Katya Alahverdzhieva and Alex Lascarides

School of Informatics
University of Edinburgh, United Kingdom

Proceedings of the HPSG11 Conference

University of Washington, Seattle, USA

Stefan Müller (Editor)

2011

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

The use of hand gestures to point at objects and individuals, or to navigate through landmarks on a virtually created map is ubiquitous in face-to-face conversation. We take this observation as a starting point, and we demonstrate that deictic gestures can be analysed on a par with speech by using standard methods from constraint-based grammars such as HPSG. In particular, we use the form of the deictic signal, the form of the speech signal (including its prosodic marking) and their relative temporal performance to derive an integrated multimodal tree that maps to an integrated multimodal meaning. The integration process is constrained via construction rules that rule out ill-formed input. These rules are driven from an empirical corporal study which sheds light on the interaction between speech and deictic gesture.

1 Introduction

The use of deixis is highly pervasive in everyday communication. Through definite referring expressions, pronouns and pointing gestures with the head and hand, people exploit the context of the communicative event in their communicative actions, and likewise interlocutors exploit this to derive an interpretation of those actions. This paper provides a formal account of deictic (pointing) gestures performed by the hand (from now on called *deixis*) and it demonstrates that standard methods from formal linguistics—namely constraint-based grammars and compositional semantics—can capture the various semantic relations between speech and deixis, and also the range of pragmatic use of deixis. To illustrate the distinct semantic relations and the distinct pragmatic uses, consider utterances (1) and (2).[1]

(1) And a as she [$_{N}$said], it's an environmentally friendly uh material . . .
The speaker extends Right Hand (RH) with palm open up towards the other participant.

(2) I [$_{PN}$enter] my [$_{N}$apartment]
RH and Left Hand (LH) are in centre, palms are open vertically, finger tips point forward; along with "enter" they move briskly downwards.

The different ways the pointing hand is engaged in the communicative event to denote the speech content gives rise to distinct interpretations of deixis: the gesture in (1) can be interpreted as demarcating the spatial location of a concrete participant salient in the communicative situation, or also as pointing at an abstract object—here, the utterance introduced by the previous speaker, located at some

[1] In the utterance transcription, the speech signal that occurs at the same time as the expressive part of the gesture, the so called *stroke*, is underlined with a straight line, and the signal that temporally co-occurs with the *hold* after the stroke is underlined with a curved line. 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).

specific spatiotemporal coordinates. In comparison, the deixis in (2) can locate an object that is physically absent from the communicative situation—an apartment or an apartment entrance door—by placing it on a virtually created map. This gesture can also identify the abstract event of entering the apartment door. In the gesture community, the use of deixis to point at physically present individuals vs. individuals absent from the communicative event is what sets apart *concrete deixis* from *abstract deixis* (McNeill, 2005). This distinction is essential since it has effects on the speech-deixis integration, as we discuss in Section 3.1 and Section 5.

With this in mind, the Logical Forms (LFs) contributed by (1) and (2) reflect the distinct gesture denotations, as well as the distinct relations between speech and deixis. We begin with the formalisation of multimodal utterance (1), with its two possible interpretations exhibited in (3a) and (3b).

(3) a. $\pi_1 : \exists m(material(m) \wedge environmentally\text{-}friendly(m))$
$\pi_2 : \exists s, g(she(s) \wedge said(e_0, s, \pi_1) \wedge loc(g, x, v(\vec{p}_x)) \wedge Identity(s, x))$

b. $\pi'_1 : \exists m(material(m) \wedge environmentally\text{-}friendly(m))$
$\pi'_2 : \exists s, g(she(s) \wedge said(e_0, s, \pi'_1) \wedge classify(g, \pi'_1, v(\vec{p_s}))$
$\wedge Acceptance(\pi'_1, g))$

To fit the current research in the broader context of formal semantics of gesture (Lascarides and Stone, 2009), (3a) and (3b) 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. Following Lascarides and Stone (2009), we use the predicates *loc* and *classify* to represent the literal and metaphorical deixis use; for instance, *loc*$(g, x, v(\vec{p}_x))$ states that the deictic gesture g introduces an individual x at the physical location $v(\vec{p}_x)$ which is the proximal space projected from the tips of the fingers in the direction of the participant.[2] In comparison, *classify*$(g, \pi'_1, v(\vec{p_s}))$ conveys the metaphorical deictic use to point at an abstract object, namely, the utterance denoted by π'_1 "contained" in the spatial coordinates $v(\vec{p_s}))$. Finally, distinct semantic relations can be inferred between the speech content and these two alternative gesture contents: we state that an *Identity* relation holds between the referents s and x in (3a). Thus the gesture physically locates the referent of "she" in physical space. In the metaphorical case, the semantic relation between speech and deixis is *Acceptance*(π'_1, g); in other words, the gesture's interpretation can be paraphrased as "I agree with what was just said" (note that π'_1 refers to the discourse segment whose content is "it's an environmentally friendly material").

We complete the range of deixis interpretations with the formalisation of (2) as displayed in (4a) and (4b).

(4) a. $\pi_1 : \exists a, g(speaker(s) \wedge apartment(a) \wedge enter(e_0, s, a)$
$\wedge loc(g, y, v(\vec{p}_y) \wedge VirtualCounterpart(a, y))$

[2] We postpone a more detailed discussion about $v(\vec{p}_i)$ until Section 4.

b. $\pi'_1 : \exists a, g(speaker(s) \wedge apartment(a) \wedge enter(e_0, s, a)$
$\wedge loc(g, e_1, v(\vec{p}_{e_1})) \wedge VirtualCounterpart(e_0, e_1))$

Whereas the LF in (4a) exemplifies one of the possible interpretations where the deictic gesture locates the apartment in a virtual map that is just in front of the speaker (through the use of *VirtualCounterpart(a,y)*), the LF in (4b) locates the event of entering an apartment in the virtual space — hence given real world knowledge about entering events it locates the apartment door. Based on that, we establish a VirtualCounterpart relation between the abstract object y and the apartment a in (4a), and between the event of entering the apartment e_0 and the deictic event e_1 in (4b).

We construct these logical forms from the underspecified semantics of deixis, the semantics of speech and the underspecified semantic relation between speech and deixis using commonsense reasoning and world knowledge. Essentially, we argue that computing how speech and deixis are integrated should happen within the *grammar* so as to capture the fact that the integration is informed by *form*. For instance, it seems anomalous to perform the deictic gesture in (2) along with the prosodically unmarked “I”, as displayed in (5), despite the multiple interpretations that can arise from this deixis use. We view utterance (5) as ill-formed where the source of ill-formedness involves the form (here, the prosodic markedness) of the linguistic signal. Ultimately in this case, we are going to capture this ill-formedness within the grammar. The alternative approach of relying only on the semantics/pragmatics interface to compute the integration of speech and deixis would involve accessing information about form disrupting thus the transition between syntax, semantics and pragmatics.

(5) * I [$_{PN}$enter] my [$_{N}$apartment]
Same gesture as in (2).

We therefore intend to provide a precise methodology for integrating speech and deixis in a single syntactic tree that maps to an (underspecified) meaning, and which also features an (underspecified) speech-deixis relation. We do this via an HPSG-based grammar of speech and deixis which defines empirically extracted construction rules for “attaching” gesture to the synchronous, semantically related speech phrase and which also introduces an underspecified $deictic\_rel(s, d)$ relation between the speech s content and the deixis d content. Resolving this relation to, say, Identity or VirtualCounterpart, is achieved at the semantics/pragmatics interface and it therefore lies outwith the scope of the grammar.

As a grammar formalism we choose HPSG because of its mechanisms to construct structured phonology in parallel with syntax (Klein, 2000), and also because the semantic composition is expressed in (Robust) Minimal Recursion Semantics ((R)MRS, Copestake et al. (2005)). (R)MRS overcomes the shortcomings of λ-calculus in that the composition is *constrained*, i.e., it does not allow a functor to pick arguments that are arbitrarily embedded in the underspecified logical form. A further advantage is that (R)MRS produces Underspecified Logical Formulae

(ULF): whereas with operations such as functional application or β-reduction, one imposes scope constraints and embeddings driven from the syntactic tree, (R)MRS produces a flat description of the possible readings without having to access the distinct readings themselves. This property is particularly useful for composing gesture meaning since even through discourse processing the semantic predications yielded by gestural form may remain unresolved as attested by the LFs in (3a), (3b) and also (4a), (4b).

We have demonstrated elsewhere that HPSG is suitable for deriving depicting gestures in parallel with speech (Alahverdzhieva and Lascarides, 2010). In this paper, we shall demonstrate that it is suitable for analysing deictic gestures as well.

2 Deixis Ambiguities

One of the major challenges for the constraint-based analysis of deixis concerns the ambiguity in form which is represented on the following two axes:

1. Gesture form features, which include the shape of the hand, its orientation, movement and location. This level of ambiguity has as an effect that the hand often underspecifies the region it points at: does an index finger (1-index) extended in the direction of a book identify the physical object book, the location of the book, e.g., the table, or the cover of the book? Despite that the region identified by the 'pointing cone' (Kranstedt et al., 2006) remains vague, it does not violate perception as speakers rely on the synchronous speech phrase to disambiguate the pointing, e.g., "the book", "the book cover", etc.

2. Attachment ambiguity, which involves the syntactic integration of the deixis daughter to the synchronous, semantically related, speech daughter. For instance, in (3a) s and x are semantically related, while in (3b) π_1' and g are related. This difference is sourced in the distinct attachments in syntax: whereas an attachment to "she" supplies an interpretation where the gesture's denotation is *identical* to the denotation of the pronoun in speech, an interpretation where the gesture signals an *acceptance* of an utterance is supported by a higher attachment in the syntactic tree. This observation is essential since the grammar needs to provide the methodology for enabling the range of possible attachment ambiguities.

Deixis displays further ambiguity with respect to the way it relates to the synchronous speech, which stems from the fact that the gesture can denote distinct features of the 'qualia structure' (Pustejovsky, 1995) of the referent. An example from Clark (1996) illustrates this: George points at a copy of Wallace Stegner's novel *Angle of Repose* and says: 1. "<u>That book</u> is mine"; 2. "<u>That man</u> was a friend of mine"; 3. "I find <u>that period of American history</u> fascinating". In 1., there is one-to-one correspondence between the deixis denotation and the physical artefact book, and they are thus bound by *Identity*. In 2., there is a reference transfer from the book to the author and the gesture denotes the creative agent of the book rather than the book itself, i.e., the gesture and speech are related through an *AgentiveRelation*, and finally in 3., the transfer is from the book to the book's content, and so deixis and speech are related through a *ContentRelation*. We shall account for these ambiguities in the grammar by a construction rule that combines synchronous speech and gesture via an underspecified relation $deictic\_rel(d, s)$ between the semantic index d of deixis and the semantic index s of speech, resolvable to a concrete value in pragmatics.

We argue that these various levels of ambiguity can be captured by standard mechanisms for producing ULFs which give a very abstract representation of what the gesture means abstracted away from context. In particular, we use Robust Minimal Recursion Semantics (Copestake, 2007) to produce highly factorised, partial meaning representations that underspecify the predicate's arity and the predicate's main variable. In so doing, we remain vague as to whether the pointing signal in (1) identifies the individual denoted by a pronoun in the synchronous speech, or it is rather a metaphor of the speech act of acceptance.

Despite the ambiguities, the process of attachment is constrained, e.g., whereas attachments to "enter", "enter my apartment" or even to the entire clause "I enter my apartment" in (2) should be enabled as they support the intended meanings in context, an attachment to the subject head daughter "I" should be ruled out since it would never produce the intended meaning in context.

3 Speech-Deixis Synchrony

Due to the lack of an accepted methodology of how to establish the synchrony of two modalities,[3] Alahverdzhieva and Lascarides (2010) defined synchrony as the *attachment of gesture to the semantically related speech phrase in the syntactic tree that, using standard semantic composition rules, yields an underspecified logical form supporting the final interpretation in the context-of-use*. Our aim is thus to constrain synchrony by exploring the linguistic properties of the multimodal action, i.e., we use information from prosody (the literature offers enough evidence that the gesture performance is intertwined with the one of speech, and that the perception of gesture depends on the synchronous prosody–e.g., Loehr (2004), Giorgolo and Verstraten (2008)), syntax (why would attachment to "enter my apartment" in (2) be allowed, but one to "I" disallowed?) and also the timing of speech relative to deixis. These constraints have been established empirically though a multimodal corpora study.

[3] As demonstrated by (1) and (2) and their corresponding logical forms, the temporal performance of one mode relative to the temporal performance of the other is insufficient for deriving the possible meaning representations.

3.1 Corpus Investigation

Autosegmental-Metrical (AM) phonology (Ladd, 1996) underpins our underlying assumptions about the interaction between speech and gesture, and hence also the annotation schema and the formalisation of grammar construction rules. In AM theory, prominence is determined by the stronger (s) or weaker (w) relation between two juxtaposed units in the metrical tree. The nuclear prominent node is the one dominated by strong nodes. In the default case of broad focus, the nuclear accent is associated with the right-most word, i.e., the metrical structure is right branching as displayed in Figure 1. This can be overridden by narrow focus where the structure can also be left-branching.

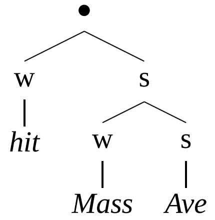

Figure 1: Metrical Tree for "hit Mass Ave"

Our choice stems from the fact that in the AM model nuclear accenting involves perception of structural prominence in relation to the metrical structure rather than to the acoustic properties of the syllable (Calhoun, 2006). In this way, we can reliably predict the gestural occurrence in relation to the metrical tree, and we can also interface the prosodic structure with the syntactic structure (Klein, 2000).

Our hypothesis about the speech-deixis interaction is as follows:

Hypothesis 1 *The relative temporal performance of deictic gesture and speech can be predicted from nuclear prominence: in case of broad-focused utterances, deixis temporally overlaps with the nuclear accent, and in case of early pre-nuclear rise, it overlaps with the pre-nuclear accent.*

The hypothesis was validated through an experimental study over two multimodal corpora: a 5.53 min recording from the Talkbank data[4] and observation IS1008c, speaker C from the AMI corpus.[5] The domain of the former is living-space descriptions and navigation giving, and the latter is a multi-party face-to-face conversation among four people discussing the design of a remote control. We augmented the corpora with annotation of prosody and of gesture. The prosody annotation was largely based on the annotation schema of the Switchboard corpus (Brenier and Calhoun, 2006) and it included an orthographic transcription, labelling of accents—nuclear, pre-nuclear (an early emphatic pitch rise), non-nuclear—and labelling of prosodic phrases. The gesture annotation included classifying the hand

[4] http://www.talkbank.org/media/Gesture/Cassell/kimiko.mov

[5] http://corpus.amiproject.org/

movements in terms of communicative vs. non-communicative, assigning them a category (depicting, deictic) and segmenting them into discrete phases. These phases are: preparation (a non-obligatory phase which involves lifting the hands from a relaxed position to the frontal space), pre-stroke hold (a non-obligatory phase, hands are held still before reaching the expressive peak), stroke (an obligatory phase, the dynamic peak of gesture that carries its meaning), a post-stroke hold (a non-obligatory phase which consists in maintaining the hands in the expressive position reached during the stroke) and retraction (a non-obligatory phase characterised by bringing the hands back to rest).

The gesture segmentation was based on formal and functional criteria. The formal ones considered the dynamic profile of the hand, i.e., the effort employed by the hand. Any sudden change in the hand dynamics signals a transition to a new phase. More specifically, preparations and retractions require minimum effort, the stroke is usually characterised by a dynamic maximum, and during the holds before/after the strokes the hand is held still (McNeill, 2005). Note that this criterion is relational — the lower or higher dynamics of a phase is determined in relation to the dynamics of the juxtaposed phase, e.g., the hand during hold is almost never absolutely still, it is still only in relation to the dynamics reached during the stroke. Further, the functional criteria involve the meaning conveyed by the gesture phase, which we established in the context of the synchronous speech: whereas the stroke and the hold after the stroke (if any) are the phases that communicate what the gesture is about, preparations and retractions are not communicative, they are the physical effort necessary to execute the stroke.

We addressed our hypothesis by searching for types of accents overlapping deixis. Since we were interested in the expressive part of the gesture, we counted the deictic strokes only. The corpora contained 87 deictic strokes (65 for the Talkbank, and 22 for AMI). 86 of them—that is, 98.85%—overlapped a nuclear and/or a pre-nuclear accented word. Deictic gestures of longer duration were often marked by a combination of a nuclear and non-nuclear and/or nuclear and pre-nuclear accented words. Essentially, the empirical analysis confirmed the expected alignment between the nuclear prominent word (not simply the nuclear accent) and the deixis stroke both in case of broad focus, and in case of narrow focus. This is attested in the broad-focused utterance (6) and in the narrow-focused utterance (7), a continuation of (6). Whereas the deixis stroke in (6) co-occurs temporally with the nuclear prominent "Mass Ave", the performance of the deixis stoke in (7) is shifted earlier to the nuclear accented "left".

(6) 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, RH 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.

(7) And then I [$_{N}$turn] *[pause]* [$_{N}$left] on [$_{NN}$Mass] Ave

LH is held in the same position as in (6); along with "left", RH opens vertically and sweeps to the left periphery close to the left shoulder.

For the formal rendition of this finding, we adopt the HPSG phonology model of Klein (2000) where the prosodic structure is specified within the PHON attribute in parallel with SYNSEM. The prosodic constituent is mapped from the metrical tree, e.g., the metrical tree in Figure 1 maps to the feature structure in Figure 2. The element dominated by *s* nodes maps to the *Designated Terminal Element* (DTE) (Liberman and Prince, 1977). Note also that the feature structure is typed as *mtr(full)* which reflects the fact that objects in the domain (DOM) are prosodic words of type *full*, which is in contrast to non-prosodic words such as conjunctions, pronouns and articles that usually form a single prosodic word with the neighbouring element.

$$\begin{bmatrix} \textit{sign} & \\ \textsc{phon} & \begin{bmatrix} \textit{mtr(full)} & \\ \textsc{dom} & \left\langle \textit{hit}, \boxed{1} \begin{bmatrix} \textit{mtr(full)} \\ \textsc{dom} \left\langle \textit{Mass}, \boxed{2}\, \textit{Ave} \right\rangle \\ \textsc{dte}\ \boxed{2} \end{bmatrix} \right\rangle \\ \textsc{dte} & \boxed{1} \end{bmatrix} \\ \textsc{synsem} & \mathsf{synsem} \end{bmatrix}$$

Figure 2: Feature Structure of the Metrical Tree for "hit Mass Ave"

Our results report on the interaction between speech and deixis on the level of *form*. Our overall aim is to account for syntactically well-formed trees which map to ULFs supporting the final interpretations in context. We therefore examined whether the syntactic attachments as constrained by prosody would produce the range of preferred interpretations in context. We encountered instances which, although syntactically well-formed, did not map to all intended meaning representations due to the fact that the semantically preferred speech element the gesture stroke overlapped with was not prosodically prominent. In (1), for instance, the gesture is produced along with the nuclear prominent "said" when one of the plausible denotations of the hand is that it is identical to the denotation of the unaccented pronoun "she" coming from speech. Moreover, this interpretation would still be available even if the deictic gesture was performed outwith the temporal span of the pronoun, as exemplified below.

(8) And a as she [$_N$said], it's an environmentally friendly uh material ...
Same gesture as in (1).

Essentially, the instances of misalignment between the semantically related, prosodically prominent word and the deictic stroke, and also between the temporal performance of the deixis and the temporal performance of the semantically related speech phrase concern cases where the visible space outlined by the deictic gesture is equal to the space it actually denoted, i.e., the individual/object was present in the communicative situation at the exact spatial coordinates identified by the deixis. This observation flags up an important finding about a multimodal grammar of speech and deixis: whereas gestures pointing at concrete individuals in the real space can be attached to elements from speech that are not necessarily prosodically prominent or that are performed outside the temporal performance of the deixis, gestures identifying abstract individuals require temporal overlap with the prosodically prominent, semantically related speech phrase. In Section 5, we propose construction rules that reflect our empirical findings.

4 Mapping Form to (Underspecified) Meaning

In Section 1 we claimed that we model gestural ambiguity by re-using standard linguistic methods for meaning underspecification. We shall now demonstrate how to express gestural meaning from form.

It is now well-established in the gesture community to formally regiment gesture in terms of Typed Feature Structures (TFSs)—e.g., Johnston (1998), Kopp et al. (2004)—since they capture the non-hierarchical gesture structure. Gestures, unlike fully-fledged language systems, are constructed by equally ranked form features—such as the shape of the hand, the palm and finger orientation—which do not compose a hierarchy (McNeill, 2005). Similarly, previous HPSG approaches to sign languages, British Sign Language in particular, incorporate the information coming from the hand shape, orientation, finger direction and movement within the PHON attribute (Marshall and Sáfár, 2004). However, in contrast to sign languages, which exhibit a combinatoric potential to combine with other arguments (Cormier et al., 1999), (Marshall and Sáfár, 2004), deictic gestures do not select obligatory arguments. Still, multiple gestures can form a hierarchical structure in the same way discourse segments do.

Recording the deixis form features is essential for identifying the region designated by the pointing hand, for instance, 1-index finger projects a line or even a cone that starts from the tip of the index finger and continues in the direction of the object pointed at. In comparison, a flat open hand can project a plane that starts from the palm and extends in a direction parallel to the palm. Furthermore, there are findings in the descriptive literature that suggest that the form of the pointing hand is significant for interpreting its meaning in context, e.g., whereas 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).

In our framework, the features appropriate for gesture include the shape of the hand, its movement, location and orientation of the palm and fingers. Their values are specified within the sort hierarchy as exemplified for *hand-shape* in Figure 3. Some values, such as *open-closed,* account for change in form.

Figure 4 regiments the form of the deixis in utterance (2) as a feature structure. It is typed as $deictic\_abstract$ so as to differentiate between feature struc-

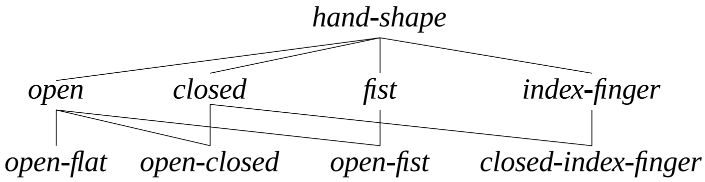

Figure 3: Fragment of the Sort Hierarchy of *hand-shape*

$$\begin{bmatrix} \textit{deictic\_abstract} & \\ \textsc{Hand-shape:} & \text{open-flat} \\ \textsc{Palm-orientation:} & \text{vertical} \\ \textsc{Finger-orientation:} & \text{forward} \\ \textsc{Hand-movement:} & \text{away-body-centre} \\ \textsc{Hand-location:} & \vec{c} \end{bmatrix}$$

Figure 4: Deixis Form Feature Structure Representation

tures contributed by abstract deixis and those contributed by concrete deixis (of type $deictic\_concrete$). This information is essential as it allows us to encode the necessary constraints between speech and concrete deictic gesture on the one hand, and between speech and abstract deictic gesture, on the other (recall our finding from Section 3.1 that relaxation between the prosodically prominent speech phrase and deixis, and also between the timing of the deixis and the timing of the speech word occurs with deictic gestures identifying concrete individuals but not abstract ones). Further, the values of the distinct features are taken from the sort hierarchies, similar to those demonstrated in Figure 3. Finally, following Lascarides and Stone (2009), we formalise the hand location in terms of the constant $\vec{c}$ which demarcates the exact location of the tip of the index finger and which, combined with the deixis form features, determines the spatial region $\vec{p}$ designated by the gesture, for instance, a stationary gesture of 1-index would make $\vec{p}$ a line (or a cone) that projects from $\vec{c}$ in the same direction as the index finger.

The compositional semantics of deictic gesture involves producing a set of underspecified predications in the RMRS notation; for instance, the RMRS representation of the deictic gesture in (2) is shown in Figure 5.

$$\begin{array}{l} l_1 : a_1 : deictic\_q(i)\ RSTR(a_1, h_1)\ BODY(a_1, h_2) \\ l_2 : a_2 : sp\_ref(i)\ ARG1(a_2, v(\vec{p})) \\ l_2 : a_3 : hand\_shape\_open\_flat(e_0)\ ARG1(a_3, i) \\ l_2 : a_4 : palm\_orient\_vertical(e_1)\ ARG1(a_4, i) \\ l_2 : a_5 : finger\_orient\_forward(e_2)\ ARG1(a_5, i) \\ l_2 : a_6 : hand\_move\_away\_body\_centre(e_3)\ ARG1(a_6, i) \\ h_1 =_q l_2 \end{array}$$

Figure 5: Deixis RMRS Representation

Each predication is associated with a not necessarily unique label (l_n) and a unique anchor (a_n): the label identifies the scopal positions of the predicate in the resolved LF and the anchor serves as a locus for adding arguments to the predicate, e.g., $l_2 : a_2 : sp\_ref(i)\ ARG1(a_2, v(\vec{p}))$ makes the predicate sp_ref take at least the two arguments i and $v(\vec{p})$ in the that order.

The deixis semantics accounts for the fact that the deictic gesture provides spatial reference of an individual or event in the physical space $\vec{p}$. Following Lascarides and Stone (2009), this is formalised in terms of the 2-place predicate $l_2 : a_2 : sp\_ref(i)\ ARG1(a_2, v(\vec{p}))$ where i is an underspecified variable (resolvable to an event e or an individual x) and $v(\vec{p})$ is the actually denoted space. To reflect the fact that the gestured space is not necessarily identical to the denoted space (which is basically the underlying difference between concrete deixis and abstract deixis), we are using the function v to map the physical space $\vec{p}$ identified by the gesture to the space $v(\vec{p})$ it denotes; e.g., in (1) the referent is at the exact coordinates in the visible space the gesture points at, i.e., v is equality, and also the deictic gesture is of type concrete. In contrast, in (2) the referent is not physically present, and so the deixis is abstract, and also v does *not* resolve to equality.

Further to this, for consistency with the English Recourse Grammar (ERG) (Copestake and Flickinger, 2000) where individuals are bound by quantifiers, the deictic referent is bound by the quantifier $deictic\_q$. Finally, to capture the semantic effects of the deixis form features, we map each feature-value pair to a predicate that, similarly to intersective modification in ERG, modifies the referent i.

5 Construction Rules

The rules for integrating deixis and speech envisage coverage of the full set of multimodal constructions found in our empirical study. These include rules that capture our findings about the interaction between nuclear prominence and deixis (rules for the integration of a single prosodic word and deixis, head-argument construction and deixis, head-modifier construction and deixis, noun-noun compounds/appositives and deixis). The rules are also based on the particular gesture type to account for the cases of prosodic and/or temporal relaxation.

In this section, we present three construction rules: a basic rule that attaches deixis to a single prosodic word (to derive a context-specific analysis of (1) as (3a)), a rule that integrates deixis with a larger spoken phrase (to derive an analysis of (1) as (3b)), and also a rule applicable to concrete deictic gestures that defeats the strict temporal condition between the stroke and the prosodically prominent spoken word.

Rule 1 *Deictic gesture can attach to the nuclear/pre-nuclear accented word of the temporally overlapping speech phrase.*

The formalisation of this rule is demonstrated in Figure 6. We shall now describe every aspect of it in turn. A prerequisite for the integration of the deictic (D) and the spoken (S) modalities is that they temporally overlap, that is,

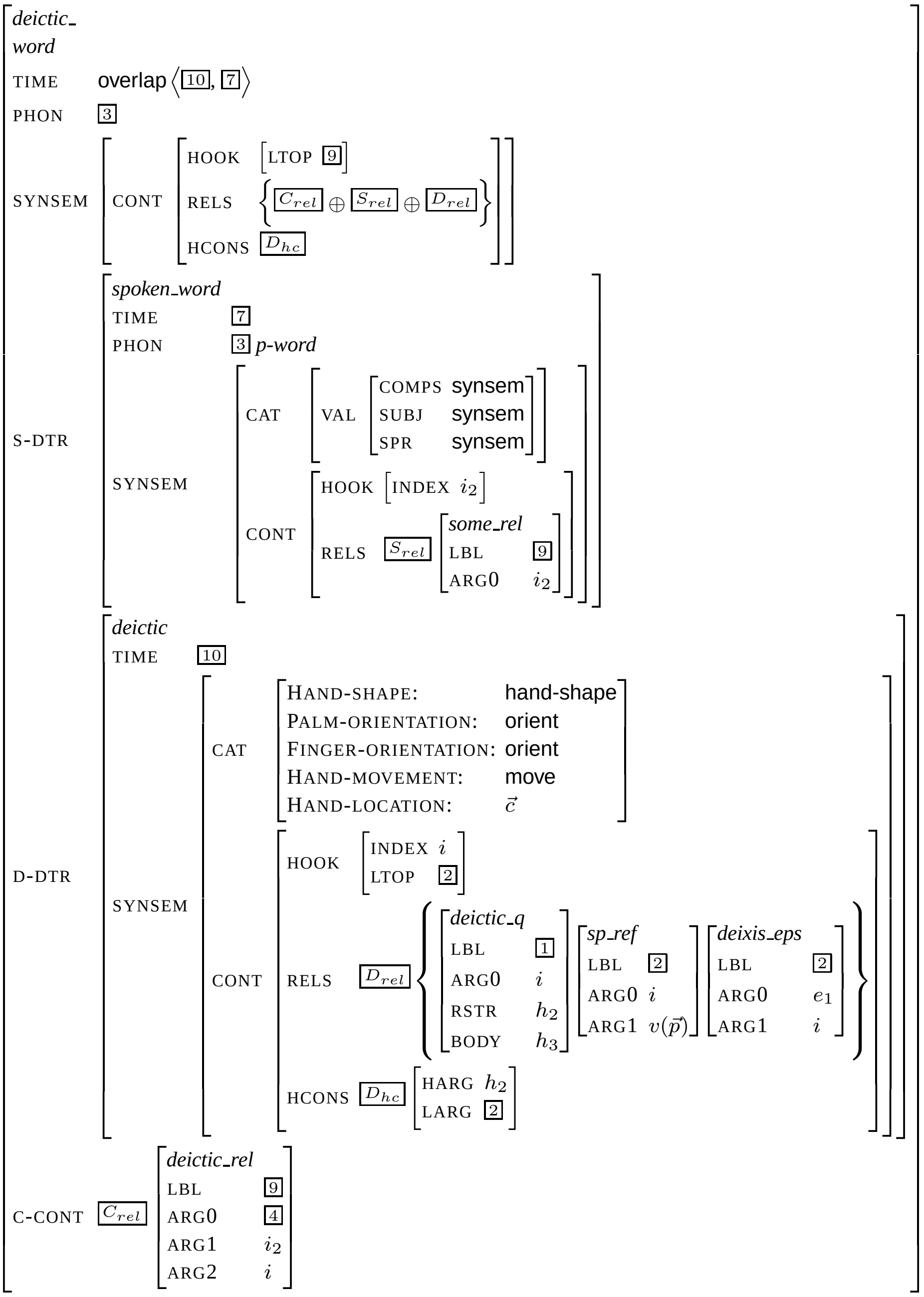

Figure 6: Deictic Prosodic Word Constraint

$end(D) > start(S)$ and $end(S) > start(D)$. Note that the application of this rule is not constrained to a particular deictic gesture type, and so it can apply to both abstract deixis and concrete deixis. The SYNSEM values of the deictic daughter are encoded as detailed in Section 4: the CAT feature contains a list of deixis' appropriate attributes and the CONT component is specified in the standard way in terms of HOOK, RELS and HCONS. We defined the pointing hand as providing a spatial reference of an individual or an event i at some position in the denoted space $v(\vec{p})$ that is determined by the physical space $\vec{p}$ and the contextually resolved mapping v from physical space to gestured space. For the sake of space, we gloss over the gesture form features as $deixis\_eps$. Following ERG where the LTOP of an intersective modifier phrase is shared with the LBLs of the head daughter and the non-head daughter, $deixis\_eps$ share the same label with sp_ref which is the LTOP of the gesture daughter. Finally, the semantic index of the gesture daughter is obtained via co-indexation with the ARG0 variable i bound by the deixis main relation sp_ref.

For the speech daughter, we similarly record its timing, syntax and semantic information, and also its prosody. Importantly, the speech head daughter should be a prosodically prominent word. We forego any details about the syntactic category of the speech daughter since it does not constrain the integration.

In Section 1 we stated that the full inventory of relations combining speech and deixis will be accounted for by an underspecified relation supporting the possible relations in context. Based on Lascarides and Stone (2009), the construction rule therefore introduces in C-CONT an underspecified relation $deictic\_rel$ between the semantic index i of the deictic gesture and the semantic index i_2 of the speech. How this relation resolves is a matter of discourse context. 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 $\oplus$). 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.

Applied to (9), this rule would produce a tree where the deixis is attached to the prosodic word “hallway”.

(9) There’s like a [$_{NN}$little] [$_{N}$hallway]
Hands are open, vertical, parallel to each other. The speaker places her hands between her centre and the left periphery.

For the sake of space, in Figure 7 we provide only the semantics of the multimodal utterance. Note that synchrony resolves the underspecified index introduced by the deictic gesture to an individual x. 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.

In Section 2 we stated that there was ambiguity with respect to attaching deixis

$$\begin{bmatrix} \textit{deictic\_word} & \\ \text{CONT} & \begin{bmatrix} \text{HOOK} & \begin{bmatrix}\text{LTOP} & \boxed{4}\end{bmatrix} \\ \text{RELS} & \left\{ \begin{array}{l} \begin{bmatrix} \textit{deictic\_rel} & \\ \text{LBL} & \boxed{4} \\ \text{ARG0} & e_0 \\ \text{ARG1} & x_2 \\ \text{ARG2} & x \end{bmatrix} \begin{bmatrix} \textit{hallway} & \\ \text{LBL} & \boxed{4} \\ \text{ARG0} & x_2 \end{bmatrix} \\ \begin{bmatrix} \textit{deictic\_q} & \\ \text{LBL} & \boxed{1} \\ \text{ARG0} & x \\ \text{RSTR} & h_2 \\ \text{BODY} & h_3 \end{bmatrix} \begin{bmatrix} \textit{sp\_ref} & \\ \text{LBL} & \boxed{2} \\ \text{ARG0} & x \\ \text{ARG1} & v(\vec{p}) \end{bmatrix} \begin{bmatrix} \textit{deixis\_eps} & \\ \text{LBL} & \boxed{2} \\ \text{ARG0} & e_1 \\ \text{ARG1} & x \end{bmatrix} \end{array} \right\} \\ \text{HCONS} & \begin{bmatrix} \text{HARG} & h_2 \\ \text{LARG} & \boxed{2} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Figure 7: Semantic Composition for Deixis + “hallway”

to the synchronous and semantically related speech phrase. We therefore introduce a further rule that takes that into account.

Rule 2 *Deictic gesture attaches to a nuclear/pre-nuclear prominent head saturated with its arguments if there is an overlap between the timing of the deixis and the timing of head.*

Unlike the non-empty VAL list of the rule in Figure 6, Rule 2 presupposes attachment to a phrase with an empty [VAL|COMPS $\langle\rangle$] and/or [VAL|SUBJ $\langle\rangle$] and/or [VAL|SPR $\langle\rangle$] list. We remain as neutral as possible about the number of saturated arguments to accommodate the fact that the deixis form can map to multiple meanings in context, and these meanings persist even in the contextually resolved discourse. Applied to multimodal utterance (2), Rule 2 would allow for combining “enter my apartment” + deixis, “I enter my apartment” + deixis, and even “I enter” + deixis. Whereas the first two derivations include standard synctactic constituents, the latter violates the HPSG principles of syntactic constituency. With this in mind, one can account for the relation between “I enter” and the deictic gesture on the semantic level by restricting the scope of $deictic\_rel$ over the elementary predicates introduced by “I” and by “enter”.

Finally, we introduce a rule that is applicable to concrete deictic gestures to account for the fact that prosodic prominence of the semantically related spoken word overlapping the concrete deixis is not necessary, and also that the spoken word can happen outwith the temporal performance of the gesture stroke as follows:

Rule 3 *Concrete deictic gesture attaches to a prosodically marked or to a prosodically unmarked spoken word whose temporal performance precedes or follows the temporal performance of the concrete deixis.*

The formal rendition of this rule is demonstrated in Figure 8. This rule remains loose about the temporal relation between the spoken word and the gesture stroke — we allow for precedence and for sequence relations (the overlap relation is also possible, and it was accounted for the rule in Figure 6). Further, the spoken word is not restricted to a particular prosodic type and in this way we can integrate a concrete deictic gesture into a non-prominent spoken word; in utterance (1), for instance, this condition enables the deixis attachment to "she". Moreover, the gesture is restricted to type *concrete_deixis*, and so this bars an attachment of the abstract deictic gesture to "I" in utterance (2). We forego any further details about the formalisation of this rule, since it remains the same as in Rule 1.

$$\begin{bmatrix} \textit{deictic\_word} & \\ \text{TIME} & \text{precede}\left\langle \boxed{10}, \boxed{7} \right\rangle \vee \text{follow}\left\langle \boxed{10}, \boxed{7} \right\rangle \\ \text{PHON} & \boxed{3} \\ \text{SYNSEM} & \text{synsem} \\ \text{S-DTR} & \begin{bmatrix} \textit{spoken\_word} & \\ \text{TIME} & \boxed{7} \\ \text{PHON} & \boxed{3}\ \textit{pros} \\ \text{SYNSEM} & \text{synsem} \end{bmatrix} \\ \text{D-DTR} & \begin{bmatrix} \textit{deictic\_concrete} & \\ \text{TIME} & \boxed{10} \\ \text{SYNSEM} & \text{synsem} \end{bmatrix} \end{bmatrix}$$

Figure 8: Concrete Deixis Prosodic Word Constraint

6 Conclusions

In this paper, we presented a constraint-based analysis of multimodal communicative signals consisting of deictic gesture signals and speech signals. Our approach re-uses standard devices from linguistics to map multimodal form to an underspecified meaning that will ultimately support reasoning on the semantic/pragmatic interface for producing a specific and context aware interpretation. We thereby account for gestural ambiguity by means of established underspecification mechanisms. To specify the form-meaning mapping, we used empirically extracted grammar construction rules which capture the conditions under which the speech-deixis signal is grammatical and semantically intended. We presented three rules: a basic rule accounting for a multimodal speech-deixis word, a rule allowing for attaching deixis to a spoken phrase, and finally, a rule that defeats the strict temporal/prosodic condition between the spoken word and the deixis stroke.

7 Acknowledgements

This work was partly funded by EU project JAMES (Joint Action for Multimodal Embodied Social Systems), project number 270435. The research of one of the authors was funded by EPSRC. The authors would like to thank the anonymous reviewers for the useful comments that have been addressed in the final version. The authors are also grateful to Nicholas Asher, Sasha Calhoun, Jean Carletta, Jonathan Kilgour, Ewan Klein and Mark Steedman. Any mistakes and inaccuracies are our own.

References

Alahverdzhieva, Katya and Lascarides, Alex. 2010. Analysing Language and Co-verbal Gesture in Constraint-based Grammars. In Stefan Müller (ed.), *Proceedings of the 17th International Conference on Head-Driven Phase Structure Grammar (HPSG)*, pages 5–25, Paris.

Asher, Nicholas and Lascarides, Alex. 2003. *Logics of Conversation*. Cambridge University Press.

Brenier, Jason and Calhoun, Sasha. 2006. Switchboard Prosody Annotation Scheme. Department of Linguistics, Stanford University and ICCS, University of Edinburgh, internal publication.

Calhoun, Sasha. 2006. *Information Structure and the Prosodic Structure of English: a Probabilistic Relationship*. University of Edinburgh, phD Thesis.

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

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

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

Cormier, Kearsy, Wechsler, Stephen and Meier, Richard P. 1999. Locus Agreement in American Sign Language. In A. Kathol, J.-P. Koenig and G.Webelhuth (eds.), *Lexical And Constructional Aspects of Linguistic Explanation*, pages 215–229, CSLI Publications.

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

Johnston, Michael. 1998. 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 '98, pages 624–630, Stroudsburg, PA, USA: Association for Computational Linguistics.

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

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

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

Kranstedt, Alfred, Lcking, Andy, Pfeiffer, Thies, Rieser, Hannes and Wachsmuth, Ipke. 2006. Deixis: How to Determine Demonstrated Objects Using a Pointing Cone. In Sylvie Gibet, Nicolas Courty and Jean-Franois Kamp (eds.), *Gesture in Human-Computer Interaction and Simulation*, volume 3881 of *Lecture Notes in Computer Science*, pages 300–311, Springer Berlin / Heidelberg.

Ladd, Robert D. 1996. *Intonational Phonology (first edition)*. Cambridge University Press.

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

Liberman, Mark and Prince, Alan. 1977. On Stress and Linguistic Rhythm. *Linguistic Inquiry* 8(2), 249–336.

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

Marshall, Ian and Sáfár, Éva. 2004. Sign Language Generation in an ALE HPSG. In Stefan Müller (ed.), *Proceedings of the HPSG-2004 Conference, Center for Computational Linguistics, Katholieke Universiteit Leuven*, pages 189–201, Stanford: CSLI Publications.

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

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