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 quels énoncés multimodaux sont bien formés, et lesquels ne pourraient jamais produire le sens voulu dans la situation communicative. 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 la performance temporelle de la parole et la deixis afin de contraindre la production d'un arbre de syntaxe combinant parole et gesture déictique ainsi que la représentation unifiée du sens pour l'action multimodale correspondant à cet arbre. 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 et parole.

Abstract. In this paper we present a constraint-based analysis of the form-meaning relation of deictic gesture and its synchronous speech signal. Based on an empirical study of multimodal corpora, we capture generalisations about which multimodal utterances are well-formed, and which would never produce the intended meaning in the communicative situation. 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 relative temporal performance of the speech and deixis to constrain the production of a single syntactic tree of speech and deictic gesture and its corresponding meaning representation for the multimodal action. In so doing, 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 gestural and speech 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, the pointing signals of their hands, etc. 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 show that co-speech deictic gesture can be integrated into a constraint-based grammar using purely linguistic information such as the prosody, syntax, semantics of speech, 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 multimodal signals to their (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 generalisations about well-formed multimodal actions and about multimodal actions that cannot 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 are expressive enough to represent the form-meaning mapping for deictic dimensions too.

2 Data

We start with an overview of deictic gesture and its relation to other co-speech gestures, and we then present the major challenges arising from the range of ambiguities and distinct performances of the pointing hand.

2.1 Deixis Background

Our focus of study are spontaneously performed co-speech deictic gestures. Compared to, say, depicting gestures where the hand literally or metaphorically *depicts* its denotation, deictic gestures designate spatial reference in Euclidean space marked by the projection of the pointing medium (finger, hand, arm, head, etc.) to a region that is proximal or distant in relation to the speaker's origo. Deictic gestures are thus anchored to the space and time of the communicative act, and so their propositional content is understood as a function that maps from a world in its contextually-specific time and space to truth values. The same is not necessarily valid for depicting gestures: uttering "What a big cake" while performing a circular motion with both hands in the frontal centre is not related to the spatial and temporal context in which the utterance occurs. We therefore argue that whereas depicting gestures provide qualitative characteristics of the referent, deictic gestures are at heart quantitative. This is the diametrical distinction that sets apart depicting and deictic gestures, and that prevails in how their semantics is defined.

Note that by "gesture" we mean the expressive part of the whole movement, the kinetic peak of the excursion that carries 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, and a non-obligatory *post-stroke hold* (the hands sustain their expressing 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 Deictic Use

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, let's consider the following example: when pointing in the direction of a book, does the space demarcated by the deictic gesture identify with the physical object book, the location of the book—e.g., the table—or with the cover of the book? Often there is not an exact

correspondence between the region identified by the pointing hand, the so called 'pointing cone' (Kranstedt *et al.*, 2006) and the reference. Our formal model does not intend to solve this ambiguity since it has no effects on multimodal perception, and certain ambiguities remain unresolved in context similarly to unimodal input. Based on Lascarides & Stone (2009), we formally regiment the location of the pointing hand with the constant $\vec{c}$, that marks the physical location of the tip of the index finger. This combines with the hand's shape, orientation and movement to determine the region $\vec{p}$ actually marked 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. We will also be using a function v to map the physical space $\vec{p}$ designated by the gesture to the actual space it denotes.

We further stated that there is a range of distinct relations between the speech signal and the pointing signal. 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 gesture space and the physical space (so v is identity), and the speech referent for "man" and the deictic referent are also bound by *identity*. In 2., the denotation of the deictic gesture and that of the synchronous speech are not identical since the individual pointed at is not present at the exact coordinates projected by the pointing fingers, and so the relation would be rather *virtual counterpart*. Finally in 3., the deictic gesture's denotation is again not equal to that in speech, and they are connected through *depiction*. Further ambiguity arises even 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* that is resolvable in context to several possible values, among them *identity, virtual counterpart, depiction*, and even *paraphrase*.

We have also observed that depending on how the hand is used in the pointing act, deictic gestures can designate regions of the visible space in two distinct ways: first, the form of the hand, including the location $\vec{c}$ of the tip of the index finger, identifies the region $\vec{p}$ in visible space that is designated by the gesture as exactly that region that is taken up by the hand itself. This use of deixis is common in living space descriptions and in direction giving dialogues; e.g., (1).[1]

(1) 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.

Second, the hand marks a distant region in the visible space to establish a real or virtual identity between the individual pointed at and the individual referred to in speech as in (2), or to perform a meta-narrative function such as offering up an instance of an object or acknowledging the addressee's statement. In this case, the form of the hand, including the physical coordinate $\vec{c}$, establish a region $\vec{p}$ in visible space that does *not* overlap with the hand.

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

§ 3.2 details how these two meanings are reflected in the formal semantic representation of deictic gesture.

It is generally assumed in the literature that deictic gesture combines with the temporally co-occurring speech signal without considering synchrony *outside* the temporal alignment (McNeill, 2005). For depicting gestures, we have shown elsewhere that synchrony is also possible beyond the strict temporal alignment of both signals (Alahverdzhieva & Lascarides, 2010). For deictic gestures, we have observed that the synchronous semantically related speech phrase can be a few milliseconds before or after the deictic stroke. In (3), 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.

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

[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).

Upon our empirical study of the temporally misaligned occurrences, we learnt that the temporal relaxation is applicable in cases where the visible space $\vec{p}$ that is designated by the gesture is identical to the space $v(\vec{p})$ that it denotes (in other words, v is identity). 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 logical form in the specific context. We shall therefore equip our grammar with rules that apply only when there is an identity function mapping the visible space to space in denotation. This will support an analysis of (3) where the deixis does *not* denote the same individual as the pronoun "I". An alternative interpretation would be where the gesture is synchronous with the temporally co-occurring speech "Thank you" in which case the gesture complements on the speech by introducing a causal relationship of the sort "Thank you for handing me the mouse".

Having introduced the main challenges that we are dealing with, we now turn to the problem of how deixis and speech interact at the level of linguistic form (prosody) and meaning.

3 Speech-Deixis Interaction

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 (McNeill, 2005). We thus analyse deixis in *synchrony* with speech, as a mapping from form to some (underspecified) meaning in the final logical form of the utterance. 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 linguistic phrase a gesture stroke is synchronous with is guided by: i. the final interpretation of the gesture in specific context-of-use; ii. the speech phrase whose content is semantically related to that of the gesture given the value of (i); and iii. the syntactic structure that, with standard semantic composition rules, would yield an underspecified logical formula supporting (ii) and hence also (i).*

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 co-occurrence between the two modalities: 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., 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

In this project, we adopt the Autosegmental-Metrical (AM) theory (term coined by Ladd (1996)) for the analysis of speech prosody. This theory views prosodic prominence as a relational property between two juxtaposed units where the prominence of unit A is determined by its (strong or weak) relation to unit B.

Based on the findings of a previous prosody study (Calhoun, 2006), we argue that it is not the lower or higher tune but rather the nuclear accenting that constrains the alignment between gesture and speech. We view nuclear accenting as the perception of phrase-level prominence which is relative to the metrical structure, and not to the acoustic properties of the syllables. In the AM model, 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—that is, the nuclear accent is associated with the right-most word. For instance, Figure 1[2] illustrates the metrical tree for "fasten a cloak" in its broad

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

focused reading with the nuclear accent being on the word entirely dominated by s nodes—"cloak". Liberman & Prince (1977) call the most prominent element of a given constituent *Designated Terminal Element* (DTE).

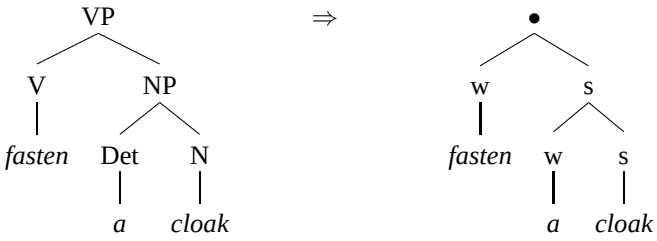

Figure 1: Syntactic Tree and Metrical tree

Strong nodes on the left of the nuclear accent can also appear, and these are known as pre-nuclear accents. Unlike the nuclear accents, pre-nuclear accents are signalled by their acoustic properties rather than their relative position in the metrical tree.

3.1.1 Empirical Study

We used empirical data to determine constraints on the interaction between deictic gestures and speech signals.

Hypothesis 1 *Deictic gestures align with the nuclear pitch 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 separate stages: annotation of speech which included word transcription, pitch accents pointing to words and prosodic phrases; and gesture annotation which included marking of gesture phrases, gesture phases, and also formless moves that beat along the speech rhythm known as beats. Both annotations were performed independently from each other.

Prosody Annotation As an annotation tool, we used Praat (Boersma & Weenink, 2003). Our annotation schema is largely based on the guidelines of the prosody annotation of the Switchboard corpus (Brenier & Calhoun, 2006). We marked 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 whole 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 (they correspond to 'plain' or 'regular' accents in Brenier & Calhoun (2006) and Calhoun (2006)); *none*: a non-discernible accent in a phrase (it corresponds to a 'Z' accent in Brenier & Calhoun (2006)); *?*: 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.

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

Gesture Annotation We used the Anvil labelling tool (Kipp, 2001) to annotate the gesture phrases, gesture phases and beats. Along the lines of Loehr (2004), we annotated gestures for the dominant *H1* hand, and for the non-dominant *H2* hand. Bi-handed gestures where the movement of *H1* was symmetrical to *H2* were coded in *H1*.

1. *Hand Movement.* The annotation of the hand movement proceeded in two main passes. The first pass aimed at marking the temporal boundaries of all hand movements, and performing a binary classification on them in terms of *communicative–non-communicative* signals. The second pass determined what dimensions the communicative signals belong to, they being *literally depicting*, *metaphorically depicting* or *deictic*. To stay consistent with the findings in the literature that a single gesture can have dimensions of, say, depicting and deictic gestures (McNeill, 2005), our annotation schema permitted for marking gestures belonging to more than one dimension.
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.
3. *Beat.* Beat movements were marked in a separate layer so as to study whether they always superimpose other gestural dimensions, or pure beats also occur.

Past annotation tasks of the Switchboard corpus (Calhoun, 2006) and of the multimodal corpus of Loehr (2004) have shown that the annotation of accents and boundaries is reliable (see Table 1), and also the annotation of gesture dimensions (see Table 2).

	All Types	**+/-**
Accents	0.800	0.800
Boundaries	0.889	0.910
Words	(752)	

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

	Coding	**Segmentation**
Hand movement	0.8536/0.8994	0.8502/0.8659
Deictic gesture	0.8605/0.8994	0.8502/0.8659
Literally depicting	0.8663/0.8916	0.8502/0.8659
Metaphorically depicting	0.8221/0.8623	0.8502/0.8659
Gesture phase	0.662/0.7	0.8864/0.8971
Beat	0.6599/0.8203	

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

Multimodal Corpora in NXT The annotated corpora were converted into Nite XML Technology (NXT) format (Carletta *et al.*, 2005), (Calhoun *et al.*, 2010) which allows for querying a corpus as a coherent set and extracting information from it by exploring the relations between the annotation layers. A corpus in NXT consists of 'observations'—our two video recordings—and annotations associated with it—orthographic transcriptions, pitch accents, prosodic phrases, gesture phrases, gesture phases and beats. Each data object is necessarily equipped with timestamps which are synchronised with the video and/or audio signal.

Data objects can be bound either by structural or by temporal relations which is specified in a meta-data file containing the annotation schema of the corpus. The type of relation also determines the query that can be executed onto these objects. The annotation of each data object is stored in a separate XML file; any relations between the annotation objects are defined in terms of stand-off links between the elements. Figure 2 illustrates the relation between the 'accents' and 'words' layers: the accent's attribute nite:pointer serves as a pointer to the unique nite:id of the relevant word. In this way, we can elegantly capture accents not overlapping a word, accents associated with two words, and also words associated with two accents.

We further specified the relationships between gestures and gesture phases, and between prosodic phrases and words as parent-child relations. This choice of representation is consistent with the essence of prosodic phrases and gesture phrases: prosodic phrases are made up by a certain number of words, and so the beginning of the first word aligns with the beginning of the prosodic phrase, and the end of the last word aligns with the end of the prosodic phrase. The same mechanism applies to gestures which are made up by at least one gesture phase. We

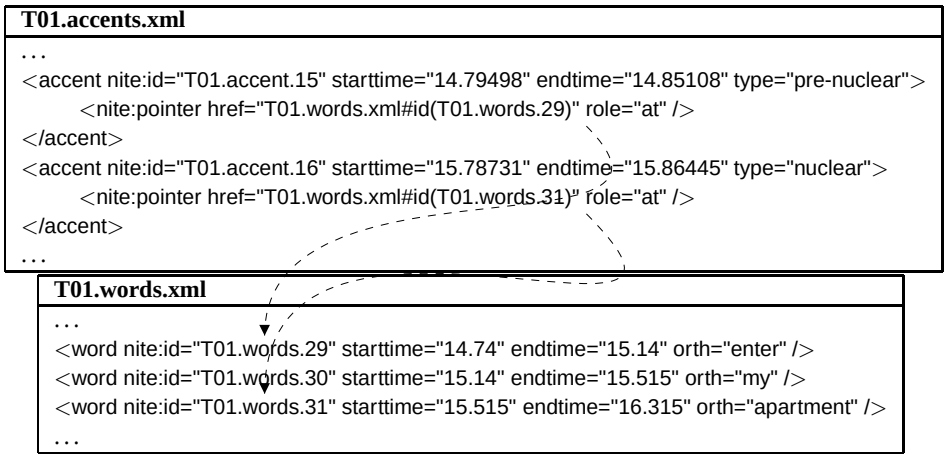

Figure 2: NXT Coding of Accents Associated with Words

forego any details about the specification of beats since they are not represented in a structural relationship with other layers.

3.1.2 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 utterances. The following two utterances illustrate our findings: (4) is a broad-focused utterance with the nuclear accent being on the right-most word. Utterance (5), which is 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: 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 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.

(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

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

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

Further to this, we looked at all-new information utterances with an initial strong acoustic pitch and then a nuclear accent on the right-most element. In these utterances, the stroke was performed along with the initial pre-nuclear accent, and there might have been a post-stroke hold on the other components of the utterance. This is exemplified in (7) where an initial meaningful speech segment aligns with the stroke, and then the content is elaborated while holding the hands in an expressive position.

(7) 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.

We use the results of this statistical analysis to define constraints on the temporal overlap between deictic gesture and speech. Also, we need to explore whether any semantic relation can be established between the temporally aligned signals.

3.2 Mapping Deixis Form to Deixis Meaning

Following previous research (Johnston, 1998), (Kopp *et al.*, 2004), the form of the pointing hand is represented using typed feature structures, where each feature value pair corresponds to an aspect of form. We use fine-grained an analysis as possible: we consider that the shape of the hand, the orientation of the palm and the fingers, the hand movement, and also the location of the hand in the spatio-temporal coordinates $\vec{c}$ are the distinct classes of form that potentially have semantic effects, e.g., the shape of the hand influences the mapping from $\vec{c}$ to $\vec{p}$. Moreover, this form representation captures the fact that the different attributes composing deictic gesture's form are not hierarchically ordered, but are rather a flat list. Figure 3 gives the form representation of the gesture in (1)—the value $\vec{c}$, which identifies the spatio-temporal coordinates of the hand, together with the other values, serve to identify the region $\vec{p}$ designated by the gesture's content (Lascarides & Stone, 2009); as explained in § 2.2, a pointing gesture (with hand shape 1-index) will make $\vec{p}$ denote a cone or line that starts at *hand-location* $\vec{c}$ and whose direction is the same value as *finger-direction* (Kranstedt *et al.*, 2006). Note also that the gesture is typed as *communicative_gesture_deictic* to distinguish between form features contributed by depicting gestures, and those contributed by pointing gestures.

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

Figure 3: TFS Representation of Form of Deictic Gesture

As a semantics description language we use Robust Minimal Recursion Semantics (RMRS) 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 — e.g., it does not determine 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. Defining flat semantics in RMRS involves defining a set of *Elementary Predications (*EPs). Each EP is associated with a label l_i that ultimately identifies the scopal position of the predicate in the context-resolved logical form. Shared labels are also possible, and they mark implicit conjunction as in intersective modifiers. Each EP is also associated with a unique anchor a_i, which serves as its locus for specifying arguments to the EP — e.g., $ARG2(a, x)$ means that the second argument to the EP whose anchor is a is the individual x. The absence of such ARG relations in the RMRS thus serves to underspecify the arguments to predicates and even the predicate's arity. Holes (h_i) are used to represent scopal arguments whose value is not fully determined by syntax. The admissible pluggings are constrained by equality conditions ($=_q$) between holes and labels ($h_i =_q l_i$ means that only 0 or more quantifiers intervene between the scopal positions). Finally, a top label h_0 is added for the whole formula.

§ 2.2 detailed the two distinct functions of deictic gestures. We will now present their compositional semantics as follows:

1. **Hand as reference.** The speaker here points to an individual/event represented by the hand which is located at the spatial coordinate $\vec{p}$ designated by the finger tips often, but not necessarily, in relation to another individual available from the discourse. The form features of the pointing hand further constrain the set of possible relations between gesture and speech, e.g., an open hand supine used for turn coordination can resolve to a *metatalk* relation (Lascarides & Stone, 2009) — roughly put, the gesture can have a meaning that can be paraphrased by the parenthetical phrase "I am telling you".
   The RMRS representation of the gesture in (1) is shown in Figure 4. Following Lascarides & Stone (2009), this RMRS semantics says that the pointing hand provides the spatial reference of an underspecified referent i (an individual or an event) at some position in the physical space $v(\vec{p})$. In context, the underspecified variable i may resolve to an individual x as in (1), or to an event e as in (7). To stay consistent with the findings in the descriptive literature, namely that the shape of the pointing hand is associated with a specific meaning (Kendon, 2004), we map each form feature-value pair to a two-place predicate. Their formal treatment is similar to the treatment of intersective modifiers in the English Resource Grammar (ERG) in that they share labels with the main predicate sp_ref. Again for consistency with ERG where individuals are bound by quantifiers, there is a quantifier outscoping the referent introduced by the deictic gesture. Following Lascarides & Stone (2009), we use the $\mathcal{G}$ operator so as to guarantee that individuals referred in speech cannot be co-referred to individuals introduced in gesture. To obtain this, $\mathcal{G}$ must outscope all gesture predications (formalised in terms of $=_q$ equality conditions).
2. **Reference is the region marked by the hand.** The hand here also points to an underspecified reference i located at $v(\vec{p})$ but unlike the previous function, the hand shape denotes not the reference itself but the region marked by it. The semantics of the gesture in (2) is shown in Figure 5, and it is similar to the one displayed in Figure 4 with the only difference being that it is the region that is modified by the various gesture form-features. Since the rest of the predications remain the same, we forego any details about them.

$l_0 : a_0 : [\mathcal{G}](h_1)$
$l_1 : a_1 : deictic\_q(i)\ RSTR(a_1, h_2)\ BODY(a_1, h_3)$
$l_2 : a_2 : sp\_ref(i)\ ARG1(a_2, v(\vec{p}))$
$l_2 : a_3 : hand\_shape\_open\_flat(e_0)\ ARG1(a_3, i)$
$l_2 : a_4 : palm\_orient\_vertical(e_1)\ ARG1(a_4, i)$
$l_2 : a_5 : finger\_orient\_forward(e_3)\ ARG1(a_5, i)$
$l_2 : a_6 : hand\_move\_away\_centre\_left(e_5)\ ARG1(a_6, i)$
$h_1 =_q l_1;\ h_1 =_q l_2;\ h_2 =_q l_2$

Figure 4: RMRS for Hand as Reference

$l_0 : a_0 : [\mathcal{G}](h_1)$
$l_1 : a_1 : deictic\_q(i)\ RSTR(a_1, h_2)\ BODY(a_1, h_3)$
$l_2 : a_2 : sp\_ref(i)\ ARG1(a_2, v(\vec{p}))$
$l_2 : a_3 : RH\_palm\_orient\_vertical(e_1)\ ARG1(a_3, \vec{p})$
$l_2 : a_4 : RH\_finger\_orient\_forward(e_2)\ ARG1(a_4, \vec{p})$
$l_2 : a_5 : RH\_hand\_move\_away\_body\_left(e_3)\ ARG1(a_5, \vec{p})$
$h_1 =_q l_1;\ h_1 =_q l_2;\ h_2 =_q l_2$

Figure 5: RMRS for the Region Marked by the Hand

4 Rules for Combining Deixis and Speech in the Grammar

We intend to augment the existing wide-coverage grammar for English—the English Resource Grammar ERG—with construction rules for combining speech and gesture. This task involves specifying the prosodic component in the grammar (we shall be using the AM theory), and also interfacing it with the syntax-semantics component.

We formally regiment our findings about the deixis-prosody interaction (§ 3.1) into the following basic construction rules:

Definition 2.1 *Deictic gesture attaches to the temporally overlapping nuclear/pre-nuclear head word.*

Definition 2.2 *Deictic gesture attaches to the temporally overlapping nuclear/pre-nuclear head word after it had been combined with the arguments and/or modifiers to the head.*

The motivation to include the latter stems from the fact that semantically the deictic signal is not strictly constrained to its temporally co-occurring word but rather it can be linked to a larger phrase. For instance, in (7), there is no information coming from the form of the hand, nor from its relative timing that it should be attached to "enter" only, and not to "enter my apartment", in which case the form of the hand would be related to the rectangular shape of, say, an entrance door to an apartment. Intuitively in this case, the gesture directs not only to the point of entering the house, but also to the entrance door which by the hand shape is rectangular.

The syntactic structure is derived in parallel with the prosodic one, and so the syntactic component would consist either of a single head word without further constraints on its syntactic category, or a larger phrase it being a head-argument, a head-modifier phrase, or an entire utterance. Generally speaking, a deictic gesture cannot be

combined with a non-prosodic word. We will come back to this point a bit later, when we will see that some exceptions to this rule can also arise. Finally, the semantic component uses the RMRS representation in § 3.2.

Semantic composition with RMRS (Copestake *et al.*, 2001), (Copestake *et al.*, 2005) is monotonic, ensuring that the semantic representations of the daughters are always subsumed by that of the mother. For each phrase, one specifies semantic entities (*sement*s) of the following parts: (1). *Top*: the global label containing the whole formula. During composition, the top labels of the daughters are equated with the top label of the mother to demonstrate the derivation of a single logical form; (2). *Hook*: placeholder that records the semantic value of the formula. It contains (a). *local top*: the label containing an EP. For instance, in Figure 4 the local top of $l_2 : a_2 : sp\_ref(i)\ ARG1(a_2, v(\vec{p}))$ is l_2; (b). *semantic index ($i_1, i_2 \ldots i_n$)*: it indicates what the phrase is about and has two subtypes: events ($e_1, e_2 \ldots e_n$) and individuals ($x_1, x_2 \ldots x_n$). It is obtained by co-indexation with the topmost EP. For instance, in Figure 4, the semantic index of the phrase is i obtained by co-indexing it with the main variable of $sp\_ref(i)$; (3). *Slots*: resources that need to be consumed so that a functor becomes semantically saturated; (4). *Rels*: a bag of EPs; (5). *Equality constraints ($=_q$)*: scopal constraints indicating the admissible plugging of a subformula into a hole.

To summarise, an RMRS sement is: $\langle Top\,[ltop, i]\,\{slots\}\{rels\}\,[=_q]\rangle$. Semantic composition of a $sement_M = op(sement_{D1}, sement_{D2})$ involves the following operations: 1. making Top of $sement_M$ = $sement_{D1}$ = $sement_{D2}$; 2. making the hook of $sement_M$ the hook of $sement_{D1}$; 3. making the remaining slots of $sement_{D1}$ and $sement_{D2}$ the slots of $sement_M$; 4. making the $rels$ and $hcons$ of $sement_M$ the union of those of the daughters.

As argued in § 2.2, deictic gesture always relates with the synchronous speech signal through some sort of relation; e.g., *identity, virtual counterpart* or a *paraphrase* relation. Based on Lascarides & Stone (2009), the construction rule therefore introduces an underspecified relation $deictic\_rel(i_1, i_2)$ between the semantic index i_1 of the deictic gesture and the semantic index i_2 of speech. How this relation resolves is a matter of discourse context. Similarly to the treatment of intersective modification in language, *deictic_rel* 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. This is similar to the treatment of appositives in ERG.

With this machinery at hand, let us now turn to the derivation of utterance (1): the deictic gesture overlaps the nuclear-accented prosodic word "hallway", and hence we could build a single situated noun out of "hallway" and deixis as demonstrated in Figure 6. In semantics, we need to extend the RMRS representation in Figure 4 with a top label, hook and slots as follows (for the sake of readability, we gloss the semantic predications contributed by the deictic form features as $l_2 : a_3 : deictic\_eps(e_0)\ ARG1(a_3, i)$):

$< h_0, [l_0, a_0, i]\,, \{\}$
$\{l_0 : a_0 : [\mathcal{G}](h_1)$
$l_1 : a_1 : deictic\_q(i)\ RSTR(a_1, h_2)\ BODY(a_1, h_3)$
$l_2 : a_2 : sp\_ref(i)\ ARG1(a_2, v(\vec{p}))$
$l_2 : a_3 : deictic\_eps(e_0)\ ARG1(a_3, i)\}$
$[h_1 =_q l_1;\ \ h_1 =_q l_1;\ \ h_2 =_q l_2] >$

The top label of the whole formula is h_0 and in derivation it is made identical with that of the mother. Further, the local top is identified with the label of the $\mathcal{G}$ operator which contains all other predications. The semantic index of the deictic gesture is the underspecified variable i introduced by the sp_ref predicate that in composition resolves to x_1. Finally, we assign no slots to the formula.

As argued above, the form of the deictic signal is not sufficient to decide whether the hand refers to "hallway", to "little hallway" or even to "a little hallway", and so our grammar does not impose constraints on the syntactic phrase, and is thus able to generate all these combinations. Prosodically, we integrate deixis into a metrical tree where the prosodically prominent element, the DTE, 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 NP "a little hallway" and deixis, both the head noun and the deictic relation would appear within the restrictor of the quantifier. The semantic composition remains the same as above.

We shall now look at some exceptions that are not covered by the temporal alignment constraint and the nuclear prominence constraint. We saw that in (3) there exists an obvious misalignment between the semantically related speech and deixis signals. Similarly in (8), the deixis denotation is identical to the denotation of "she" despite it not being prosodically prominent.

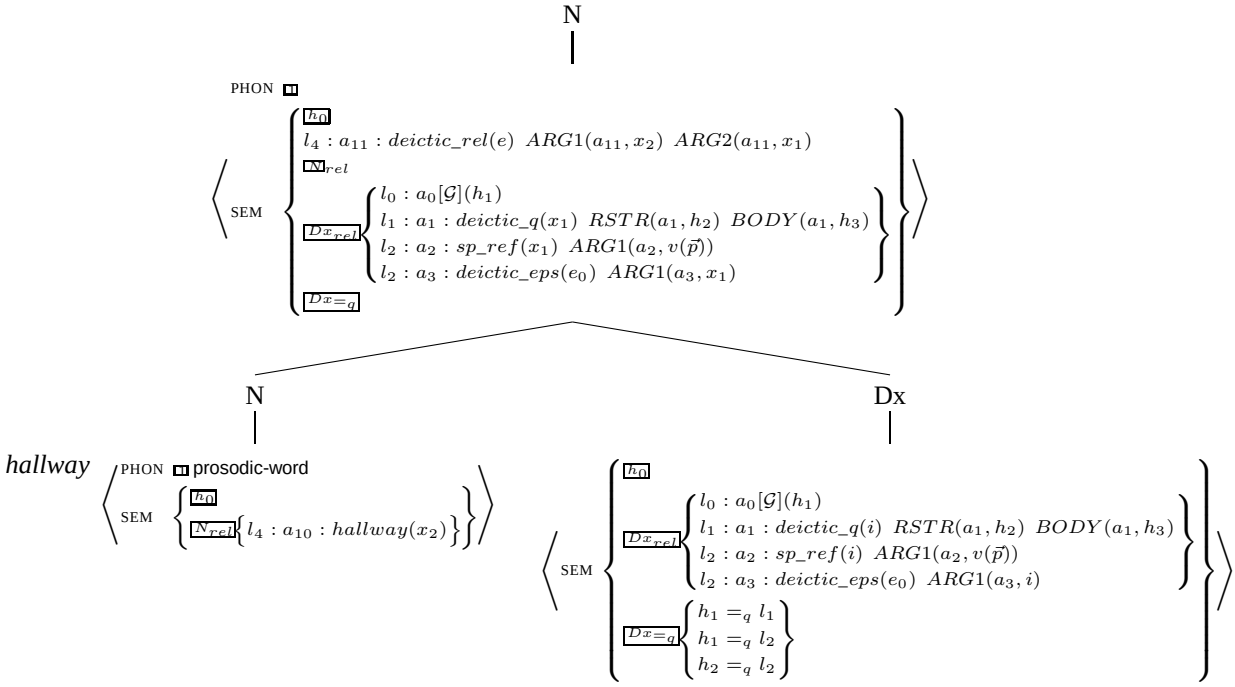

Figure 6: Derivation Tree for Deictic Gesture and the N "hallway"

(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

To cope with these exceptions, we studied all utterances where the semantically preferred attachment of the deictic gesture is an element beyond its temporal performance and/or a non-prosodically prominent element. This temporal/prosodic 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, that is, there is an identity between the physical space that the hand points at and the actual denotation of the gesture's referent. Of course, this does not mean that *deictic_rel* would always in this case resolve to identity. 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" there is an identity between the visible space that the hand points at and the denotation of the gesture since the novel is salient in the physical space gesture points at. However, the denotation of the gesture is not identical to the one of speech, and we therefore claimed that the relation between speech and deixis is rather *depiction*. In our grammar, we therefore spell out the following rule:

Definition 2.3 *Deictic gesture attaches to an item (prosodically prominent or non-prominent) whose temporal performance is adjacent to that of the gesture if the mapping v resolves to identity.*

Importantly, this relaxation is not applicable in cases where the hand serves as an abstract reference pointing to an object not present in the communicative act. If the gesture in (4) was related to the speech head daughter "I", the logical form would fail to resolve.

5 Conclusions

In this paper, we demonstrated that well-established methods from linguistics are sufficient to provide the form-meaning mapping of multimodal communicative actions consisting of speech and deictic gesture. This goal was achieved by integrating them into a multimodal grammar thereby using constraints coming 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. This paper contributed to the existing resources by setting the theoretical framework for a multimodal grammar, and also by extending the existing corpora with prosody and gesture annotation in the NXT format which can further be used for various studies of multimodal communication.

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.

CALHOUN S., CARLETTA J., BRENIER J., MAYO N., JURAFSKY D., STEEDMAN M. & BEAVER D. (2010). The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, **44**, 387–419.

CARLETTA J., EVERT S., HEID U. & KILGOUR J. (2005). The nite xml toolkit: Data model and query language. *Language Resources and Evaluation*, **39**, 313–334.

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

COPESTAKE A., FLICKINGER D., SAG I. & POLLARD C. (2005). Minimal recursion semantics: An introduction. *Journal of Research on Language and Computation*, **3**(2–3), 281–332.

COPESTAKE A., LASCARIDES A. & FLICKINGER D. (2001). An algebra for semantic construction in constraint-based grammars. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL/EACL 2001)*, p. 132–139, Toulouse.

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. (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.

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.