogue. *Personality and Social Psychology Bulletin*, 21:394–405.

[Beckman and Elam1997] Beckman, Mary E. and Gayle Ayers Elam. 1997. *Guidelines for ToBI labelling*. The Ohio State University Research Foundation. Version 3 available from `http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/`.

[Bender and Oepen2002] Bender, Emily M., Dan Flickinger and Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan.

[Bergmann and Kopp2007] Bergmann, Kirsten and Stefan Kopp. 2007. Coexpressivity of speech and gesture: lessons for models of aligned speech and gesture production. In *Proceedings of the AISB'07 Symposium Language, Speech and Gesture for Expressive Characters*, Newcastle, UK.

[Boersma and Weenink2003] Boersma, Paul and David Weenink. 2003. *'Praat:doing phonetics by computer'*. `http://www.praat.org`.

[Bögel et al.2009] Bögel, Tina, Miriam Butt, Ronald M.n Kaplan, Tracy Holloway King, and John T. Maxwell. 2009. Prosodic phonology in lfg: a new proposal. In *Proceedings of the LFG 2009 Conference held 2009 July 13-16 at Trinity College, Cambridge, UK*, Stanford, CA. CSLI Publications.

[Bolinger1986] Bolinger, Dwight. 1986. *Intonation and Its Parts: Melody in Spoken English*. Stanford University Press, Stanford.

[Bos2004] Bos, Johan. 2004. Computational semantics in discourse: Underspecification, resolution, and inference. *J. of Logic, Lang. and Inf.*, 13(2):139–157, March.

[Brenier and Calhoun2006] Brenier, Jason and Sasha Calhoun. 2006. Switchboard prosody annotation scheme. Department of Linguistics, Stanford University and ICCS, University of Edinburgh. Internal publication.

[Bressem2007] Bressem, Jana. 2007. Recurrent Form Features in Coverbal Gestures. Unpublished Manuscript.

[Bressem2008] Bressem, Jana. 2008. Notating gestures – Proposal for a form based notation system of coverbal gestures. Unpublished Manuscript.

[Buisine and Martin2007] Buisine, Stéphanie and Jean-Claude Martin. 2007. The effects of speech-gesture cooperation in animated agents' behavior in multimedia presentations. *Interact. Comput.*, 19(4):484–493.

[Bulwer1644] Bulwer, John. 1644. *Chirologia or the Natural Language of the Hand*.

[Bunt2007] Bunt, Harry. 2007. Semantic underspecification: Which technique for what purpose? In *Computing Meaning*, volume 83. Springer Netherlands, pages 55–85.

[Butt and King1998] Butt, Miriam and Tracy Holloway King. 1998. Interfacing phonology with lfg. In *Proceedings of the LFG98 Conference*.

[Calhoun2006] Calhoun, Sasha. 2006. *Information Structure and the Prosodic Structure of English: a Probabilistic Relationship*. University of Edinburgh. PhD Thesis.

[Calhoun2010] Calhoun, Sasha. 2010. The Centrality of Metrical Structure in Signaling Information Structure: A Probabilistic Perspective. *Language*, 86(1):1–42.

[Callmeier2000] Callmeier, Ulrich. 2000. PET — A platform for experimentation with efficient HPSG processing techniques. *Natural Language Engineering*, 6 (1) (Special Issue on Efficient Processing with HPSG):99 – 108.

[Carletta1996] Carletta, Jean. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22:249–254, June.

[Carletta2006] Carletta, Jean. 2006. Announcing the AMI meeting corpus. *The ELRA Newsletter*, 11(1):3–5, January.

[Carletta et al.2005] Carletta, Jean, Stefan Evert, Ulrich Heid, and Jonathan Kilgour. 2005. The nite xml toolkit: Data model and query language. *Language Resources and Evaluation*, 39:313–334.

[Cassell2000] Cassell, Justine. 2000. Nudge nudge wink wink: elements of face-to-face conversation for embodied conversational agents. *Embodied conversational agents*, pages 1–27.

[Cassell et al.1998] Cassell, Justine, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1998. Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. pages 582–591.

[Cassell, Stone, and Yan2000] Cassell, Justine, Matthew Stone, and Hao Yan. 2000. Coordination and context-dependence in the generation of embodied conversation. In *Proceedings of the International Natural Language Generation Conference*, pages 171–178, Mitzpe Ramon, Israel.

[Cassell, Vilhjálmsson, and Bickmore2001] Cassell, Justine, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2001. Beat: The behavior expression animation toolkit. In *Proceedings of SIGGRAPH 01*, Los Angeles.

[Clark1996] Clark, Herbert H. 1996. *Using Language*. Cambridge University Press, Cambridge.

[Cohen et al.1997] Cohen, Philip R., Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen, and Josh Clow. 1997. Quickset: multimodal interaction for distributed applications. In *Proceedings of the fifth ACM international conference on Multimedia*, MULTIMEDIA '97, pages 31–40, New York, NY, USA. ACM.

[Copestake2002] Copestake, Ann. 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford, CA.

[Copestake2007] 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 and Briscoe1995] Copestake, Ann and Ted Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of Semantics*, 12:15–67.

[Copestake and Flickinger2000] Copestake, Ann and Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second Linguistic Resources and Evaluation Conference*, pages 591 – 600, Athens, Greece.

[Copestake et al.2005] Copestake, Ann, 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.

[Copestake, Lascarides, and Flickinger2001] Copestake, Ann, 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)*, pages 132–139, Toulouse.

[De Jorio1832] De Jorio, Andrea. 1832. *La mimica degli antichi investigata nel gestire napoletano*. Naples.

[De Ruiter1998] De Ruiter, Jan Peter. 1998. *Gesture and speech production*. Doctoral dissertation at Catholic University of Nijmegen, Netherlands.

[De Ruiter2000] De Ruiter, Jan Peter. 2000. The production of gesture and speech. In David McNeill, editor, *Language and Gesture: Window into Thought and Action*. Cambridge: Cambridge University Press, pages 284–311.

[De Ruiter, Bangerter, and Dings2012] De Ruiter, Jan Peter, Adrian Bangerter, and Paula Dings. 2012. The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science*, 4:232–248.

[Dipper, Goetze, and Skopeteas2007] Dipper, Stefanie, Michael Goetze, and Stavros Skopeteas. 2007. *Information structure in cross-linguistic corpora: Annotation guidelines for phonology, morphology, syntax, semantics and information structure*. Universitätsverlag Potsdam, Potsdam.

[Doherty-Sneddon2003] Doherty-Sneddon, Gwyneth. 2003. *Children's Unspoken Language*. Jessica Kingsley Publishers.

[DuncanUnder perpetual revision] Duncan, Susan. Under perpetual revision. *CODING MANUAL*. University of Chicago.

[Duncan and McNeill2000] Duncan, Susan and David McNeill. 2000. Growth points in thinking-for-speaking. In David McNeill, editor, *Language and Gesture*. Cambridge University Press, Cambridge.

[Ebert, Evert, and Wilmes2011] Ebert, Cornelia, Stefan Evert, and Katharina Wilmes. 2011. Focus marking via gestures. In I. Reich et al., editor, *Proceedings of Sinn & Bedeutung 15*, Saarbrücken, Germany. Universaar - Saarland University Press.

[Egg, Koller, and Niehren2001] Egg, Markus, Alexander Koller, and Joachim Niehren. 2001. The constraint language for lambda structures. *Journal of Logic, Language and Information*, 10:457–485, September.

[Engle2000] Engle, Randi. 2000. *Toward a Theory of Multimodal Communication: Combining Speech, Gestures, Diagrams and Demonstrations in Structural Explanations*. Stanford University. PhD thesis.

[Firbas1992] Firbas, Jan. 1992. *Functional Sentence Perspective in Written and Spoken Communication*. Cambridge University Press, Cambridge.

[Flickinger2000] Flickinger, Dan. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6 (1) (Special Issue on Efficient Processing with HPSG):15 – 28.

[Flickinger, Bender, and Oepen2003] Flickinger, Dan, Emily M. Bender, and Stephan Oepen. 2003. Mrs in the lingo grammar matrix: A practical user's guide. Technical report.

[Fricke2008] 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.).

[Gardent2008] Gardent, Claire. 2008. Integrating a unification-based semantics in a large scale Lexicalised Tree-Adjoining Grammar for French. In *Proceedings of the 22nd International Conference on Computational Linguistics – Volume 1*, COLING '08, pages 249–256, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Giorgolo and Asudeh2011] Giorgolo, Gianluca and Ash Asudeh. 2011. Multimodal Communication in LFG: Gestures and the Correspondence Architecture. In *Proceedings of the LFG11 Conference*, pages 257–277, Stanford, CA: CSLI Publications.

[Giorgolo and Verstraten2008] Giorgolo, Gianluca and Frans Verstraten. 2008. Perception of speech-and-gesture integration. In *Proceedings of the International Conference on Auditory-Visual Speech Processing 2008*. pages 31–36.

[Giuliani and Knoll2007] Giuliani, Manuel and Alois Knoll. 2007. Integrating multimodal cues using grammar based models. In *HCI (6)*, pages 858–867.

[Givón1985] Givón, Talmy. 1985. Iconicity, Isomorphism and Non-arbitrary Coding in Syntax. In John Haiman, editor, *Iconicity in Syntax*. John Benjamins, Amsterdam, pages 187–219.

[Goffman1963] Goffman, Erving. 1963. *Behavior in Public Places: Notes on the Social Organization of Gatherings*. The Free Press.

[Goldin-Meadow et al.2008] Goldin-Meadow, Susan, Wing Chee So, Asli Özyürek, and Carolyn Mylander. 2008. The natural order of events: How speakers of different languages represent events nonverbally. In *Proceedings of the National Academy of Sciences*, volume 105, July.

[Grice1975] Grice, H. P. 1975. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*. Academic Press, San Diego, CA, pages 41–58.

[Grosz and Sidner1986] Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12:175–204.

[Hadar and Butterworth1997] Hadar, Uri and Brian Butterworth. 1997. Iconic gestures, imagery, and word retrieval in speech. *Semiotica*, 115:147–172.

[Hahn and Rieser2010] Hahn, Florian 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*. Polish Society for Cognitive Science, Poznań, pages 99–109.

[Haji-Abdolhosseini2003] Haji-Abdolhosseini, Mohammad. 2003. A constraint-based approach to information structure and prosody correspondence. In Stefan Müller, editor, *Proceedings of the HPSG-2003 Conference, Michigan State University, East Lansing*, pages 143–162. CSLI Publications. `http://cslipublications.stanford.edu/HPSG/4/`.

[Harrison2010] Harrison, Simon. 2010. Evidence for node and scope of negation on coverbal gesture. *Gesture*, 10(1):29–51.

[Helm, Marruitt, and Odersky1991] Helm, Richard, Kim Marruitt, and Martin Odersky. 1991. Building visual language parsers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '91, pages 105–112, New York, NY, USA. ACM.

[Hobbs1985] Hobbs, Jerry R. 1985. On the Coherence and Structure of Discourse. Technical report csli-85-37, Stanford University, Center for the Study of Language and Information, October.

[Hobbs, Stickel, and Martin1993] Hobbs, Jerry R., Mark Stickel, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63:69–142.

[Hockenmaier and Steedman2005] Hockenmaier, Julia and Mark Steedman. 2005. Ccgbank: Users manual. Technical report ms-cis-05-09, Department of Computer and Information Science, University of Pennsylvania.

[Holler and Beattie2007] Holler, Judith and Geoffry Beattie. 2007. Gesture use in social interaction: how speakers' gestures can reflect listeners' thinking. In *Proceedings of the 2nd Conference of the International Society of Gesture Studies Lyon, France 15-18 June 2005*.

[Hostetter, Alibali, and Kita2007] Hostetter, Autumn, Martha Alibali, and Sotaro Kita. 2007. I see it in my hands' eye: Representational gestures reflect conceptual demands. 22:313–336(24).

[Johnston1998a] Johnston, Michael. 1998a. Multimodal language processing. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia.

[Johnston1998b] Johnston, Michael. 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, pages 624–630, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Johnston2000] Johnston, Michael. 2000. Deixis and conjunction in multimodal systems. In *Proceedings of the 18th conference on Computational linguistics - Volume 1*, COLING '00, pages 362–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Johnston and Bangalore2000] Johnston, Michael and Srivinas Bangalore. 2000. Finite-state multimodal parsing and understanding. In *Proceedings of COLING-2000*, pages 369–375, Saarbrücken, Germany.

[Johnston et al.1997] Johnston, Michael, 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*, pages 281–288, Somerset, New Jersey. Association for Computational Linguistics.

[Kamp and Reyle1993] Kamp, Hans and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers, Dordrecht, NL.

[Kehler2002] Kehler, Andrew. 2002. *Coherence, Reference, and the Theory of Grammar*. CSLI Publications.

[Kendon1972] Kendon, Adam. 1972. Some relationships between body motion and speech. In A. Seigman and B. Pope, editors, *Studies in Dyadic Communication*. Pergamon Press, Elmsford, New York, pages 177–216.

[Kendon1980] Kendon, Adam. 1980. Gesticulation and speech: Two aspects of the process of utterance. *The relationship between verbal and nonverbal communication*.

[Kendon1988] Kendon, Adam. 1988. How gestures can become like words. *Cross-Cultural Perspectives in Nonverbal Communication*, pages 131–141. New York: C. J. Hogrefe.

[Kendon2004] Kendon, Adam. 2004. *Gesture. Visible Action as Utterance*. Cambridge University Press, Cambridge.

[Kipp2001] Kipp, Michael. 2001. Anvil — a generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg, September. Georgetown University.

[Kipp2008] Kipp, Michael. 2008. Anvil 5.0: user's manual. `http://www.anvil-software.de`.

[Kita1990] Kita, Sotaro. 1990. *The temporal relationship between gesture and speech: a study of Japanese-English bilinguals*. Department of Psychology, University of Chicago. Masters Thesis.

[Kita et al.2007] Kita, Sotaro, Asli Özyŕek, Shanley Allen, Amanda Brown, Reyhan Furman, and Tomoko Ishizuka. 2007. Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. *Language and Cognitive Processes*, 22(8):1212–1236.

[Kita and Özyürek2003] Kita, Sotaro and Asli Özyürek. 2003. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48:16–32.

[Kita, van Gijn, and van der Hulst1998] Kita, Sotaro, Ingeborg van Gijn, and Harry van der Hulst. 1998. Movement phases in signs and co-speech gestures, and their transcription by human coders. In Ipke Wachsmuth and Martin Frhlich, editors, *Gesture and Sign Language in Human-Computer Interaction*, volume 1371 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, pages 23–35. 10.1007/BFb0052986.

[Klein and Manning2003] Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Klein2000a] Klein, Ewan. 2000a. A constraint-based approach to english prosodic constituents. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 217–224, Morristown, NJ, USA. Association for Computational Linguistics.

[Klein2000b] Klein, Ewan. 2000b. Prosodic constituency in hpsg. In *Grammatical Interfaces in HPSG, Studies in Constraint-Based Lexicalism*, pages 169–200. CSLI Publications.

[Koller, Regneri, and Thater2008] Koller, Alexander, 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.

[Kopp, Tepper, and Cassell2004] Kopp, Stefan, 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*, pages 97–104, New York, NY, USA. State College, PA, USA, ACM.

[Kopp and Wachsmuth2004] Kopp, Stefan and Ipke Wachsmuth. 2004. Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15(1):39–52.

[Kranstedt et al.2006] Kranstedt, Alfred, Andy Lücking, Thies Pfeiffer, Hannes Rieser, and Ipke Wachsmuth. 2006. Deixis: How to determine demonstrated objects using a pointing cone. In Sylvie Gibet, Nicolas Courty, and Jean-Franois Kamp, editors,

*Gesture in Human-Computer Interaction and Simulation*, volume 3881 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, pages 300–311.

[Krauss, Chen, and Gottesman2000] Krauss, Robert M., Yihsiu Chen, and Rebecca F. Gottesman. 2000. Lexical gestures and lexical access: A process model. In David McNeill, editor, *Language and Gesture*. New York: Cambridge University Press, pages 261–283.

[Krauss and Hadar1999] Krauss, Robert M. and Uri Hadar. 1999. The role of speech-related arm/hand gestures in word retrieval. In L. Messing and R. Campbell, editors, *Gesture, Speech and Sign*. New York: Oxford University Press.

[Kruijff-Korbayová and Steedman2003] Kruijff-Korbayová, Ivana and Mark Steedman. 2003. Discourse and information structure. *Journal of Logic, Language and Information*, 12:249–259.

[Kühnlein, Nimke, and Stegmann2002] Kühnlein, Peter, 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.

[Ladd1996] Ladd, Robert D. 1996. *Intonational Phonology (first edition)*. Cambridge University Press.

[Ladd2008] Ladd, Robert D. 2008. *Intonational Phonology (second edition)*. Cambridge University Press.

[Ladewig2010] Ladewig, Silva. 2010. "It has a certain [gesture]" – Syntactic integration of gestures into speech. In *Proceedings of the Fourth Conference of the International Society for Gesture Studies (ISGS)*, Frankfurt an der Oder.

[Lakin1987] Lakin, Fred. 1987. Visual grammars for visual languages. In *Proceedings of the sixth National conference on Artificial intelligence - Volume 2*, AAAI'87, pages 683–688. AAAI Press.

[Lakoff and Johnson1980] Lakoff, George and Mark Johnson. 1980. *Metaphors We Live By*. The University of Chicago Press, Chicago and London.

[Lascarides and Stone2006] Lascarides, Alex 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)*, pages 125–132, Potsdam, Germany. Universitätsverlag Potsdam.

[Lascarides and Stone2009a] Lascarides, Alex and Matthew Stone. 2009a. Discourse coherence and gesture interpretation. *Gesture*, 9(2):147–180.

[Lascarides and Stone2009b] Lascarides, Alex and Matthew Stone. 2009b. A formal semantic analysis of gesture. *Journal of Semantics*.

[Lausberg and Sloetjes2009] Lausberg, Hedda and Han Sloetjes. 2009. Coding gestural behavior with the neuroges-elan system. *Behavior Research Methods*, 41(3):841–849.

[Levelt and Melinger2004] Levelt, Willem J. M. and Alissa Melinger. 2004. Gesture and the communicative intention of the speaker. *Gesture*, 4(2):119–141.

[Liberman and Prince1977] Liberman, Mark and Alan Prince. 1977. On stress and linguistic rhythm. *Linguistic Inquiry*, 8(2):249–336.

[Loehr2004] Loehr, Daniel. 2004. *Gesture and Intonation*. Georgetown University, Washington DC. Doctoral Dissertation.

[Lücking, Rieser, and Staudacher2006] Lücking, Andy, Hannes Rieser, and Marc Staudacher. 2006. 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*, pages 106–113, Potsdam, 9. Universitätsverlag Potsdam.

[Marcu and Echihabi2002] Marcu, Daniel and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 368–375, Stroudsburg, PA, USA. Association for Computational Linguistics.

[McClave1991] McClave, Evelyn. 1991. *Intonation and Gesture*. Georgetown University, Washington DC. Doctoral Dissertation.

[McCullough2010] McCullough, Karl-Erik. 2010. Gestural holds, superimposed beats, and the issue of cross-modal alignment. In *Proceedings of the Fourth Conference of the International Society for Gesture Studies (ISGS)*, Frankfurt an der Oder.

[McNeill1979] McNeill, David. 1979. *The Conceptual Basis of Language*. Hillsdale: Erlbaum.

[McNeill1992] McNeill, David. 1992. *Hand and Mind. What Gestures Reveal about Thought*. University of Chicago Press, Chicago.

[McNeill2005] McNeill, David. 2005. *Gesture and Thought*. University of Chicago Press, Chicago.

[Morrel-Samuels and Krauss1992] Morrel-Samuels, Palmer and Robert M. Krauss. 1992. Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Language, Memory and Cognition*, 18:615–622.

[Müller2004] Müller, Cornelia. 2004. The palm-up-open-hand. a case of a gesture family? In *The semantics and pragmatics of everyday gestures. Proceedings of the The Berlin conference*, pages 233–256, Berlin: Weidler.

[Nunberg1995] Nunberg, Geoffrey. 1995. Transfers of meaning. *Journal of Semantics*, 12:109–132.

[Oepen2001] Oepen, Stephan. 2001. [incr tsdb()] — competence and performance laboratory. User manual. Technical report, Computational Linguistics, Saarland University, Saarbrücken, Germany.

[Oepen, Netter, and Klein1997] Oepen, Stephan, Klaus Netter, and Judith Klein. 1997. TSNLP — Test Suites for Natural Language Processing. In John Nerbonne, editor, *Linguistic Databases*. CSLI Publications, Stanford, CA, pages 13 – 36.

[Oviatt1999] Oviatt, Sharon. 1999. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81.

[Oviatt, DeAngeli, and Kuhn1997] Oviatt, Sharon L., Antonella DeAngeli, and Karen Kuhn. 1997. Integration and synchronization of input modes during multimodal human-computer interaction. *CHI*, pages 415–422.

[Paggio and Navarretta2009] Paggio, Patrizia and Costanza Navarretta. 2009. Integration and representation issues in the annotation of multimodal data. In *Proceedings of the NODALIDA 2009 workshop Multimodal Communication — from Human Behaviour to Computational Models*, volume 6, pages 25–31. Northern European Association for Language Technology (NEALT).

[Pelachaud et al.1995] Pelachaud, Catherine, Justine Cassell, Norman Badler, Mark Steedman, Scott Prevost, and Matthew Stone. 1995. Synthesizing cooperative conversation. In *Proceedings of the International Conference on Cooperative Multimodal Communication*, pages 237–256.

[Pinkal1996] Pinkal, Manfred. 1996. Radical underspecification. In *Proceedings of the 10th Amsterdam Colloquium*, pages 587–606, Amsterdam: University of Amsterdam.

[Polanyi1985] Polanyi, Livia. 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.

[Pollard and Sag1994] Pollard, Carl and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press and Stanford: CSLI Publications.

[Pustejovsky1995] Pustejovsky, James. 1995. *The Generative Lexicon*. MIT Press, Cambridge.

[Quintilianus1992] Quintilianus, Marcus Fabius. 1992. *Institutio Oratoria*, volume XI/3. Translator: E.H. Butler.

[Reyle1993] Reyle, Uwe. 1993. Dealing with ambiguities by underspecification: Construction, representation and deduction. *Journal of Semantics*, 10:123–179.

[Saussure1916] Saussure, Ferdinand de. 1916. *Cours de linguistique générale*. Paris ; Lausanne : Payot.

[Seyfeddinipur2006] Seyfeddinipur, Mandana. 2006. *Disfluency: Interrupting Speech and Gesture*. MPI Series in Psycholinguistics, Radboud Universiteit Nijmegen. Ponsen & Looijen. PhD thesis.

[Seyfeddinipur and Kita2001] Seyfeddinipur, Mandana and Sotaro Kita. 2001. Gesture as an indicator of early error detection in self-monitoring of speech. In *Proceedings of the Disfluency in Spontaneous Speech ISCA Tutorial and Workshop*, University of Edinburgh.

[Shieber1986] Shieber, Stuart M. 1986. *An Introduction to Unification-Based Approaches to Grammar*, volume 4 of *CSLI Lecture Notes Series*. Center for the Study of Language and Information, Stanford, CA.

[Sidner and Lee2007] Sidner, Candace and Christopher Lee. 2007. Attentional Gestures in Dialogues between People and Robots. In T. Nishida, editor, *Engineering Approaches to Conversational Informatics*. Wiley and Sons.

[Slama-Cazacu1976] Slama-Cazacu, Tatiana. 1976. Nonverbal components in message sequence: 'mixed syntax'. *Language and Man: Anthropological Issues*, pages 217–222.

[Slobin1996] Slobin, Dan I. 1996. From thought and language to thinking for speaking. In *Rethinking linguistic relativity*. Cambridge: Cambridge University Press, pages 70–96.

[Steedman2000] Steedman, Mark. 2000. *The Syntactic Process*. The MIT Press.

[Webber1991] Webber, Bonnie Lynn. 1991. Structure and ostension in the interpretation of discourse deixis. In *Language and Cognitive Processes*, pages 107–135.

[Wittenburg, Weitzman, and Talley1991] Wittenburg, Kent, Louis Weitzman, and Jim Talley. 1991. Unification-based grammars and tabular parsing for graphical languages. *J. Vis. Lang. Comput.*, 2(4):347–370, December.

[Zwicky1982] Zwicky, Arnold. 1982. Stranded *to* and phonological phrasing in english. *Linguistics*, 20(1/2):3–57.