Semantic Composition of Multimodal Actions in Constraint-based Grammars

Katya Alahverdzhieva and Alex Lascarides

School of Informatics, University of Edinburgh

Abstract. The use of speech-accompanying hand gestures to depict objects, to structure the discourse or to give directions is ubiquitous in face-to-face interaction. In this paper, we analyse multimodal signals consisting of speech and gesture from the perspective of semantic composition in constraint-based grammars: we elevate standard methods from linguistics to description of multimodal input so as to connect the semantics of the gestural signal to the semantics of the speech signal and to produce an integrated logical form.

1 Introduction

The past few decades have witnessed substantial research in spontaneous, improvised co-speech gestures performed in synchrony with speech, e.g., [5], [9]. The vast majority of the descriptive, cognitive and formal studies of gesture unanimously acknowledge the fact that speech and gesture function within a single communicative system to convey an integrated meaning through spoken and visual material.

In this paper, we take the integrated nature of the speech-gesture action as a starting point, and we demonstrate that well-established mechanisms for semantic composition from linguistics can be applied to multimodal actions consisting of speech and communicative co-speech hand gestures. In particular, we use the form of the gesture signal, the form of the speech signal, and their relative timing to define constraints within the HPSG framework [10] on which parts of the speech are semantically related to the gesture. We further use the semantic framework of Robust Minimal Recursion Semantics (RMRS) [3] to map the form of the gesture signal to an underspecified meaning representation, providing an abstract representation of what the signal means in any context. Following earlier definitions we will say that a gesture g is semantically synchronous with a speech phrase s if their contents are connected by a meaningful relation that serves to establish the coherence of the speech-gesture ensemble. We intend to capture semantic synchrony by composing an intergrated multimodal semantics: a construction rule that indicates that the speech-gesture ensemble is well-formed introduces an underspecified semantic relation between the meaning of the gesture signal g and the meaning of the synchronous speech signal s.

The formal modelling of gesture lies on the interface between form (prosody and syntax), semantics and discourse. The grammaticality of the multimodal utterance is licensed through linguistically informed construction rules that integrate speech and gesture in a single syntactic tree. We use this tree to compose the (underspecified) logical form (ULF) of the utterance. Finally, through reasoning with this ULF and contextual information, we establish a semantic relation between the speech and gesture which is similar to relating two clauses in discourse.

The main challenge for modelling gesture arises from its *ambiguity*, both syntactic and semantic. We use *attachment ambiguity* to reflect the fact that the choice of which speech phrase a gesture is semantically synchronous with is not unique, with each choice having potentially distinct effects on the gestural interpretation in context. In (1),[1] for instance, does the gesture attach to "books", in which case the hands' denotation is a container containing the books? Or does it attach to "give you other books" so that the content of the forward movement of the open hands is the metaphor of giving and offering? Alternatively, the gesture could also attach to the whole clause in which case, the agent, the recipient and the books offered all serve to resolve the values of the participants

[1] http://www.talkbank.org/media/ClassBank/Lecture-unlinked/feb07/feb07-1.mov The speech signal aligned with the expressive part of the gesture, the *stroke*, is underlined. 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).

expressed by the gesture. Finally, the gesture might be a paraphrase of "I am informing you" which is possible only by attachment to the entire clause.

(1) I can give you other books ....
Hands are in parallel with palms open vertical. They move quickly forward to the frontal centre.

The semantic ambiguity of co-speech gestures refers to the fact that within various contexts and also within the same context one gesture form can map to distinct semantic representations, which can resolve to distinct values on the semantics/pragmatics interface. For instance, in the context of (1) the forward hand movement can resolve to the 4-place predicate $literal\_giving(e, x, y, z)$; also the vertical hand shape can resolve to the one-place predicate $books(x)$. It is perfectly acceptable for the hand movement from (1) to be performed in the context of "The church is in front of the bus station", in which case the same hand shape would contribute the 2-place predicate $in\_front\_of(x, y)$ between the church x and the station y.

Despite the ambiguities in form, we argue that speech-gesture synchrony is constrained by the prosodic properties of the speech signal; e.g., we consider the constructed utterance in (2) ill-formed even though the gesture is semantically related to the act of calling. We view ill-formedness as the temporal performance of one mode relative to the other at a place where it could never happen. In this instance, the grammaticality of the multimodal utterance depends on the gesture temporally overlapping the nuclear-accented item or a larger phrase containing it.

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

Following [7], we assume that resolving gesture's meaning to a specific value is logically co-dependent with inferring a rhetorical relation between gesture and its synchronous speech. The set of relations are constrained by the use that the hand makes out of the referent: for depicting gestures which literally or metaphorically depict the referent, we expect one set of relations distinct from those of deictic gestures which provide the referent's spatial characteristics; e.g., interpreting the gesture in (1) as denoting the books suggests an inference where the hand signal and the speech signal are related through *Depiction*. The alternative interpretation where the gesture is rather a paraphrase of "I am informing you" supports a *Metatalk* relation. In comparison, the deixis in (3) is related to the synchronous speech through *Identity* since there is one-to-one correspondence between the gestural denotation and the spoken denotation. If however, the hand pointed in the virtual space while uttering "This guy comes from tropical countries" without the referent being at the spatial coordinates identified by the pointing, the relation would be rather *VirtualCounterpart*.

(3) [$_{PN}$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.

In our model, we capture the various relations by introducing in semantics an underspecified relation $vis\_rel(s, g)$ between the content g of the depicting gesture and the content s of speech, and also an underspecified relation $deictic\_rel(s, g)$ between the content g of the deictic gesture and the content s of speech. The possible resolutions are established outwith the grammar as they are informed by discourse context.

2 Gesture Form

It is now well-established in the formal models of gesture to represent its form by Typed Feature Structures (TFSs)—e.g., [6]—which capture the fact that gesture, unlike language, is not hierarchically structured [9] but it rather contains a list of features such as hand shape, the orientation of the palms and fingers, location and movement.

To illustrate, consider the TFS in Fig.1 of the gesture in (1). The TFS is typed as *depicting* so as to differentiate between feature values contributed by depicting gestures and those of deictic gestures. This distinction is essential as it allows us to construct the appropriate LFs for gesture: whereas depicting gestures require qualitative characteristics represented by (underspecified) predications,

deictic gestures need quantitative values in that they denote the spatial coordinates of the referent (see §3). In comparison, the TFS of deixis (Fig.2 shows the TFS of the deixis in (3)) contains the coordinate $\overrightarrow{c}$ which marks the exact location of the tip of the index finger and which, together with the deixis form features, constrains the region $\overrightarrow{p}$ actually designated by the gesture [7]. We record the form features of the pointing hand since they effect the designated areas; e.g., an extended index finger marks a line (or a cone) that starts from the tip of the index finger and continues in the direction of the finger orientation; an open flat hand, on the other hand, can designate a region $\overrightarrow{p}$ that starts from the palm and continues in the perpendicular direction to the palm orientation.

depicting	
HAND-SHAPE	open-vertical
RH-PALM-ORIENTATION	left
LH-PALM-ORIENTATION	right
FINGER-ORIENTATION	forward
HAND-LOCATION	centre-low
HAND-MOVEMENT	straight-down

Fig. 1: TFS of Depicting Gesture

deictic	
RH-HAND-SHAPE	open-flat
RH-PALM-ORIENTATION	upwards
RH-FINGER-ORIENTATION	forward
RH-HAND-MOVEMENT	straight-right
RH-HAND-LOCATION	$\overrightarrow{c}$

Fig. 2: TFS of Deictic Gesture

3 Semantic Underspecification

In §1 we stated that the semantic ambiguity of gesture is persistent even within the same context of use. A standard approach for handling cases where the disambiguated representation about form is insufficient to determine a complete interpretation is *semantic underspecification*. We adopt the framework of RMRS due to the following factors: first, in RMRS, the argument sort can be left underspecified which is useful since the hand signal is ambiguous with respect to the main property depicted through gesture — a gesture in the same context can denote an event e or an individual x. We also need underspecification to produce an underspecified semantic relation between the speech content and the gesture content. Finally, in RMRS the predicate's arity can be left underspecified: recall from §1 that the same hand movement can resolve to predicates of distinct arity.

For depicting gestures, producing LFs in RMRS involves mapping each feature value pair to an *elementary predication* (EP) with underspecified scope, arity and variable; e.g., the form features in Fig.1 map to the RMRS representation in Fig. 3. An EP is associated with a (not necessarily unique) label ($l_0 \dots l_n$) which marks the scopal position of the constructor that the EP resolves to in the final LF, a unique anchor ($a_0 \dots a_n$) which is used as a locus for adding arguments to the EP's predicate symbol, and an underspecified variable ($i_1 \dots i_n$), which abstracts over the argument sort of the predication. Following [7], we also use the $\mathcal{G}$ operator to limit the scope of the gesture within its modality. This is for the sake of co-reference in discourse, so that individuals introduced by a depicting gesture are not co-referred to by subsequent pronouns in speech. This scopal constraint is formalised through the scopal conditions $=_q$ between the argument h_1 of the operator $\mathcal{G}$ and the labels of the predications $l_1 \dots l_6$. The resolution of the underspecified predicates happens outside the grammar by a hierarchy of increasingly specific predicates that all relate to the original EP through iconicity, e.g., $hand\_shape\_open\_vertical(i_1)$ can resolve to the more specific content $holding\_event(e)$ which in turn can resolve to $literal\_giving(e, x, y, z)$ [7].

Mapping deixis form to meaning (the RMRS of the gesture in Fig.2 is shown in Fig.4) follows the principles introduced above: we map each feature value pair to an EP associated with labels, anchors and arguments. Since the role of the deictic predications is to constrain the region denoted by the pointing hand, we treat them as intersective modifiers in the English Resource Grammar (ERG) [1], i.e., 2-place predicates with the second argument being the area denoted by the gesture. Deictic gestures provide the spatial reference of an individual or event i in the physical space $\overrightarrow{p}$, and we therefore augment the deixis' compositional semantics with the 2-place EP $l_1 : a_1 : sp\_ref(i)\ ARG1(a_1, v(\overrightarrow{p}))$ where i is an underspecified referent introduced by the gesture and v is a function that maps the gestured space $\overrightarrow{p}$ to the actual space $v(\overrightarrow{p})$ in denotation. This function is important, since there is not always an exact correspondence between the space the gesture points at and the space it actually denotes; e.g., in the navigation domain the frontal space is often used as a virtual map for setting up and navigating through landmarks that have no physically accessible counterparts. For consistency with ERG where individuals are bound by quantifiers, we introduce the quantifier *deictic_q* that takes scope over the spatial referent. This is

$l_0 : a_0 : [\mathcal{G}](h_1)$
$l_1 : a_1 : hand\_shape\_open\_vertical(i_1)$
$l_2 : a_2 : RH\_palm\_orient\_left(i_2)$
$l_3 : a_3 : LH\_palm\_orient\_right(i_3)$
$l_4 : a_4 : finger\_orient\_forward(i_4)$
$l_5 : a_5 : hand\_location\_centre\_low(i_5)$
$l_6 : a_6 : hand\_move\_straight\_down(i_6)$
$h_1 =_q l_i \ \ where \ \ 1 \leq i \leq 6$

Fig. 3: RMRS of Depicting Gesture

$l_0 : a_0 : deictic\_q(i) \ RSTR(a_0, h_1) \ BODY(a_0, h_2)$
$l_1 : a_1 : sp\_ref(i) \ ARG1(a_1, v(\overrightarrow{p}))$
$l_1 : a_2 : RH\_hand\_shape\_open\_flat(e_1) \ ARG1(a_2, v(\overrightarrow{p}))$
$l_1 : a_3 : RH\_palm\_orient\_upwards(e_2) \ ARG1(a_3, v(\overrightarrow{p}))$
$l_1 : a_4 : RH\_finger\_orient\_forward(e_3) \ ARG1(a_4, v(\overrightarrow{p}))$
$l_1 : a_5 : RH\_hand\_move\_straight\_right(e_4) \ ARG1(a_5, v(\overrightarrow{p}))$
$h_1 =_q l_1$

Fig. 4: RMRS of Deictic Gesture

ensured by adding $=_q$ constraints between the RSTR (restrictor) of the quantifier and the label of sp_ref. Unlike the compositional semantics for depicting gestures, the EPs are not outscoped by the gestural modality. In this way, we could capture co-reference in multimodal discourse where an individual introduced by a pointing gesture can be anaphorically referred to with "it".

4 Semantic Composition

We start off by identifying constraints on the speech-gesture interaction so as to produce a single multimodal tree which guides the process of semantic composition. Based on previous studies about the interaction between speech prosody and gestural performance [5], [8], [4], we used multimodal corpora annotated for speech (orthographic transcription, pitch accents and prosodic phrases) and gesture (gesture boundaries, segmentation of gesture into distinct phases and gestural dimensions such as depicting and deictic) to extract generalisations about multimodal well-formedness. For depicting gestures, we found that 96% of the gesture strokes overlapped at least one nuclear and/or pre-nuclear accented word, and for deictic gesture, this number was 99%. In other words, the performance of the meaningful part of the gesture can be reliably predicted from the nuclear prominence in speech which in the default case of broad focus is a right-branching structure. Any occurrence of early pre-nuclear rise (marked by a high pitch contour) is also predictive for the occurrence of a stroke. Unlike prosody, in syntax, the gesture is not constrained to a particular syntactic category — the temporal co-occurrence of a gesture with, say, a sentence, a noun phrase, a verb phrase, or a verb itself is equally probable.

We used the empirical findings to spell out grammar construction rules that account on the one hand for the interaction between gesture and the (pre-)nuclear prominence of the temporally overlapping speech signal and on the other for the fact that the form of the hand does not uniquely determine its semantically synchronous phrase. Through attachment ambiguities, the gesture can be synchronised with a speech phrase that includes elements whose temporal performance is outside the temporal performance of the gesture. In this way, we approach the finding from the descriptive literature that gestures are synthetic, which contrasts to the analytic nature of spoken words; e.g., in (1) a single gestural movement conveys information about the event of giving, the object being given, and also about the recipient. We propose two basic rules that take this into account: Rule 1 allows for combining (depicting or deictic) gesture with a temporally overlapping prosodically prominent word. The feature structure representations of the construction rules for depicting and deictic gesture are shown in Fig.5 and Fig.6, respectively. Note that the rules contribute the underspecified relations *vis_rel* and *deictic_rel* between the speech content and the gesture content. Furthermore, Rule 2 involves attaching gesture to a constituent larger than the single prosodically prominent word. Since the feature structure schemata of Rule 1 and Rule 2 are almost identical, we do not show the feature structure representation of Rule 2.

Rule 1 *A depicting or deictic gesture can attach to a spoken word w if (a.) there is an overlap between the timing of the gesture stroke and w and (b.) w bears a pre-nuclear or nuclear accent.*

Rule 2 *A depicting or deictic gesture can attach to a constituent larger than the nuclear/pre-nuclear prominent word w, where w is the syntactic head, upon partially or fully saturating the head with the arguments and/or modifiers it selects if there is a temporal overlap between the performance of the gesture stroke and the performance of w.*

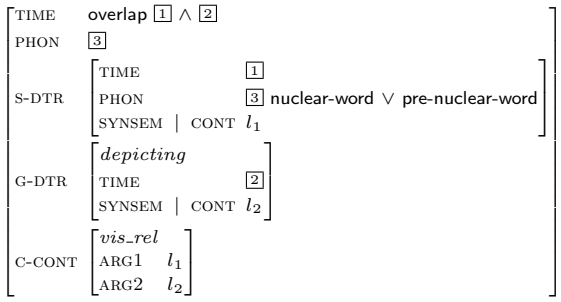

Fig. 5: Rule for Depicting Gesture and Speech Word

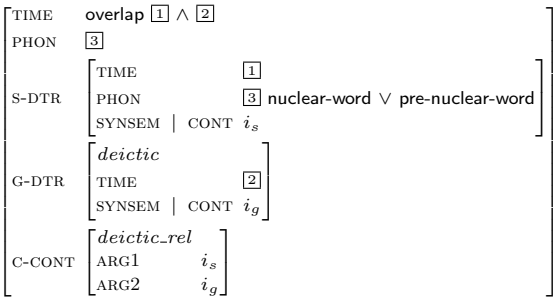

Fig. 6: Rule for Deictic Gesture and Speech Word

To demonstrate the syntactic derivation and the semantic composition, we shall be using utterance (4) where the speaker uses a deictic gesture to set up the location of her apartment.[2]

(4) 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.

Since we use RMRS for semantic composition within HPSG grammars, we first translate the deixis semantics in Fig.4 into a feature-structure representation (see the deixis feature structure in Fig.7). The EPs are encoded within the RELS feature where every predication introduces its own type and feature-value pairs. For the lack of space, we gloss over the relations produced by the deixis form features as *deixis_eps* and these include $RH\_hand\_shape\_open\_flat$, $RH\_palm\_orient\_upwards$, etc. To make the deixis structure available for composition, we augment it with TOP, HOOK and HCONS constraints as follows: the TOP label is a global label containing the whole formula and in composition the top label h_0 of the mother is identified with the top labels h_0 of the daughters to demonstrate the derivation of a single LF; the HOOK is a placeholder for missing information similar to a λ-abstracted term which contains: (a.) an LTOP: a local top which is equated with the label of an EP in a ULF, e.g., the deixis ltop is l_1[3] and (b). INDEX: a variable that indicates what the LF is about and it has two subtypes: events e and individuals x. The semantic index of the deixis corresponds to the main variable of sp_ref—the underspecified i; HCONS: scopal constraints which correspond to the $=_q$ scopal constraints.

Following Rule 1, the gesture could attach to the verb "enter": it bears a pre-nuclear accent and its temporal performance overlaps the temporal performance of the gesture (see Fig.7). The verb semantics is also encoded in terms of TOP, HOOK and RELS values and we forego any details about them since they are entirely based on the ERG. The semantic composition of the multimodal verb proceeds as follows: the top labels of the daughters are identified with that of the mother; the gesture relations are appended (as demonstrated by $\oplus$) to that of the speech daughter making thus the composition monotonic. In composition, the underspecified semantic index i of the gesture resolves to an event e_1. The construction rule contributes an underspecified relation *deictic_rel* between the semantic index e_2 of the speech and the semantic index e_1 of the gesture. Similarly to appositives in ERG, the label of this relation is shared with the label of the verb's predicate; in this way, anything outscoping the verb would also outscope the deictic relation.

Further attachments for utterance (4) are licensed by Rule 2: the rule uses partial/full saturation to remain as neutral as possible about the number of selected arguments, that is, the gesture can attach to the verb saturated with its complement only ("enter my apartment") or to the verb saturated with both the subject and the complement ("I enter my apartment"). Whereas an attachment to the VP supports an interpretation where the gesture denotes a salient feature of the apartment such as the rectangular shape of the entrance door, a higher attachment to the root node supports an interpretation from the speaker's viewpoint where the speaker performs the event of entering the apartment door.

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

[3] In ERG, quantifiers can float among scope bearing elements and this is why the LTOP is a label distinct of that of the quantifier.

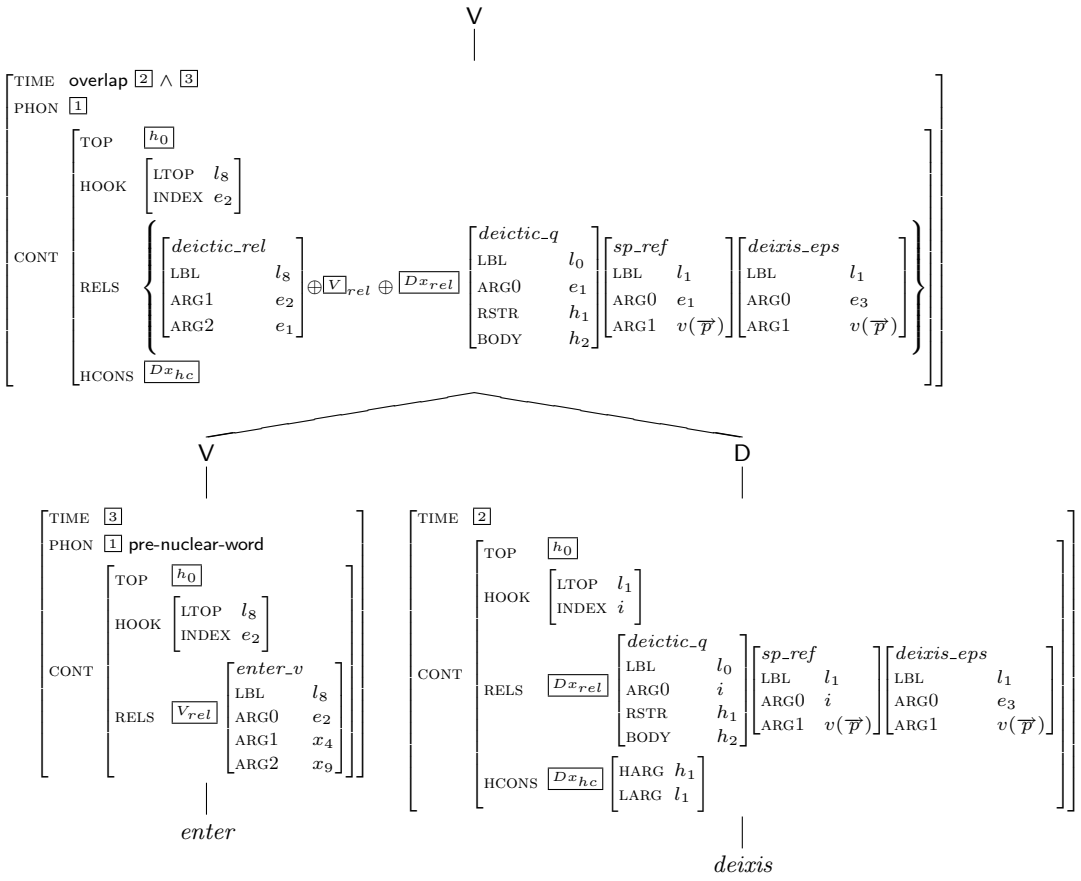

Fig. 7: Deriving Compositional Semantics for Multimodal Utterance in TFS

5 Resolving Underspecified Semantics

The problem of resolving the underspecified semantics of gesture is based on the compositional semantics of the multimodal signal and real-world knowledge. Whereas the specific mechanisms have been discussed elsewhere [7], it suffices to say that resolving *deictic_rel* is achieved by axioms of the sort: $(deictic\_rel(e_2, e_1) \wedge sp\_ref(e_1, v(\overrightarrow{p})) \wedge (\overrightarrow{p} \neq v(\overrightarrow{p}))) > VirtualCounterpart(e_1, e_2)$, which stipulates that if there is a deictic relation between the content denoted by gesture e_1 and the content denoted by the synchronous speech e_1, and that the spatial reference of e_1 is $v(\overrightarrow{p})$ but there is no identity between the physical space $\overrightarrow{p}$ and the gestural denotation $v(\overrightarrow{p})$, then the relation resolves to *VirtualCounterpart*.

6 Conclusions

In this paper, we demonstrated that standard linguistic methods for semantic composition can be applied to multimodal actions. We saw how the form of the gestural signal maps to highly underspecified predications resolvable to preferred values in discourse. The semantic composition in constraint-based TFS grammars supports firstly, underspecification in that it build an abstract representation supporting the final interpretations in context-of-use and secondly, monotonicity in that the semantics of the daughters is consistently appended to the semantics of the mother.

Acknowledgements. We are grateful to the anonymous reviewers, Ewan Klein, Jean Carletta, Jonathan Kilgour, Daniel Loehr and Mark Steedman. All errors remain our own.

References

1. Copestake, A. and Flickinger, D. An open-source grammar development environment and broad-coverage English grammar using HPSG. In Proceedings of LREC (2000)
2. Copestake, A., Lascarides, A. and Flickinger, D. An Algebra for Semantic Construction in Constraint-based Grammars. In Proceedings of the 39th Annual Meeting of ACL/EACL, Toulouse (2001)
3. Copestake, A. Slacker Semantics: Why Superciality, Dependency and Avoidance of Commitment can be the Right Way to Go. In EACL, pp 19 (2009)
4. Giorgolo, G. and Verstraten, F. Perception of speech-and-gesture integration. In Proceedings of the International Conference on Auditory-Visual Speech Processing (2008)
5. Kendon, A. Some relationships between body motion and speech. In Studies in Dyadic Communication (1972)
6. Kopp, S., Tepper, P. and Cassell, J. Towards integrated microplanning of language and iconic gesture for multi-modal output. In Proceedings of the 6th International Conference on Multimodal Interfaces, USA (2004)
7. Lascarides, A. and Stone, M. A Formal Semantic Analysis of Gesture. Journal of Semantics (2009)
8. Loehr, D. Gesture and Intonation. Washington DC: Georgetown University, Doctoral Dissertation (2004)
9. McNeill, D. Hand and Mind. What Gestures Reveal about Thought. Chicago: University of Chicago Press. (1992)
10. Pollard, C. and Sag, I. A. Head-Driven Phrase Structure Grammar. University of Chicago Press & CSLI Publications (1994)