TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Integration of Speech and Deictic Gesture in a Multimodal Grammar

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

Résumé. Dans cet article, nous présentons une analyse à base de contraintes de la relation forme-sens des gestes déictiques et de leur signal de parole synchrone. En nous basant sur une étude empirique de corpus multimodaux, nous définissons des généralisations décrivant les énoncés multimodaux bien formés qui soutiennent le sens voulu dans le contexte final. Plus précisément, nous formulons une grammaire multimodale dont les règles de construction utilisent la prosodie, la syntaxe et la sémantique de la parole, la forme et le sens du signal déictique, ainsi que le timing de la parole et la deixis afin de contraindre la production d'un arbre de syntaxe qui corresponde á une représentation unifiée du sens pour l'action multimodale. La contribution de notre projet est double : nous ajoutons aux ressources existantes pour le TAL un corpus annoté de parole et de gestes, et nous créons un cadre théorique pour la grammaire au sein duquel la composition sémantique d'un énoncé découle de la synchronie entre geste déictique et parole.

Abstract. In this paper we present a constraint-based analysis of the form-meaning mapping of deictic gesture and its synchronous speech signal. Based on an empirical study of multimodal corpora, we capture generalisations about well-formed multimodal utterances that support the preferred interpretations in the final context-of-use. More precisely, we articulate a multimodal grammar whose construction rules use the prosody, syntax and semantics of speech, the form and meaning of the deictic signal, as well as the temporal performance of speech relative to the temporal performance of deixis to constrain the derivation of a single multimodal tree and to map it to a meaning representation. The contribution of our project is two-fold: it augments the existing NLP resources with annotated speech and gesture corpora, and it also provides the theoretical grammar framework where the semantic composition of an utterance results from its speech-and-deixis synchrony.

Mots-clés : Deixis, parole et geste, grammaires multimodales
.

Keywords: Deixis, speech and gesture, multimodal grammars.

1 Introduction

Through the physical co-location of people known as *co-presence* (Goffman, 1963), individuals convey information to each other using various meaningful and visibly accessible channels such as the arrangements of the bodies in the shared space, the bodily orientations and the pointing signals of their hands and heads. In recent years, it has become commonplace to integrate input from different modalities of interaction, such as natural language and deictic gesture, in multimodal systems for the purposes of human-robot interaction (Giuliani & Knoll, 2007), or pen-based applications (Oviatt *et al.*, 1997), (Johnston, 1998).

In this paper, we demonstrate that speech and co-speech deictic gesture can be integrated into a constraint-based grammar by exploring the linguistic information of the speech signal (i.e., its prosody, syntax and semantics), the form and meaning of the deictic signal, and their relative temporal performance. Our overall aim is to articulate the mapping from the form of the multimodal action to its (underspecified) meaning, using established methods from linguistics such as constraint-based syntactic derivation and semantic composition. To specify this mapping, we develop a grammar for speech and co-speech deictic gesture (referred to as *deixis*) which captures empirically extracted generalisations about syntactically and semantically well-formed multimodal actions that convey the intended meaning in the specific context. We have already captured constraints on depicting dimensions via a constraint-based grammar (Alahverdzhieva & Lascarides, 2010). Here were are going to demonstrate that constraint-based grammars can represent the form-meaning mapping for deictic dimensions too.

This paper is structured as follows: in §2, we start off with an overview of deictic gesture, including its range of usage that is accounted for in the formal modelling. In §3 we put forth the empirical investigation aimed at extracting generalisations about the speech-deixis interaction. Finally, in §4 we introduce the representation of gesture form and its mapping to meaning, and we formalise the empirical findings in construction rules for the integration of speech signals and deictic gestures.

2 Data

2.1 Deixis Background

Our focus of study are deictic (or pointing) gestures performed spontaneously by the hand along with the speech signal. Deictic gestures demarcate spatial reference in Euclidean space by projecting the hand to a region that is proximal or distal in relation to the speaker's origo. Through deixis, 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. We shall come back to this property of deixis in §4 when detailing multimodal grammaticality.

Note that by "gesture" we mean the kinetic peak of the hand movement that conveys the gesture's meaning—the so called *stroke*. What is intuitively recognised as a gesture, is known as a *gesture phrase*. It contains the following *phases*: a non-obligatory *preparation* (the hands are lifted from the rest position to the frontal space to perform the semantically intended motion), a non-obligatory *pre-stroke hold* (the hands are sustained in a position before reaching the kinetic peak), an obligatory stroke, a non-obligatory *post-stroke hold* (the hands sustain their expressing position) and an obligatory *retraction* to a rest position. The deictic stroke might be static (the pointing forelimbs are stationary in the expressive position) or dynamic (gesture's meaning is derived from a movement of the pointing forelimbs).

2.2 Range of Usage

The deictic signal on its own is ambiguous with respect to the region pointed out and the syntactic and semantic relation between speech and deixis. To clarify the region's ambiguity, consider the following example: when pointing in the direction of a book with an extended index finger (1-index), does the region demarcated by 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 referent. Our formal model does not intend to solve this

ambiguity since it has no effects on multimodal perception. Certain ambiguities in the interpretation of deixis remain unresolved even in context, just as certain ambiguities can be tolerated in purely linguistic utterances.

Following Lascarides & Stone (2009), we formally regiment the location of the tip of the index finger with the constant $\vec{c}$, and $\vec{c}$ combined with the hand shape, orientation and movement determines the region $\vec{p}$ designated by the gesture—e.g., a stationary stroke with hand shape 1-index will make $\vec{p}$ a line (or even a cone) that starts at $\vec{c}$ and continues in the direction of the index finger. To account for the fact that the gestured space is not necessarily equal to the denoted space, we are using the function v to map the physical space $\vec{p}$ designated by the gesture to the actual space $v(\vec{p})$ it denotes; e.g., in (1),[1] taken from a longer multi-party conversation, v would resolve to equality since the referent identified by the hand is at the exact coordinates in the visible space the gesture points at. In contrast, in (2), extracted from a conversation where the speaker describes the layout of her flat, v would *not* resolve to equality since the referent "apartment" is not available in the communicative context.

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

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

Further ambiguities arise from the fact that the choices of syntactic "attachment" of the gesture to the synchronous, semantically related, linguistic phrase are not unique, and this has effects on the gestural interpretation. In (2), for instance, there is no information coming from the form of the hand, nor from its timing relative to speech, whether it should attach to "enter" only or "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 house, but also to the entrance door which by the hand shape is rectangular. This observation flags up an important claim in this work, namely that the attachment of gesture to the temporally co-occurring speech elements is too restrictive since it bars the possibility of fully exploring the semantic content of speech, the semantic content of deixis, and their mutual relations.

We further stated that there is a range of relations between the speech signal and the pointing signal which results from the fact that deixis can denote distinct features of the qualia structure (Pustejovsky, 1995) of the referent, i.e., the gesture relates through a range of relations with the various roles of polysemous words. 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 an identity between the gesture denotation and the physical artifact book, so we assume they are bound by *FormIdentity*. In 2., there is a reference transfer from the book to the author, and so the gesture denotes the creative agent of the book rather than the book itself. We therefore say that there is an *AgentiveRelation* between the deixis and the speech NP, and finally in 3., the deixis refers to the content of the book, and so the deixis denotation is rather related through a *ContentRelation* with the speech denotation. More ambiguities can be found in the context of the co-occurring speech: does the pointing gesture while uttering "We turn right" identify the event e of turning or the direction x? Our formal model fully supports ambiguity and partial meaning since we map deictic form to an underspecified meaning representation whose main variable can resolve to either e or x in context, and we also connect speech and deictic referents in the grammar through an underspecified relation *deictic_rel(s,d)* between the content s of speech and the content d of deixis where there is a grammar construction rule that says that s is synchronous with d. The way this relation resolves is a matter of discourse context, and some of its possible values are *FormIdentity, AgentiveRelation* and *ContentRelation*.

In this section we gave an overview of deictic gestures, and we also introduced the main challenges arising from deixis ambiguity. In §3 we turn to the problem of how deixis and speech interact at the level of linguistic form (prosody) and meaning.

[1]For the utterance transcription, we have adopted the following convention: the speech signal aligned with the stroke is underlined, and the signal aligned with a post-stroke hold 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).

3 Empirical Investigation

Our motivation for unifying speech and gesture into a grammar stems from the descriptive accounts that gesture takes an integral part in language production and language comprehension (Kendon (2004), McNeill (2005) inter alia). We thus analyse deixis in *synchrony* with speech, as a mapping from form to an underspecified logical form (ULF) with the ULF being resolved in context into a complete and specific interpretation through complex pragmatic processing (resolving a ULF into a complete interpretation is beyond the scope of this paper, but see Lascarides & Stone (2009)). Due to the controversial findings concerning the temporal alignment of speech and gesture, Alahverdzhieva & Lascarides (2010) proposed the following definition of synchrony, which considers only qualitative factors coming from form and meaning:

Definition 1 Synchrony. *The choice of which speech phrase a gesture stroke is synchronous with is guided by: i. the final interpretation of the gesture in specific context-of-use; ii. the speech phrase whose content is semantically related to that of the gesture given the value of (i); and iii. the syntactic structure that, with standard semantic composition rules, would yield an underspecified logical formula (*ULF*) supporting (ii) and hence also (i).*

The gestural signal and the spoken signal are closely related on both the level of form and of meaning. We view form as a matter of temporal performance of one mode relative to the temporal performance of the other mode: there is increasing evidence in the literature that gesture performance is constrained by the prosody of speech, both speech and gesture are integrated into a common rhythmical system, and the perception of one mode is dependent on the performance of the other—e.g., Kendon (1972), Loehr (2004), Giorgolo & Verstraten (2008). We shall perform some experiments to validate these claims, and hence equip our grammar with the constraints on the mapping between form and meaning of co-speech deictic actions that stem from the relative temporal performance of gesture and speech, and prosody (among other factors), where these constraints model our empirical findings in multimodal corpora.

3.1 Prosody Background

We adopt the Autosegmental-Metrical (AM) theory (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, 2000).

In the AM framework, nuclear prominence results from the following operations: (a). mapping a syntactic structure to a binary metrical tree; (b). 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 (Liberman & Prince, 1977, p.257) as shown in Definition 2; and (c). tracing the path dominated by s nodes.

Definition 2 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. For instance, (3)[2] illustrates the metrical tree for "fasten a cloak" in its broad focused reading with the nuclear accent being on the word entirely dominated by s nodes—"cloak". Early pre-nuclear rise on the left of the nuclear node is also possible, and it is signalled through its acoustic properties rather than its relative position in the metrical tree.

(3)

[2]The example is taken from Klein (2000)

3.2 Hypothesis and Data Annotation

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

Hypothesis 1 *Deictic gestures align with the nuclear accents in speech both in the default case of broad focus, and in case of narrow focus. In case of early pre-nuclear rise, deictic gestures align with the pre-nuclear pitch accents.*

To test the validity of our hypothesis, we used two multimodal corpora: a 5.53 min recording from the Talkbank Data,[3] and observation IS1008c, speaker C from the AMI corpus.[4] 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. Annotation on both corpora proceeded in two independent stages: annotation of prosody and annotation of gesture.

Prosody Annotation The annotation of prosody was done in Praat (Boersma & Weenink, 2003) and it was consistent with the guidelines of the prosody annotation of the Switchboard corpus (Brenier & Calhoun, 2006). It included marking the following layers:

1. *Orthographic Transcription.*
2. *Pitch Accents.* Words were unambiguously associated with at least one accent of the following type: *nuclear:* the accent of the prosodic phrase that is structurally, and not phonetically perceived as the most important one; *pre-nuclear:* an early emphatic high rise characterised by a high pitch contour; *non-nuclear:* unlike nuclear accents, non-nuclear accents are perceived on the basis of their phonetic properties, and the rhythm of the sentence; *none*: a non-discernible accent in a phrase; *?*: uncertainty concerning the presence of an accent.
3. *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: *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; *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; *backchannel*: short phrases containing only fillers such as “er”, “um”, “you know”, etc.

Past annotation tasks of the Switchboard corpus (see Table 1) have shown that this annotation strategy is reliable: the κ measurement is calculated from the number of times the annotators agree plus the number of times they are expected to agree by chance; it is believed that $0.67 < \kappa < 0.80$ is fair, and $\kappa > 0.8$ shows good reliability (Carletta, 1996).

	All Types	**Absence/Presence**
Accents	0.8	0.8
Boundaries	0.889	0.91
Words	(752)	

Table 1: Inter-coder reliability for accents and phrase boundaries & for the presence/absence of an accent/boundary in *kappa (κ)* (Calhoun, 2006)

Gesture Annotation We used the Anvil labelling tool (Kipp, 2001) to annotate the hand movements and gesture phases. 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.

1. *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 whether the communicative signal belonged to a deictic or to a different dimension.

[3] `http://www.talkbank.org/media/Gesture/Cassell/kimiko.mov`
[4] `http://corpus.amiproject.org`

2. *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, that is, 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. 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; contrarily, post-stroke holds are more likely to occur with fluent speech when the speaker elaborates on the content reached during the stroke.

We used this gesture annotation schema on a single observation of the multimodal corpus of Loehr (2004) where we reached inter-annotator agreement as shown in Table 2. The segmentation column shows agreement over the presence/absence of an element within a certain time slice, and the coding column refers to agreement over 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).

	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%
Gesture phase	0.8864	0.8971	0.662	0.7	75%

Table 2: Inter-coder reliability for gesture coding agreement & segmentation agreement in *Cohen's kappa (κ) and in corrected kappa (κ)*

3.3 Results and Discussion

In relation to our hypothesis, we searched for the types of accents overlapping a deictic gesture stroke. 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. Strokes overlapping a combination of non-nuclear and nuclear accented words were also common. Essentially, the empirical analysis confirmed the expected alignment between the nuclear prominent word (not simply the nuclear accent) and the gesture stroke both in case of broad focus, and in case of narrow-focused utterance; e.g., (4) is a broad-focused utterance with the nuclear accent being on the right-most word, and (5), a continuation of (4), displays narrow focus with the nuclear accent pointing to the first word of the prosodic phrase–"left". The interaction between prosodic prominence and gesture stroke appears to be on the level of Information Structure (IS): nuclear prominence along with gesture stroke aligns with the focused (kontrastive)[5] elements that push the communication forward, and not with those available from the background. This prediction has its grounds in the descriptive literature of gesture where "a break in the continuity" (Givón, 1985) of the narrative implies "highest degree of gesture materialisation" (McNeill, 2005, p.55).

(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
Hands are held in the same position as in (4), then along with "left" RH moves to the left periphery over LH, RH is vertically open

The single counterexample in the corpus to Hypothesis 1 concerns the first gesture in (6): at this stage we remain agnostic as to why this misalignment occurred. As long as it is not a recurrent feature found over a larger amount of data, we would rather attribute it to impreciseness of annotation than to a general phenomenon to be considered in a model of multimodal actions.

[5] In the IS literature kontrast designates "parts of the utterance—actually, words—which contribute to distinguishing its actual content from alternatives the context makes available." (Kruijff-Korbayová & Steedman, 2003)

(6) [$_{NN}$Between] the living [$_{N}$room] and *[pause]* the [$_{N}$study] and the *[pause]* [$_{N}$bedroom]
Hands are in the front centre, bent in elbows, palms are open, vertical, facing each other; along with "between", they perform a loose sweeping movement to the right periphery, then LH moves away to the left upper centre with palm vertical, finger tips oriented forward; along with "the study", RH is moved in parallel to LH, as if both hands place a rectangular object in space

Our results report on the interaction between speech and deixis on the level of *form*. Our overall aim is to account both for the syntactic and the semantic well-formedness of the multimodal signal. In other words, the ULFs 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 six 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 (7) 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". An 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.

   (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

2. The speech signal that is semantically related to the gesture is not prosodically prominent; e.g., in (8) the deictic gesture aligns temporally with the nuclear prominent "said", when in fact, it identifies the individual pointed at and it thus resolves the pronoun coming from speech.

   (8) And a as she [$_{N}$said], it's an environmentally friendly uh material
   Speaker C extends right hand palm supine towards the speaker B

These instances of temporal/prosodic 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, i.e., v was equality. Otherwise, any synchronicity between a deictic gesture and an individual not present at the exact coordinates of the gesture space would fail to produce the intended LF in the specific context. For instance, it is perfectly acceptable for the gesture stroke in (1) to be performed a few milliseconds later so that it aligns with "come" or even with "tropical countries" without blocking the interpretation where the hand denotes the addressee. In (2), however, if the deixis were performed along with "I", the LF would fail to resolve to "apartment".

In this section we presented an empirical study that intended to shed light on the deixis-speech interaction at the prosodic level. Using annotated multimodal corpora, we established that the nuclear and pre-nuclear prominence in speech are predictive for the deixis realisation. We also learnt that the occurrence of temporal/prosodic misalignment is restricted to marking salience of individuals present in the communicative act. In §4.2, we provide grammar rules that reflect these generalisations about the corpus data.

4 Formal Modelling of Speech and Deixis

This section details the theoretical framework for the integration of spoken and deictic signals. We start off with the formal representation of deixis, and how its form maps to meaning. We then proceed with construction rules for the speech-deixis integration.

4.1 Deixis Form and Meaning

It is now commonplace to formally regiment gesture form with Typed Feature Structures (TFSs), where each feature value pair corresponds to an aspect of form (Johnston, 1998), (Kopp *et al.*, 2004). This representation

captures the fact that gesture, unlike language, is not hierarchically structured and its meaning cannot be computed from the meaning of its parts (McNeill, 2005). We use as fine-grained an analysis as possible: we consider that the shape of the hand, the orientation of the palm and fingers, the hand movement, and also the location of the tip of the index finger at the spatio-temporal coordinates $\vec{c}$ are the distinct classes of form that potentially have semantic effects; e.g., the TFS representation of the deixis in (9) is shown in (10).

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

(10)

communicative_gesture_deictic	
HAND-SHAPE:	open-flat
PALM-ORIENTATION:	vertical
FINGER-ORIENTATION:	forward
HAND-MOVEMENT:	away-centre-left
HAND-LOCATION:	$\vec{c}$

To capture the deictic ambiguities (see §2.2), we use the semantics description language of Robust Minimal Recursion Semantics (RMRS) (Copestake, 2007) since it is highly flexible about the semantic underspecification it supports: in RMRS, one can leave the main predicate underspecified until resolved by further context. In this way, we can elegantly capture the fact that the form of a deictic gesture alone does not fully determine its content. Form does not determine, for instance, whether the gesture denotes an individual or an event, but rather contextual information is needed as well to infer this aspect of the gesture's (pragmatic) interpretation.

For deictic gestures, producing ULFs in RMRS involves defining a set of *Elementary Predications (*EP*s*) with underspecified scope and main variable; e.g., the RMRS representation of the gesture in (9) is shown in (11). Each EP is associated with a label ($l_1 ... l_n$) and an anchor ($a_1 ... a_n$). The label is not necessarily unique and it identifies the scopal positions of the predicate in the context-resolved LF (EPs that share a label are joined by conjunction with the label when specifying the scopal position of the conjunction). The anchor, which is unique to each EP, is used as a locus for adding arguments to the main predicate so that in case of shared labels, an argument can be uniquely associated with its predication.

(11) h_0
$l_1 : a_1 : deictic\_q(i)\ RSTR(a_1, h_1)\ BODY(a_1, h_2)$
$l_2 : a_2 : sp\_ref(i)\ ARG1(a_2, v(\vec{p}))$
$l_2 : a_3 : hand\_shape\_open\_flat(e_0)\ ARG1(a_3, i)$
$l_2 : a_4 : palm\_orient\_vertical(e_1)\ ARG1(a_4, i)$
$l_2 : a_5 : finger\_orient\_forward(e_3)\ ARG1(a_5, i)$
$l_2 : a_6 : hand\_move\_away\_centre\_left(e_5)\ ARG1(a_6, i)$
$h_1 \text{QEQ} l_2$

We defined deictic gestures as providing spatial reference of an individual or event in the physical space $\vec{p}$. This is expressed by the two-place EP $l_2 : a_2 : sp\_ref(i)\ ARG1(a_2, v(\vec{p}))$ where the first argument is an underspecified referent i, and the second argument (linked through the anchor a_2) is $v(\vec{p})$ with v being a function that maps the physical space to the actual space in denotation. In context, the underspecified variable i may resolve to an individual x as in (9), or to an event e as in (2). Further, we map the feature-value pairs to EPs which serve as intersective modifiers of the referent. The deixis form features are needed as they have effects on how the predication may resolve in context: whereas an open hand supine often serves a meta-narrative function such as giving the floor or offering an instance on the open hand, 1-index finger rather individuates the object pointed at (Kendon, 2004). For consistency with the English Resource Grammar (ERG) (Copestake & Flickinger, 2000) where individuals are bound by quantifiers, we use the quantifier *deictic_q* to quantify over the spatial referent. Holes (h_i) are used to represent scopal arguments whose value is not fully determined by syntax. The admissible pluggings are specified in terms of scopal constraints (QEQ) between holes and labels. Finally, a top label h_0 is added to the whole formula.

4.2 Rules for Combining Deixis and Speech in the Grammar

The rules for integrating speech and deixis envisage full coverage of the multimodal constructions in the corpora. Our rules balance between the strict constraints imposed by prosody and the non-unique attachments permitted by syntax which reflect the possible interpretations in context (recall (2) and the related discussion).

We specify the following constraints on speech-deixis well-formedness:

Definition 2.1 Deictic Prosodic Word Constraint. *Deictic gesture attaches to the nuclear or pre-nuclear accented word whose temporal performance overlaps with the temporal performance of the deictic stroke.*

This rule accounts for cases where the deictic stroke overlaps a single nuclear or pre-nuclear accented word which is semantically related to the stroke (recall that 98.85% of our corpus examples feature such overlap). Applied to (9), this rule involves building a single tree out of "hallway" and deixis as displayed in Figure 1. For the sake of readability, we gloss the EPs yielded by the deictic form features as $l_2 : a_3 : deictic\_eps(e_0)\ ARG1(a_3, i)$.

In derivation, the PHON(ology) value of the N daughter is identified with that of the mother. Semantic composition with RMRS involves the following operations over the semantics of the mother (S_M) and the semantics of the two daughters (S_{D1} and S_{D2}):

- $h_0(S_M) = h_0(S_{D1}) = h_0(S_{D2})$ to demonstrate the derivation of a single LF
- EP(S_M) = EP(S_{D1}) $\oplus$ EP(S_{D2}) where $\oplus$ is the append operator
- QEQ(S_M) = QEQ(S_{D1}) $\oplus$ QEQ(S_{D2})

In composition, the underspecified variable i of *sp_ref* resolves to x_1. As argued in §2.2, deictic gesture relates with the synchronous speech through some relation whose resolution is logically co-dependent with the gesture's denotation (Lascarides & Stone, 2009). The construction rule therefore introduces an underspecified relation $deictic\_rel(x_2, x_1)$ between the main variable x_2 of the speech EP and the main variable x_1 of the deixis EP. 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 in speech. The way it resolves is a matter of context and commonsense reasoning, where some of its possible values are *FormIdentity, AgentiveRelation* or *ContentRelation*.

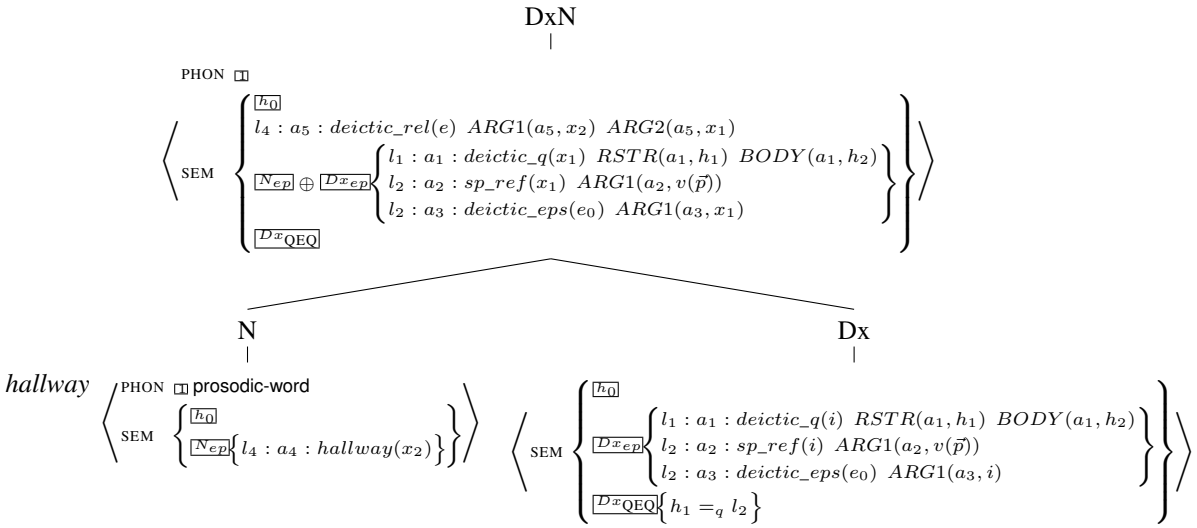

Figure 1: Derivation Tree for Deictic Gesture and the N "hallway"

In §2.2, we claimed that speech-deixis synchrony cannot be determined on the sole basis of the strict temporal alignment since this does not explore the full content of the speech daughter, the gesture daughter and their semantic relation. Moreover, the form of the hand is too ambiguous with respect to the phrase in speech it denotes. To account for the possible meanings of deixis, we provide a rule that extends synchrony beyond the temporal alignment as follows:

Definition 2.2 Deictic Head-Argument Constraint. *Deictic gesture attaches to a nuclear prominent head saturated with the arguments it selects (the external and/or the internal arguments to the head) if there is an overlap temporal relation between the performance of the gesture and the performance of the head-argument construction.*

Applied to (2), this rule would permit attachment to "enter my apartment" and even to "I enter my apartment": the verb head is nuclear accented, its temporal performance overlaps the temporal performance of the gesture, and so the deixis can attach to it upon combining it with the selected internal and/or external arguments. The extension beyond the strict temporal performance is motivated in the synthetic nature of gesture vs. the analytic nature of spoken words (McNeill, 2005), e.g., the event of entering an apartment in (2) is surface realised by a single gesture movement and several linearly ordered lexical items ("I", "enter", "my", "apartment").

The same principle of extending synchrony applies to head-modifier constructions; e.g., neither the form of the deictic signal in (9), nor its temporal performance provide sufficient information as to whether the hand refers to "hallway", to "little hallway" or even to "a little hallway". We therefore augment the grammar as follows:

Definition 2.3 Deictic Head-Modifier Constraint. *Deictic gesture attaches to a nuclear prominent head which had been combined with the selected modifiers if there is an overlap temporal relation between the performance of the gesture and the performance of the head-modifier construction.*

Note that the temporal condition does not impose constraint on whether the gesture overlaps temporally only the modifier, the head or both. We also do not constrain the depth of the modifier phrase, i.e., a gesture could attach to a phrase of n-number recursively ordered adjectives; e.g., the third stroke in (12) could attach both to "small study" and to "rectangular small study".

(12) And [$_N$then] on the [$_N$left] side there's a kind of a rectangular [pause] small [$_N$study]
LH is open flat vertical, along with "then" begins sweeping to the left periphery and is interrupted at a position parallel to the body. The movement is resumed along with "left" when the hand moves further down to the left; palm is still open flat, with fingers slightly extended.

This constraint is also loose with respect to the prosodically prominent element in the head-modifier construction, i.e., we do not restrict attachment to right prominence only. Importantly, the unification of the speech head-daughter and the speech non-head-modifier results in a metrical tree where one of the elements carries the structural accent, which could be the head as in "marketing [$_N$ strategy]" or the non-head as in "[$_N$ right] here". Applied to (9), we integrate deixis into the metrical tree "little hallway" where the prosodically prominent element is "hallway" , and syntactically into a head-modifier construction with "hallway" being the head daughter. Since *deictic_rel* shares the same label as the head, when combining the deictic N "little hallway" and the quantifier "a", both the head noun and the deictic relation would appear within the restriction of the quantifier.

Definition 2.4 Deictic Prosodic Word with Defeasible Constraint. *Deictic gesture attaches to an item whose temporal performance is adjacent to that of the gesture if the mapping v from gestured space $\vec{p}$ to space in denotation $v(\vec{p})$ resolves to equality.*

This temporal/prosodic relaxation rule integrates defeasible constraint with the view of producing LFs that in context would resolve to the intended meaning. As attested by (7) and (8), 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 actual denotation of the gesture's referent. This rule accounts for the fact the grammaticality of the speech-deixis ensemble is informed by the context in which the utterance takes place. In so doing, the grammar architecture is not strictly pipelined since the output of the pragmatics module serves as input to the syntax.

Of course, equality between the gestured space and the space in denotation does not mean that *deictic_rel* would always resolve to *FormIdentity*. Let us illustrate this by reusing the example from Clark (1996) in §2.2: when pointing to the novel while uttering "*This man* was a friend of mine" the visible space that the hand points at and the space denoted by the deixis are equal since the novel is salient in the physical space gesture points at, i.e., v is equality. However, the denotation of the gesture is not identical to the one of speech, and we therefore claimed that the deixis denotation is related through an *AgentiveRelation* with the speech denotation.

Further construction rules account for the integration of noun-noun compounds and appositive constructions. For the sake of space, we forego any details about them. The underlying conditions of prosodic prominence of the temporally overlapped speech phrase remain unchanged.

4.3 Coverage and Non-Coverage. Issues and Future Work

The constraints presented above cover 83.33% of the multimodal actions encountered in our corpora.[6] Since our goal was not only to account for multimodal grammaticality but also to produce LFs supporting the final interpretations in context, we considered as uncovered those instances which although syntactically well-formed, did not correspond to the intended meaning representations.

In this section we shall report on the uncovered and the problematic instances. Firstly, the grammar rules do not account for integrating deixis to a non-nuclear item as the first stroke in (6). Secondly, any sequence of rule applications would result in massive overgenerations, e.g., in (8), the rule in Definition 2.1 would build a synchronous tree out of "said" and deixis when the preferred attachment to "she" is rendered only through the rule in Definition 2.4. Further issues are related to the fact whether we should allow for gesture projection from the internal arguments to the head accounting thus for the focus-projection principle. We still remain agnostic as to whether an attachment to, say, a verb phrase, should be barred in the right-most prosodic prominence architecture where the complement is by default associated with the nuclear prominence. These issues remain unresolved and are subject to future investigation.

5 Conclusions

In this paper, we demonstrated that well-established methods from linguistics are expressive enough to produce the form-meaning mapping of multimodal communicative actions. This goal was achieved by integrating speech and deictic gesture into a multimodal grammar, thereby using constraints from the form of the speech signal, the form of the gesture signal and their relative temporal performance so as to map them to a single meaning representation in the final logical form of the utterance. Using multimodal corpora, we extracted generalisations about multimodal well-formedness accounting for the intended meaning in context. The conditions that made a multimodal action well-formed were driven from the prosody of speech: we established that the prosodic prominence of the speech signal aligned with the deictic stroke. Exceptions to this generalisation were also possible but these could be constrained by using contextual information, i.e., the salience of individuals in the gestured space was indicative as to whether prosodic and/or temporal alignment was anticipated. We formally regimented our empirical findings into constraint-based construction rules that accounted for the speech-deixis attachments in the multimodal corpora. These rules were designed independently from a particular grammar framework but they are suitable for any constraint-based framework that interfaces structured phonology, syntax and semantics, e.g., Head-Driven Phrase Structure Grammar (HPSG), Lexical Functional Grammar or Combinatory Categorial Grammar. Despite that the current analysis was driven from one multi-party conversation and from one one-sided conversation, we believe that it has laid down the first steps towards a large-scale formal analysis of multimodal input. In future, we envisage to implement the grammar rules into the online implementation of HPSG and to test the rules against unseen multimodal data.

A significant contribution of this project was also the extension of the existing multimodal resources with annotated speech and gesture corpora which can be further used for various studies of multimodal communication.

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 Sasha Calhoun, Jean Carletta, Jonathan Kilgour, Ewan Klein and Mark Steedman.

[6] The tests were performed on 42 multimodal utterances out of 87 in the annotated data.

References

ALAHVERDZHIEVA K. & LASCARIDES A. (2010). Analysing speech and co-speech gesture in constraint-based grammars. In S. MÜLLER, Ed., *The Proceedings of the 17th International Conference on Head-Driven Phrase Structure Grammar*, p. 6–26, Stanford: CSLI Publications.

BOERSMA P. & WEENINK D. (2003). *'Praat:doing phonetics by computer'*. `http://www.praat.org`.

BRENIER J. & CALHOUN S. (2006). Switchboard prosody annotation scheme. Department of Linguistics, Stanford University and ICCS, University of Edinburgh. Internal publication.

CALHOUN S. (2006). *Information Structure and the Prosodic Structure of English: a Probabilistic Relationship*. University of Edinburgh. PhD Thesis.

CARLETTA J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, **22**, 249–254.

CLARK H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

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

COPESTAKE A. & 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*, p. 591 – 600, Athens, Greece.

GIORGOLO G. & VERSTRATEN F. (2008). Perception of speech-and-gesture integration. In *Proceedings of the International Conference on Auditory-Visual Speech Processing 2008*, p. 31–36.

GIULIANI M. & KNOLL A. (2007). Integrating multimodal cues using grammar based models. In *HCI (6)*, p. 858–867.

GIVÓN T. (1985). Iconicity, Isomorphism and Non-arbitrary Coding in Syntax. In J. HAIMAN, Ed., *Iconicity in Syntax*, p. 187–219. Amsterdam: John Benjamins.

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

JOHNSTON M. (1998). Multimodal language processing. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*.

KENDON A. (1972). Some relationships between body motion and speech. In A. SEIGMAN & B. POPE, Eds., *Studies in Dyadic Communication*, p. 177–216. Elmsford, New York: Pergamon Press.

KENDON A. (2004). *Gesture. Visible Action as Utterance*. Cambridge: Cambridge University Press.

KIPP M. (2001). Anvil — a generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg: Georgetown University.

KIPP M. (2008). Anvil 5.0: user's manual. `http://www.anvil-software.de`.

KLEIN E. (2000). Prosodic constituency in hpsg. In *Grammatical Interfaces in HPSG, Studies in Constraint-Based Lexicalism*, p. 169–200: CSLI Publications.

KOPP S., TEPPER P. & CASSELL J. (2004). Towards integrated microplanning of language and iconic gesture for multimodal output. In *ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces*, p. 97–104, New York, NY, USA: State College, PA, USA ACM.

KRANSTEDT A., LÜCKING A., PFEIFFER T., RIESER H. & WACHSMUTH I. (2006). Deixis: How to determine demonstrated objects using a pointing cone. In S. GIBET, N. COURTY & J.-F. KAMP, Eds., *Gesture in Human-Computer Interaction and Simulation*, volume 3881 of *Lecture Notes in Computer Science*, p. 300–311. Springer Berlin / Heidelberg.

KRUIJFF-KORBAYOVÁ I. & STEEDMAN M. (2003). Discourse and information structure. *Journal of Logic, Language and Information*, **12**, 249–259.

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

LASCARIDES A. & STONE M. (2009). A formal semantic analysis of gesture. *Journal of Semantics*.

LIBERMAN M. & 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.

MCNEILL D. (2005). *Gesture and Thought*. Chicago: University of Chicago Press.

OVIATT S. L., DEANGELI A. & KUHN K. (1997). Integration and synchronization of input modes during multimodal human-computer interaction. *CHI*, p. 415–422.

PUSTEJOVSKY J. (1995). *The Generative Lexicon*. MIT Press, Cambridge.