An HPSG Approach to Synchronous Deixis and Speech

Katya Alahverdzhieva Alex Lascarides
K.Alahverdzhieva@sms.ed.ac.uk alex@inf.ed.ac.uk
School of Informatics, University of Edinburgh

1 Introduction

The usage of deixis is highly pervasive in everyday communication. Through definite referring expressions, pronouns and pointing gestures with the head and hand, people engage the context of the communicative event in the utterance interpretation. In this paper, we concentrate on deictic (pointing) gestures performed by the hand,[1] and we demonstrate that deixis, along with speech, can be described in terms of standardly used formal methods in linguistics, namely constraint-based formal grammars and compositional semantics. In particular, we use the form of the deictic gesture, the form of the synchronous speech and the utterance-specific time and space to combine speech and gesture into a single tree and to map them them to an underspecified meaning representation. 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). RMRS 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 ULF; also, RMRS 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 gestural meaning since even through discourse processing the semantic predications yielded by gestural form may remain unresolved.

Deictic gestures demarcate spatial reference by projecting the hand to a region that is proximal or distal in relation to the speaker's origo. The pointing does not necessarily identify a concrete referent that is present in the communicative situation. It can identify an abstract individual or object placed by the speaker on a virtually created map—also known as *abstract pointing* (McNeill, 2005)—or it can highlight a word or phrase from the simultaneously produced speech—known as *nomination pointing* (Kendon, 2004). Since these distinct types of deictic gestures might have distinct semantic effects, we can represent them in a typed hierarchy, thereby allowing a type to share by inheritance information with its supertypes (Pollard and Sag, 1994).

[1] From now on, we shall call this *deixis*.

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; and 2. attachment ambiguity, which involves the syntactic integration of a deixis daughter to the synchronous, semantically related, speech daughter. The form features 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 the ambiguities in the region identified by the 'pointing cone' (Kranstedt et al., 2006), we do not aim to resolve them as they have no effects on multimodal perception.

Following Lascarides and Stone (2009), we formalise the location of the tip of the index finger with the constant $\vec{c}$ which, combined with the deixis form features, determines the spatial region $\vec{p}$ designated by the gesture; e.g., a stationary gesture of 1-index would make $\vec{p}$ a line or even a cone that projects from $\vec{c}$ in the same direction as the index finger. To account for the fact that the gestured space is not necessarily identical to the denoted space, we are using the function v to map the physical space $\vec{p}$ identified by the gesture to the actual space $v(\vec{p})$ it denotes; e.g., in (1)[2] the referent is at the exact coordinates in the visible space the gesture points at, i.e., v is equality. In contrast, in (2) the referent is not physically present and so v does *not* resolve to equality. Deixis is also used by American Sign Language to set up nominals in the virtual space: if the individual demarcated by the nominal is physically present, the speaker identifies him by a pointing gesture to its location, and otherwise the speaker points at an arbitrary location in the frontal space (Cormier et al., 1999).

(1) [$_{PN}$You] guys come from tropical [$_{N}$countries]
Speaker C turns to the right towards participant D pointing at him using Right Hand (RH) with palm open up

(2) I [$_{PN}$enter] my [$_{N}$apartment]

[2] In the utterance transcription, the speech signal aligned with the expressive part of the gesture, the so called *stroke*, is underlined with a straight line, and the signal aligned 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).

RH and Left Hand (LH) are in centre, palms are open vertically, finger tips point forward; along with "enter" they move briskly downwards.

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. "*That book* is mine"; 2. "*That man* was a friend of mine"; 3. "I find *that period of American history* fascinating". In 1., there is one-to-one correspondence between the deixis denotation and the physical artefact book, and they are thus bound by *FormIdentity*. 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.

The choices of attaching gesture to speech are also not unique and affect the gestural interpretation. In utterance (2), for instance, there is no information coming from the form of the hand, nor from its relative timing to determine whether it should attach to "enter" only, or 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.

We argue that these various levels of ambiguity can be captured by well-established mechanisms for producing *underspecified logical forms* (ULFs) which give a very abstract representation of what the gesture means abstracted away from context. In particular, we use *Robust Minimal Recursion Semantics* (RMRS) (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 to the right while saying "I turn right on Ames Street" identifies the street x or the event of turning e.

Despite the ambiguities, the process of attachment is constrained—e.g., despite that the subject head daughter in (2) is performed with the same distance from "enter" as the object, an attachment to it is disallowed since it would never produce the intended meaning in context.

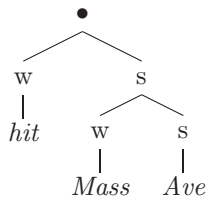
2 Speech-Deixis Synchrony

Due to the lack of an accepted methodology of how to establish the synchrony of two modalities,[3] we assume that synchrony consists in *attaching gesture to the semantically related speech phrase in the syntactic tree that, using standard semantic composition rules, yields a* ULF *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.

2.1 Corpus Investigation

The Autosegmental-Metrical (AM) phonology (Ladd, 1996) underpins our underlying assumptions about speech-gesture interaction, and hence also the annotation schema and the formalisation of grammar construction rules. In the 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, it is the rightmost one, i.e., the metrical structure is right branching as displayed in (3). This is overridden by narrow focus where the structure can also be left-branching. 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).

(3)



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

Hypothesis 1. *Deictic gesture can be predicted from the nuclear prominence in speech: in case of broad-focused utterances, it aligns with the nuclear accent, and in case of early pre-nuclear rise, it aligns with the pre-nuclear accent.*

[3] As demonstrated by (2) the temporal performance of one mode relative to the temporal performance of the other is insufficient for deriving the possible meaning representations.

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 movements in terms of communicative vs. non-communicative, assigning them a category–depicting, deictic—and segmenting them into discrete phases—preparation, stroke, hold and retraction to rest.

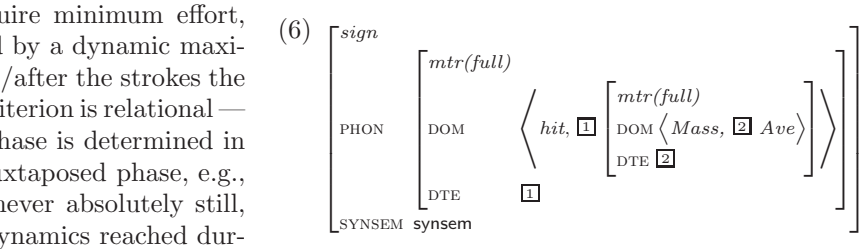
The gesture segmentation was based on formal and functional criteria. The formal criteria involved the dynamic profile of the hand, i.e., the effort employed by the hand. Any sudden change in the hand dynamics signals a transition to a new phase. More specifically, preparations and retractions require minimum effort, the stroke is usually characterised by a dynamic maximum, and during the holds before/after the strokes the hand is held still. Note that this criterion is relational — the lower/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 104 deictic gesture strokes,[6] 103 of which were overlapped by at least one nuclear/pre-nuclear accented word (not simply the accent itself). Gestures of longer duration were often marked by a combination of a nuclear and non-nuclear and/or nuclear and pre-nuclear accented words. Importantly, the results confirmed our hypothesis: the part of the gesture carrying the meaning is constrained by the meaningful accent in speech. This is attested in the broad-focused utterance (4) and in the narrow-focused utterance (5), a continuation of (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, 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

(5) And then I [$_N$turn] *[pause]* [$_N$left] on [$_{NN}$Mass] Ave
LH is held in the same position as in (4); 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 (3) maps to the feature structure in (6). 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.

(6)


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 preferred interpretations in context. We encountered six instances which, although syntactically well-formed, did not map to the intended meaning representations due to the fact that the gesture stroke was performed with a few milliseconds positive or negative delay in relation to the semantically preferred speech element. In (7), for instance, the gesture is produced while uttering "Thank you" when obviously the denotation of the hand is identical to that of the computer mouse.

(7) [$_N$Thank] you. [$_{NN}$I'll] take the [$_N$mouse]
RH is loosely closed, index finger is loosely extended, pointing at the computer mouse

These instances of temporal misalignment occurred only in cases where the visible space $\vec{p}$ designated by the gesture was equal to the space $v(\vec{p})$ it denoted; e.g., whereas the temporal misalignment in (7) and (1) is acceptable, the temporal misalignment in (2) would fail to produce the intended LF.

In §4, we propose construction rules that reflect our empirical findings.

[4] http://www.talkbank.org/media/Gesture/Cassell/kimiko.mov
[5] http://corpus.amiproject.org/
[6] Two strokes spanning a single gesture were also possible.

3 Underspecified Semantics

In §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 structure of gesture. Gestures, unlike fully-fledged language systems, are constructed by equally ranked features 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. By recording the deixis form features (the TFS of the deixis in (2) is given in (8)), we stay consistent with the findings in the descriptive literature that the form of the pointing hand is significant for interpreting its meaning in context, e.g., whereas 1-index hand 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).

(8)
$$\begin{bmatrix} \textit{deictic\_abstract} & \\ \text{HAND-SHAPE:} & \text{open-flat} \\ \text{PALM-ORIENTATION:} & \text{vertical} \\ \text{FINGER-ORIENTATION:} & \text{forward} \\ \text{HAND-MOVEMENT:} & \text{away-body-centre} \\ \text{HAND-LOCATION:} & \vec{c} \end{bmatrix}$$

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 in Figure 1. Some values such as *open-closed* account for a change in the form.

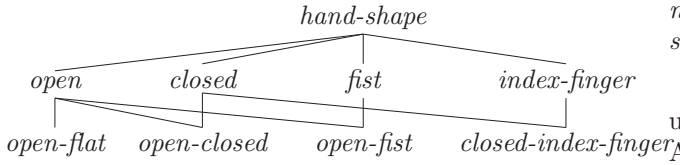


Figure 1: Fragment of the Sort Hierarchy of *hand-shape*

The compositional semantics of deictic gesture involves producing a set of underspecified predications in the RMRS notation (the RMRS of the deixis in (2) is shown in (9)). 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., in (9) the fact that the $sp\_ref(i)$ predicate takes an ARG1 argument is licensed through a_2.

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 is a function that maps the physical space $\vec{p}$ to the space $v(\vec{p})$ in denotation. 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.

(9)
$$\begin{aligned} & l_1 : a_1 : \textit{deictic\_q}(i)\ RSTR(a_1, h_1)\ BODY(a_1, h_2) \\ & l_2 : a_2 : sp\_ref(i)\ ARG1(a_2, v(\vec{p})) \\ & l_2 : a_3 : \textit{hand\_shape\_open\_flat}(e_0)\ ARG1(a_3, i) \\ & l_2 : a_4 : \textit{palm\_orient\_vertical}(e_1)\ ARG1(a_4, i) \\ & l_2 : a_5 : \textit{finger\_orient\_forward}(e_2)\ ARG1(a_5, i) \\ & l_2 : a_6 : \textit{hand\_move\_away\_body\_centre}(e_3)\ ARG1(a_6, i) \\ & h_1 =_q l_2 \end{aligned}$$

4 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), and also rules that account for the fact that although syntactically well-formed, some multimodal utterances need the output of a pragmatics processor to resolve to the intended LF in context. In this section, we detail two construction rules: a basic rule that attaches deixis to a single prosodic word and a rule that integrates defeasible constraints with the view of producing the right LFs.

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 2. 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 are in an overlap temporal relation, i.e., $end(D) > start(S)$ and $end(S) > start(D)$. The SYNSEM values of the deictic daughter are encoded as detailed in §3: 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 $\vec{p}$ 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*.

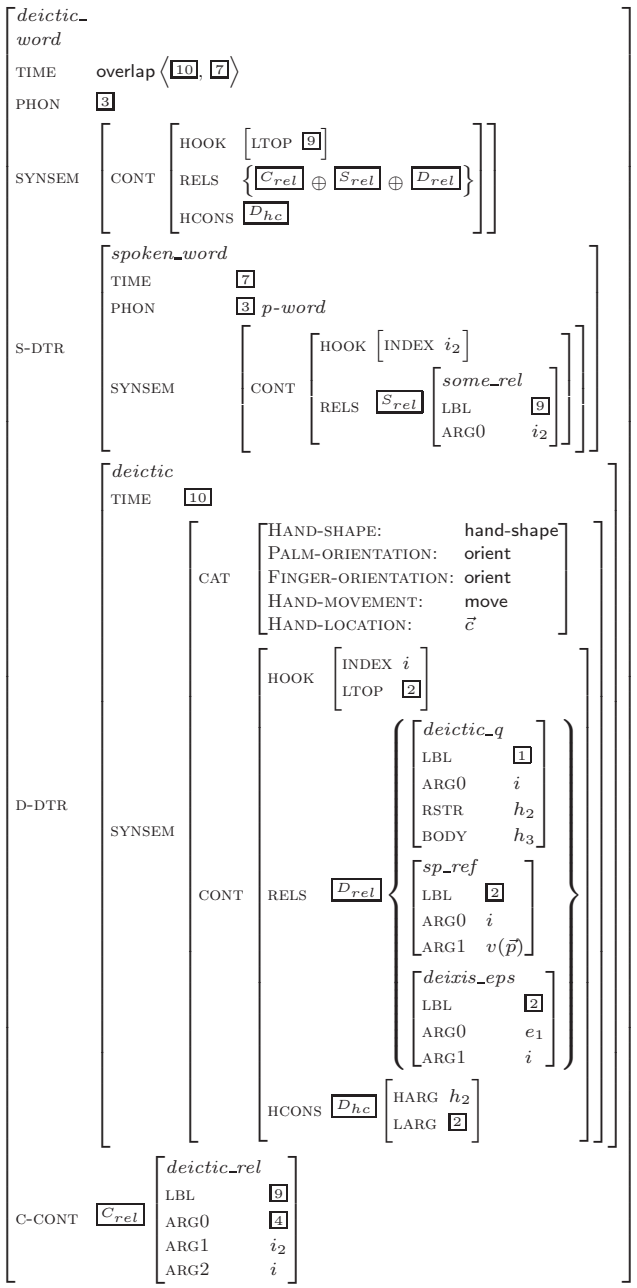


Figure 2: Deictic Prosodic Word Constraint

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 §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 mouse, 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 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 (10), this rule would produce a tree where the deixis is attached to the prosodic word "hallway".

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

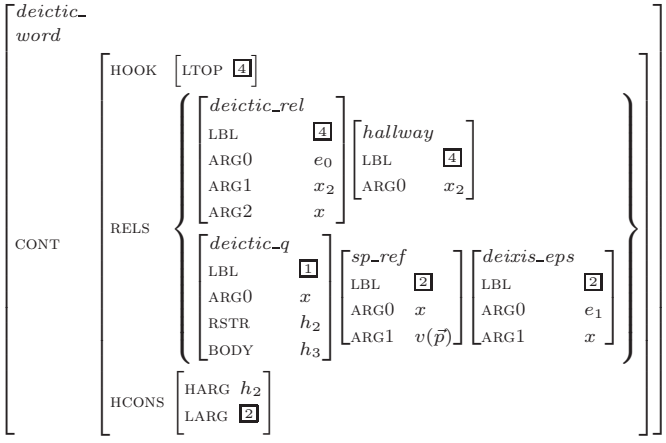


Figure 3: Semantic Composition for Deixis + "hallway"

For the sake of space, in Figure 3 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.

The occurrence of instances like (7) necessitates the introduction of a rule that makes the constraint in Figure 2 defeasible as follows:

Rule 2. *Deictic gesture attaches to a spoken word whose temporal performance is adjacent to that of deixis if* **v** *resolves to equality.*

As attested by (7), this temporal relaxation is applied only with salience of individuals in the communicative event and it is thus necessary in utterances such as (1) and (7) where the gesture's denotation is physically present in the visible space. Applied to (7), this rule would account for the multimodal utterance "the mouse" + deixis. An alternative interpretation following Rule 1 would involve integrating the gesture with the temporally co-occurring prosodically prominent "Thank you". This, however, cannot resolve to "Thank you for the mouse" since "Thank you" is related to the previous discourse—projecting the presentation in slide show mode in response to the speaker's request.

The temporal misalignment between the performance of deixis and the performance of speech also illustrates an important finding about deictic gesture— it is not only that its truth conditions depend on the context in which the sentence was uttered (this is true for all deixis expressions, not only pointing gestures), but also its grammaticality is informed by the context, i.e., the temporal relaxation is permitted only with salience of individuals. In so doing, the grammar architecture is not strictly pipelined since the output from pragmatics is input to the syntax.

5 Conclusions

In this paper, we presented a constraint-based analysis of multimodal communicative signals consisting of deictic gestures and speech. Our approach re-uses standard devices from linguistics to map multimodal form to meaning, thereby accounting for the gestural ambiguity by means of established underspecification mechanisms. To specify this 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 two rules: a basic rule accounting for a multimodal word, and a defeasible constraint accommodating the fact that it is not prosody or syntax but rather pragmatics that informs the grammaticality of certain multimodal expressions. In the full paper, we shall present the full scope of the theoretical framework, as well as the coverage against multimodal data. Essentially, with this paper we demonstrated that the constraint-based grammar framework of HSPG is expressive enough to produce multimodal LFs from syntax.

6 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 current version. The authors are also grateful to Sasha Calhoun, Jean Carletta, Jonathan Kilgour, Ewan Klein and Mark Steedman. Any mistakes and inaccuracies are our own.

References

Brenier, J. and Calhoun, S. 2006. Switchboard Prosody Annotation Scheme. Internal publication.

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

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

Copestake, A. 2007. Semantic composition with (robust) minimal recursion semantics. In *DeepLP '07: Proceedings of the Workshop on Deep Linguistic Processing*, ACL.

Copestake, A. and Flickinger, D. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second Linguistic Resources and Evaluation Conference*, Greece.

Copestake, A., Flickinger, D., Sag, I. and Pollard, C. 2005. Minimal Recursion Semantics: An introduction. *Journal of Research on Language and Computation* 3(2–3), 281–332.

Cormier, K., Wechsler, S. and Meier, R. 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*, pp 215–229, CSLI Publications.

Giorgolo, G. and Verstraten, F. 2008. Perception of speech-and-gesture integration. In *Proceedings of the International Conference on Auditory-Visual Speech Processing*.

Johnston, M. 1998. Unification-based multimodal parsing. In *Proceedings of the 36th Annual Meeting of ACL and 17th International Conference on CL*, ACL.

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

Klein, E. 2000. Prosodic Constituency in HPSG. In *Grammatical Interfaces in HPSG, Studies in Constraint-Based Lexicalism*, CSLI Publications.

Kopp, S., Tepper, P. and Cassell, J. 2004. Towards integrated microplanning of language and iconic gesture for multimodal output. In *Proceedings of the 6th International Conference on Multimodal interfaces*, State College, USA.

Kranstedt, A., Lücking, A., Pfeiffer, T., Rieser, H. and Wachsmuth, I. 2006. Deixis: How to Determine Demonstrated Objects Using a Pointing Cone. In *Gesture in Human-Computer Interaction and Simulation*, Springer Berlin/ Heidelberg.

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

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

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

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

Marshall, I. and Sáfár, É. 2004. Sign Language Generation in an ALE HPSG. In S. Müller (ed.), *Proceedings of the HPSG-2004 Conference, Center for Computational*

Linguistics, Katholieke Universiteit Leuven, pp 189–201, Stanford: CSLI Publications.

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

Pollard, C. and Sag, I. A. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press & CSLI Publications.

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