Interactive Gesture in Dialogue: a PTT Model

Hannes Rieser
Bielefeld University
Hannes.Rieser@uni-Bielefeld.de

Massimo Poesio
Università di Trento/University of Essex
poesio@essex.ac.uk

Abstract

Gestures are usually looked at in isolation or from an intra-propositional perspective essentially tied to one speaker. The Bielefeld multi-modal Speech-And-Gesture-Alignment (SAGA) corpus has many interactive gestures relevant for the structure of dialogue (Rieser 2008, 2009). To describe them, a dialogue theory is needed which can serve as a speech-gesture interface. PTT (Poesio and Traum 1997, Poesio and Rieser submitted a) can do this job in principle, how this can be achieved is the main topic of this paper. As a precondition, the empirical research procedure from systematic corpus annotation via gesture typology to a partial ontology for gestures is described. It is then explained how PTT is extended to provide an incremental modelling of speech plus gesture in an assertion-acknowledgement adjacency pair where grounding between dialogue participants is obtained through gesture.

1 Introduction and Overview

We present work combining experimental methods, body-movement tracking techniques, corpus linguistics and theoretical modelling in order to investigate the role of iconic gesture in dialogue. We propose to map speech meaning and gesture meaning into a single compositional meaning which is then used in grounding and up-dating of information states in discourse, using PTT (Poesio & Traum 1997, Poesio & Rieser submitted 2009a) to account for the speech-gesture interface. We argue that several design features of PTT are essential for this purpose, such as accepting sub-propositional inputs, extracting information from linguistic surface, using dynamic semantics, basing the dialogue engine on a theory of groundedness and grounding, and allowing for the resolution of anaphora across turns.

The structure of the paper is as follows. Section 2 looks at the Bielefeld Speech-and-Gesture-Alignment corpus SAGA from which the data comes. Section 3 then deals with multi-modal acts using one example from SAGA (Dial 1 p.**??**). In section 4 a short introduction into PTT is provided. Sections 5 and 6 explain how a gesture typology and a partial ontology can be extracted from the annotated data. Both (see Appendix) serve as the basis for the integration of gesture meaning and verbal meaning. In section 7 PTT is developed as an interface for verbal and gestural meaning. First a PTT description of Dial 1 is provided using (Poesio and Rieser submitted b, Poesio to appear) dealing *inter alia* with anaphora resolution (7.1). Secondly, PTTs interface properties are detailed (7.2), the semantic defaults for combining speech and gesture meaning are set up (7.3), and a gestural dialogue act is described (7.4). Section 8 contains some preliminary insights into the grounding of multi-modal content.

2 The Multi-modal SAGA Corpus

The SAGA corpus contains 25 route-description dialogues taken from three camera perspectives using body tracking technologies.[1] The setting comes with a Router "riding on a car" through a virtual landscape passing five landmarks. The landmarks are connected by streets. Fig. 1a in Appendix B shows the Router, Fig. 1b the site, Fig.

[1] cf. Bergmann, K. *et al.* (2007, 2008)

Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pages 87–96,
Queen Mary University of London, September 2009.

1c the town hall. After the ride the Router reports his trip in detail to a Follower. We collected audio and body movement data as well as eye-tracking data from the Router. The dialogues have all been annotated, use of functional predicates like INDEXING, MODELLING, SHAPING[2] etc. was rated.

3 An Example from the SAGA corpus

In the dialogue passage (Dial 1) the Router uses gestures to explain the looks of the town-hall. We'll focus on the numbered utterances in this paper; utterances omitted in the reconstruction are reported in italics, omitted phrases in brackets.

DIAL 1 [ROUTER:] [...] *und [du] folgst dann dem*
[...] *and [you] follow then the*
Straßenverlauf einfach nur bis du ah vor nem
street simply until you ah before a
größeren Gebäude stehst.
larger building stand.

(2.1) Das ist dann das Rathaus.
That is then the townhall.

(2.2) [Ahm] das ist ein U-förmiges Gebäude.
[Ahm] that is a U-shaped building.

(2.3) Du blickst [praktisch] da rein.
You look [practically] into it.

(2.4) [Das heißt] es hat vorne zwei
[That is] it has to the front two
Buchtungen und geht hinten
bulges and closes in the rear
zusammen dann.
then.

[FOLLOWER:] OK.

In (Dial 1) Router's gestures first come with two BEATS.[3] Shortly after, the BEATs extend into an ICONIC gesture overlapping *town hall* in (2.1)(stills in Apendix B), cf. the still Two-Handed-Prism-Segment-1. Then the Router's DRAWN U-*shaped* gesture (still One-Handed-U-Shape) intersects the word U-shaped. Next his SHAPING the sides of a prism (still Two-Handed-U-Shaped-Prism-Segment) aligns with *[look pactically] into it.* The gesture following is two-handed: one hand SHAPES the U's left branch and the other both the U's right branch and its rear bend linking up to the left branch (stills Two-Handed-Prism-Segment-2A and 2B). The STROKE overlaps with the words *and closes in the rear.* The Follower copies the two-handed *town hall* gesture of the Router in his acknowledgement (still Two-Handed-Prism-Segment-3). In other words: the Follower's gesture is aligned to the Router's. Being copies of each other, the semantics of the Router's and the Follower's gesture can enter the common ground (*cf.* 7.4 and 8). In the reconstruction we will use the translation with the English word order standardised.[4]

4 A Short Introduction to PTT

Explanation of dialogue rests on three things: making clear how the succession of speakers' contributions emerges, stating what the impact of contributions on speakers' minds is and specifying how information is extracted incrementally from the contributions. Turning to emerging structure, PTT assumes that participants perform (often fragmentary) contributions, discourse units (DUs), which are dynamic propositions (DRSs in the sense of (Muskens, 1996)). They contain locutionary acts, conversational events/dialogue acts plus their propositional contents/DRSs. DUs may be sub-propositional micro-conversational events. Dialogue acts are either core speech acts or grounding acts. Core speech acts can be related to the present like *assert*, towards the past like *accept* or towards the future like *commit*. Grounding acts are *acknowledge* or *repair* (Traum 2009). Putting the distinctions above to work, we obviously can already model adjacency pairs. For the problems at issue we do not need more, cf. (Dial 1).

Which attitudes are assumed in current PTT and which changes of participants' minds are accounted for? Agents can have individual and private or common and public intentions. All sorts of actions, verbal or domain ones, are as a rule intended, at the outset of changes we have individual intentions. Common intentions are for example needed in order to explain completions and repairs (Poesio and Rieser submitted a). Most of the cooperation facts investigated in Clark (1996) need common intentions, most prominently, the intention to carry out a communicative task felicitously. Frequently, the vehicle for these types of intentions are (partial) plans. Plans can also be individual or shared. In (Dial 1) for example, the Router has an individual plan how to best map out his ride and the intention to communicate it to the

[2]Annotation PREDICATES are written in capital letters. Cf. also fn. 5.

[3]BEATS largely rest on supra-segmentals and would demand a paper of their own.

[4]We will end up with a mixture of German gesture and English wording here. However, for didactic purposes (sketch the main ideas) this seems acceptable. Sometimes we will simulate German constructions in English.

Follower. The Follower in turn intends to let the Router control her beliefs. Both have the collective intention to enable the Follower to follow the Router's route. Information presupposed or generated is contained in the discourse situation which, in PTT, is just a normal situation with objects and events, i.e., a DRS.

Conversational participants have command over information states. An information state is updated whenever a new event is perceived, including events such as sub-sentential utterances, and non-verbal events such as gestures or nods. Hence the possibility is already implemented in PTT to model accumulation of information due to gesture. Information common to the dialogue participants can be considered as grounded by default. This assumption connects PTT with other dialogue theories, for example Clark's (cf. Clark and Marshall, 1981, Clark and Schaefer, 1989) and Traum's (Traum 2009). Acknowledged information is at the heart of the grounding process. What is grounded is mutually believed *ceteris paribus.* Therefore, grounded information is part of the pragmatic machinery driving a dialogue forward (Rieser 2009). Grounding acts are taken as meta-discoursive devices and not included in discourse units proper. Besides beliefs and intentions we have obligations as mental attitudes. In PTT every conversational action induces an obligation on the participant indicated to address that action.

Information states raise the question of how changes of information are brought about on the basic grammatical level, viz. the interpretation of incrementally produced locutionary acts. The grammar in which syntactic and semantic interpretation is implemented is LTAG (Abeilleé & Rambow (eds), 2000). LTAG is a tree-grammar encoding syntactic projections which do the duty of, say, HPSGs rules, principles and constraints. Nodes and projecting leaves are decorated with semantic information based on Compositional DRT as developed in (Muskens, 1996, 2001). A specific trait of PTT is working with semantic non-monotonicity at all compositional levels: PTT hypothesizes that semantic computation is the result of defeasible inferences over DRSs obtained concatenating updates of single contributions. These default inference rules have the effect of semantic composition rules. Due to the impact of interpreted LTAG one can say that PTT is well founded in a bottom up fashion. Especially the default mechanism of PTT is used to make it a workable interface for speech and gesture (*cf.* 7.2 - 7.4).

5 Setting up the Speech-gesture Interface: Typology and Partial Ontology

As mentioned, this paper is based on the systematic annotation of SAGA carried out over the years 2007-2009 (Rieser 2009). Like many gesture researchers we assume that the semantic and pragmatic centre of a gesture is its **stroke**. The stroke overlaps as a rule with part of a complex constituent, for example the head or the logical subject. The range of speech-gesture overlap usually marks the functional position where the gestures meaning has to be merged into the speech content. Technically, the annotation is an ELAN-grid. From the annotation, a set of gesture types has been factored out in the following way (Rieser 2009). *AGENCY*[5] is installed as a root feature dominating the role features *ROUTER* and *FOLLOWER*. Next come the Router's and the Follower's *LEFT* and *RIGHT HAND* and *BOTH* their *HANDS*. *HANDEDNESS* in turn is mapped onto single annotation features like HANDSHAPE, WRISTMOVEMENT, PATHOFWRISTMOVEMENT etc. Bundles of features make feature *CLUSTERs* which yield classes of objects like curved, straight etc. entities. These build up *SHAPES* of different dimensions:[6] *ABSTRACT OBJECTs* of *0 DIMENSION* and *LINEs,* one-dimensional entities of different curvature. Among the two dimensional entities are *LOCATIONs, RECTANGLEs, CIRCLEs*[7] etc. Then three dimensional sorts come up: *CUBOIDSs CYLINDERs, PRISMs* and so on. In the end we get *COMPOSITEs* of *SHAPES,* for example a *BENT LINE* in a *SPHERE,* and S*EQUENCES OF COMPOSITEs.*[8] The central issue of 'How does a gesture acquire meaning'? is answered in the following way: A gesture type is mapped onto a partial ontology description, a stipulation encoding the content attributed to a gesture by raters. As a rule, gesture content is underspec-

[5]Gesture types, organised in an inheritance hierarchy working with defaults (cf. Rieser 2009), are written in *CAPITAL ITALICS*.

[6]In the following *geometry terms* are used mnemonically.

[7]*SHAPEs* can in general be fully developed or come in *SEGMENTs*. We do not deal with *SEGMENTs* here.

[8]*SEQUENCES* encode evolution of *SHAPEs* in time.

ified and will be completed to some extent when interfacing with verbal meaning. As an example of a gesture type and its partial ontology, see e.g. *TwoHandedPrismSegment1* and *'Partial OntologyTwoHandedPrismSegment1'* in Appendix A.

6 Setting up the Speech-gesture Interface: Levels of Interaction

Our starting point is the hypothesis detailed in (Rieser 2008) that a genuine understanding of dialogues like (Dial 1) requires integration of multi-modal meaning at different levels of discourse, from fine grained lexical definitions up to rhetorical relations. In the rest of the paper, we will specify how information from spoken utterances merges with information from gestures, using (Dial 1) as an example. Omitting the two BEATS on *that is [then],* we have the following gestures on the Router's side (see stills in Appendix B):

6.1 the *PRISM SEGMENT* covering *the town hall;* cf. still Two-Handed-Prism-Segment

6.2 the DRAWN U-shape overlapping the adjective *U-shaped;* still One-Handed-U-Shape

6.3 the *PRISM SEGMENT* affiliated to [practically] *look into it;* still Two-Handed-U-Shaped-Prism-Segment

6.4 the two-handed U-shaped *PRISM SEGMENT* going with *and closes in the rear;* stills Two-Handed-Prism-Segment-2A and 2B.

The Follower uses a variant of

6.5 the Router's *PRISM SEGMENT* in (6.4) followed by OK; still Two-Handed-Prism-Segment-3.

The key observation from Rieser (2009) is that gestures interact with verbal contributions at different levels. (6.1) to (6.4) must be integrated at the level of the semantic interpretation of LTAG. (6.3) is involved since the stroke covers three constituents in the German wording, the modal adverb *[practically],* the pronoun it, and the separable prefix *da rein/into* of the verb *blickst/you look.* We will develop a simplified solution here using the "verb" *look-into.* Similarly, in (6.4) the gesture contains information relevant for *closes in the rear,* i.e. for the whole VP. The gesture information has to be integrated into the Router's dialogue acts at the interface points mentioned. Therefrom several side issues arise, for example the treatment of anaphora across Router's or Follower's contributions. In (Dial 1) the Follower uses gestural information only to acknowledge. It is a multi-modal example of acknowledging by imitating the Router's multi-modal acts. Her gesture and the OK form a kind of "complex acknowledgement". This way the Router's contributions (6.2) to (6.4) and the Follower's contribution (6.5) show the interactive role of gesture, more specifically, gesture content in its use for grounding. We will briefly comment upon that in section 8.

7 Using PTT as an Interface for Verbal Meaning and Gestural Meaning

7.1 The verbal part of (Dial 1)

According to PTT, the discourse situation after the verbal updates brought about by (Dial 1) would be as follows.[9] (We only represent one aspect of the content of the initial utterances of (Dial 1).):

[DU0, DU1, DU2, DU3, DU4, DU5 |
DU0 **is** [. . . K1, . . . |
K1 **is** [b1 | **building**(b1), **large**(b1)],
. . .]
DU1 **is** [u2.1, K2, ce2.1 |
u2.1: **utter**(Router,"Das ist das Rathaus"),
sem(u2.1) is K2,
K2 is [th1, tnhl |
th1 **is** ιy1. K1; [| y1 is b1],
tnhl **is** ιu. [| **town hall** (u)],
th1 **is** tnhl,
ce2.1: **assert**(Router, Follower, K2),
generate(u2.1, ce2.1)],
DU2 **is** [u2.2, K4, ce2.2 |
u2.2: **utter**(Router, "das ist ein
U-förmiges Gebäude."),
sem(u2.2) **is** K4,
K4 **is** [th2 | th2 **is** ιy2. K5; [|s: y2 **is** b1],
building(th2), U-shaped(th2),
K5 is K1],
ce2.2: **assert**(Router, Follower, K4),
generate(u2.2, ce2.2)],
DU3 **is** [u2.3, K7, ce2.3|
u2.3: **utter**(Router, "Du blickst da rein"),
sem(u2.3) **is** K7,
K7 **is** [th3, s1 | th3 **is** ιy3. K8; [|s: y3 **is** b1],
s1: look-into(Follower, th3),[10]
K8 **is** K4],
ce2.3: **assert**(Router, Follower, K7),
generate(u2.3, ce2.3)],
DU4 **is** [u2.4, K9, ce2.4|
u4: **utter**(Router, "es hat vorne
zwei Buchtungen und geht hinten zus. dann"),
sem(u2.4) **is** K9,
K9 **is** [th4,bu1,bu2,s2,s3,s4,s5,s6,
re1,re2 |
th4 **is** ιy4. K10;
[| y4 **is** th3], K10 **is**
K7,
bulge(bu1), bulge(bu2),
s2: has(th4, bu1),

[9] Abbreviations used in the PTT-fragment: The prefixes are usually followed by a number n $\geq$ 0. DU = discourse unit, ce = conversational event, K = DRS, u = utterance, sem = semantic function, x, y, z . . . = DRs, e: event, s: = situation. In the DRSs ',' stands for conjunction an ';' between DRSs for composition of DRSs.

s3: has(th4, bu2),
to-front-of(bu1, th4),
to-front-of(bu2, th4),
rear(re1), s4: has(bu1,
re1), rear(re2)[11],
s5:has(bu2, re2),
s6: meet(re1, re2)],
ce2.44: **assert**(Router, Follower, K9),
generate(u2.4, ce2.4)].

The model of anaphora resolution accounting for the anaphoric cases is developed in (Poesio and Rieser, submitted 2009 b). The anaphoric *Das/this* in DU1 depends on the discourse entity *a larger building* introduced at the beginning of the conversation in DRS K1: K1 is the resource situation for the anaphoric definite. The second *das/this* still depends on the same resource situation. The pronouns, however, behave differently: Pronoun *da/there* in DU3 takes up the antecedent a *U-shaped building,* whereas the *es/it* in DU4 in turn refers to the *it* in DU3. Observe that the verbal part of (Dial 1) alone would already specify the interpretation completely: nothing essential is missing. As it will become clear below, what gestures do in this example is to add details to the verbally determined models and restrict the model set.

7.2 Tying in Gestures with Utterances

What we have got so far is a PTT-representation of the verbal part of (Dial 1). We now move on to how the information coming from the Router's gestures gets integrated with the verbal information – in particular, how this integration can take place below the sentential level. Our account builds on two key ideas from PTT. First of all, gestures are part of the discourse situation – i.e., the occurrence of gestures is recorded in the information state's representation of the discourse situation. Second, every occurrence of a sentence constituent counts as a conversational event – a MICRO CONVERSATIONAL EVENT (MCE). With these assumptions in place, the interaction of speech meaning and gesture meaning – how the two types of meanings combine to specify the overall meaning of a contribution – can be specified using the same mechanisms that specify the meaning of MCEs: i.e., with (prioritized) defaults in the sense of (Reiter, 1980, Brewka 1989). One example of a default specifying semantic composition is the BINARY SEMANTIC COMPOSITION (BSC) developed in (Poesio to appear, Poesio and Rieser submitted a) to specify the default way in which MCEs meanings can be derived from the meanings of their constituents. (We use the notation $>$ to indicate defeasible inference, $\uparrow$ to indicate 'dominated by'.)

BSC: u1$\uparrow$u, u2$\uparrow$u, sem(u1) **is** $\alpha_{\langle\sigma\iota\rangle}$ sem(u2) **is** β_σ, **complete**(u,u1,u2)
$>$ sem(u) **is** $\alpha(\beta)$

BSC can however be overridden in a number of circumstances: most notably, when anaphora interpretation processes identify a referent for a definite description like u_{NP1}: **utter**("the building"), in which case **sem**(u_{NP1}) will be the referent as opposed to a set of properties; or in cases of metonymy such as those studied by Nunberg (2004), in which the meaning of a MCE may be derived even more indirectly. We hypothesize that the integration of utterance meaning and gesture meaning is specified by **interface defaults** that may override the general meaning in a similar way by enriching the normal meaning of MCEs. We provide several examples of interface defaults below. For reasons of space, we only specify the results of default inference, without providing full derivations of the multi-modal meanings. For the gestures only the semantics[12] is specified, abstracted from the description of the partial ontology (cf. Appendix A for details). Utterance meaning then operates on the partial ontology information. MM abbreviates "multi-modal"; "lex-entry" means the word-form at stake, "lex-definition" means an explict dictionary definition for the word, for example in the style of the OED, cast into PL1.

7.3 The Interface Defaults

The general heuristic strategy for setting up interface defaults designed to combine verbal meaning and gesture meaning is to probe into the PTT structure as deep as you need in order to fit in the gestural content properly. Gestures may be relevant at any level of discourse, as shown in (Rieser, 2008) and demonstrated below; this means that sometimes gestural content has to be stored "deep

[10]Observe that the town-hall and the U-shaped building are the same.

[11]Observe that the gesture dynamically shapes two rears which meet.

[12]This is due to the fact that we do not integrate gestures into the discourse situation here. If these are integrated one will use their type description as syntax in AVM format. Gestures do not have the normal category syntax.

in" the lexical definition of a word, at other times one has to remain on the top level of semantic composition or even follow up the contributions produced so far. The interface defaults mostly follow the general schema:

λ-prefix mentioning the open parameters + lexicon definition + open parameters applied to iconic meaning = λ-abstracted partial ontology description where the λ-bound parameters secure binding.

An exception to this is (7.3.5.1) which uses the notion of satisfaction (see stills in Appendix B).

7.3.1 The PRISM SEGMENT aligned with *[the] town hall* (6.1). To begin with, gestural meaning can enrich the meaning of a nominal utterance. The interface default allowing this is called **Noun meaning extended (NMExt)**[13]

> **NMExt**: Noun(u), sem(u) **is** λx **lex-definition**(x), u↑u', N'(u'), u overlaps g, gesture(g), iconic-meaning(g) **is** λp partial ontology(p)
> >sem(u') is λx (lex-definition(x)) iconic-meaning(g)(x)

For instance in the dialogue under consideration **lex-definition** is the predicate 'large building used for the administration of local government' abbreviated as 'λPλx [[|s: large building(x), used for the administration of local government(x)]; P(x)]' and the Partial Ontology *TwoHandedPrismSegment1* from the Appendix A, resulting in the following meaning for the utterance of 'town hall' accompanied by the gesture:

> (7.3.1.1) λx [ls, rs, loc|s: large building(x), used for the administration of local government(x), side(ls, x), left(ls, Router), side(rs, x), right(rs, Router), location(loc, x)]

Observe that the fine-grained local information is provided by the gesture.

7.3.2 The DRAWN U-shape overlapping the adjective *U-shaped* is an example of gesture enriching an adjectival meaning through the interface default **Adjective meaning extended (AdjMExt)**

> **AdjMExt**: Adjective(u), sem(u) **is** λPλx [|lex-entry(x), P(x)], u↑u', N'(u'), u overlaps g, gesture(g), iconic-meaning(g) **is** λp partial ontology(p)
> > **sem**(u') **is** λPλQλx([|lex-entry(x), P(x)]; Q(x)) iconic- meaning(g)(x).

[13] 'λp partial ontology (p)' in NMExt and the following defaults is used in the following way: The expression 'partial ontology' refers to information from the partial ontology list in the Appendix A. What has to be chosen can be seen from the application of the default below.

Using **AdjMExt** and the meaning of the gesture *OneHanded-U-shape* in the Partial Ontology we obtain (7.3.2.1) as an enriched meaning for "U-shaped", '$\oplus$' denoting mereological composition:

> (7.3.2.1) λQλx([|U-shaped(x), λus(strai- ght-line(lr, us), arc(lb, us), straight-line(ll, us), us = lr $\oplus$ lb $\oplus$ ll)(x)]; Q(x))

After fitting in the noun modified by the multimodal content into position 'Q', the DRs will have to be correctly bound.

Observe that we could apply (NMext) and (AdjMext) iteratively to arrive at a complex MM Nom-meaning.

7.3.3 The PRISM SEGMENT affiliated to *[practically] look into it* is computed using the interface default **Verb meaning extended (VMExt)**.

> **VMExt**: VP(u), V(u1), NP(u2), u1↑ u, u2↑ u, sem(u1) **is** λPλx([|s: lex-definition(x), P(x)], u overlaps g, gesture(g), iconic-meaning(g) **is** λp partial ontology(p)
> > **sem**(u) **is** λPλx([|s: lex-definition(x), P(x)]) iconic-meaning(g)(x)

VMExt gives us, again using the information from the Partial Ontology *TwoHanded-U-shapedPrism* from the Appendix:

> (7.3.3.1) λx([|s: focus(agent, x), space(x), bounded(x), empty(x), λp[hl, ls, lel, fs, hr, rs, ler, d| prism(p), height(hl, ls), left-side(ls, p), front-side(fs, p), left(ls, Router), height(hr, rs), right-side(rs, p), length(ler, rs), right(rs, Router), length(lel, ls), distance(d, ls, lr), lel = ler](x)])

Again we see that fine-grained information is provided by the gesture, especially the pragmatic anchoring of the space looked into from the Router's position.

7.3.4 Finally, the two-handed U-shaped PRISM SEGMENT going with *and closes in the rear* needs a default **VP meaning extended (VPMextended)**. The gesture information is distributed among the verb "closes" and the PP "in the rear", the assumption being that the object closing does so at a particular location which is part of the object itself. So we have:

> **VPMExt**: V(u) ↑ VP(uph1), P(u) ↑ PP(uph2), Det(u) ↑ NP(uph3), Nom(uph4) ↑ NP(uph3), PP(uph2) ↑ VP(uph1), **sem**(u) **is** λPλx([|lex-definition(x)]; P(x)), u overlaps g, gesture(g), iconic-meaning(g) **is** λp partial ontology(p)
> > **sem**(uph1) = λPλx([|lex-definition(x), P(x)]; iconic-meaning(g)(x)

The default using Appendix A, Partial Ontology *TwoHanded-U-shapedPrism,* generates the following MM meaning:

(7.3.4.1) λx([|s: close(x), at(s, loc), prism(leftp), prism(rightp), part(leftp, x), part(rightp, x), section(sectl, leftp), leftside(lefts, leftp), length(ll, lefts), left(lefts, Router), section(sectr, rightp), rightside(rights, rightp), frontside(fronts, rightp), bent(rightp), meet(lefts, rights, loc), right(rights, Router), parallel(lefts, rights), distance(d, lfts, rhts)]).

7.3.5 The Follower's U-shaped gesture: So far, gesture meaning constrained word meanings or constituent meanings. In contrast, the Follower's U-shaped gesture invades dialogue structure. The Follower's reply has two steps. Her iconic gesture yields a predicate U-shaped in much the same way as the Router's contribution in DU2 and DU4 does. This is combined with a DR anaphorically linked to the Router's preceding *its* and *thats*. The gesture in turn takes up the Router's U-shapes from DU2 and DU4. So we get an anaphora related to antecedent multi-modal information.[14] Her "OK" then simply acknowledges her own DU5 filled up. Acknowledgement of the Router's contributions is achieved indirectly. In order to model all that, we have to Hook up the Gesture's Content with a DR. This is simply

(7.3.5.1) λp(iconic-meaning(p))DR for some preceding discourse referent DR satisfying iconic-meaning.

The relevant iconic meaning is taken from Partial Ontology *TwoHandedPrismSegment3:* section(sect, p), leftpart(lftp, p), lengthl(lftp), left(leftp, Follower), rightpart(rtp, p), right(rightp, Follower), lengthr(rtp), lftp = rtp, p = lftp $\oplus$ rtp.

7.4 A Gestural Dialogue Act of Assertion

Concerning dialogue structure, we have concentrated on the verbal part of (Dial 1) in 7.1. In the SAGA corpus there are many data showing how dialogue structure interfaces with gesture meaning. In 7.3.5 a default for the follower's U-shaped gesture was given. Its embedding into the PTT-description of (Dial 1) is shown in DU5 below:

(7.4) DU5 **is** [g1, K10|
  g1: gesticulate(Follower, Router, U-shape),
  sem(g1) **is** K10,
  K10 **is** [|s: th5 is th4, λp(section(sect, p),
    leftpart(lftp, p), lengthl(lftp),
    left(leftp, Follower), rightpart(rtp, p),
    right(rightp, Follower), lengthr(rtp),
    lftp is rtp, p is lftp $\oplus$ rtp)(th5))]
  ce5: **assert**(Follower, Router, K10),
  generate(g1, ce5)],],

[ce6, u6|
  u6: **utter**(Follower,"OK"),
  ce6: **ack**(Follower, DU5),
  textbfgenerate(u6, ce6)]]

In the multi-modal dialogue passage we have 'gesticulate' instead of 'utter'. The semantics, using the default (7.3.5.1) 'Hook up the Gesture's Content with a DR' and material from Appendix A is provided in the standard way by K10. It is assumed that gestural content can be generated and asserted. The Follower's acknowledgement is a sort of self-acknowledgement that percolates up through anaphora.

8 Grounding by Gesture: a Genuine Case of Gestural Alignment

The different defaults, Noun-meaning extended (**NMextended**), Adjective meaning extended (**AdjMextended**), Verb meaning extended (**VMextended**), VP meaning extended (**VPMextended**) and **Hook up the Gesture's Content with a DR**, clearly indicate that integration of gesture meaning has to operate on levels of different grain. Gesture can operate on a sub-word level if one has to attach its meaning to parts of a lexical definition, on the word level, on the level of constituents, and, as a consequence of all that, on specific dialogue acts. Furthermore, we have seen gesture at two inter-propositional levels at work, at the interface among the contributions of one agent (see Router's contributions which are all "united" by communicating the appearance of the town hall) and at the interface among contributions of different agents (Router-Follower).The Follower acknowledges by imitating gestures of the Router; this is a genuine case of gestural alignment. Alternatively, she could also acknowledge verbally, uttering 'U-shaped' but she chooses a gestural content. Obviously, speakers think that this works. Her 'OK' furthermore shows that verbal and gestural means can work in tandem. So, in the end, the U-shape of the town hall is rooted in the common ground by default and the Router can continue with describing the route leading to the next landmark.

Acknowledgements

Support by the SFB 674, Bielefeld University, is gratefully acknowledged. We also want to than three anonymous reviewers for carful reading

[14]These anaphorical relations are not reconstructed here but delegated to a follow-up paper.

and suggestions for improvement. Hannes Rieser wants to thank Florian Hahn for common work on gesture typology starting in 2007.

References

Abeillé, A & Rambow, O. (eds) 2004. *Tree Adjoining Grammars. CSLI Publ.* Stanford, CA.

Bergmann, K., Fröhlich, C., Hahn, F., Kopp, St., Lücking, A. and Rieser, H. June 2007. *Wegbeschreibungsexperiment: Grobannotationsschema.* Univ. Bielefeld, MS.

Bergmann, K., Damm, O., Fröhlich, C., Hahn, F., Kopp, St., Lücking, A., Rieser, H. and Thomas, N. June 2008. *Annotationsmanual zur Gestenmorphologie.* Univ. Bielefeld, MS.

Brewka, G. 1989. *Nonmonotonic reasoning: from theoretical foundation towards efficient computation.* Hamburg, Univ., Diss.

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

Clark, H. H. and Marshall, C. R. 1981. Definite Reference and Mutual Knowledge. In A. K. Joshi, B. Webber, and I. A. Sag (eds.),*Elements of Discourse Understanding.* CUP

Clark, H. H. and Schaefer, E. F. 1989. Contributing to Discourse. *Cognitive Science,* 13, 259-294.

Muskens, R. 1996. Combining Montague Semantics and Discourse Representation. *Linguistics and Philosophy 19,* pp. 143-186.

Muskens, R. 2001. Talking about Trees and Truth-conditions. *Journal of Logic, Language and Information, 10(4),* pp. 417-455.

Nunberg, G. 2004. The Pragmatics of Deferred Interpretation. In: Horn, L.R. and Ward, G.: *The Handbook of Pragmatics.* Blackwell Publishing Ltd.

Poesio, M. to appear. *Incrementality and underspecification in semantic interpretation.* CSLI Publications.

Poesio, M. February 2009. *Grounding in PTT.* Talk given at Bielefeld Univ.

Poesio, M. and Rieser, H. submitted a. *Completions, Coordination, and Alignment in Dialogue.*

Poesio, M. and Rieser, H. submitted b. *Anaphora and Direct Reference: Empirical Evidence from Pointing.*

Poesio, M. and Traum, D. 1997. "Conversational Actions and Discourse Situations", *Computational Intelligence,* v. 13, n.3, 1997, pp.1- 45.

Rieser, H. 2008. Aligned Iconic Gesture in Different Strata of MM Route-description Dialogue. In *Proceedings of LONdial 2008,* pp. 167-174

Rieser, H. 2009. On Factoring out a Gesture Typology from the Bielefeld Speech-And- Gesture-Alignment Corpus. Talk given at the *GW 2009*, ZiF Bielefeld, to appear in the Proceedings of GW 2009.

Traum, D. 2009. Computational Models of Grounding for Human-Computer Dialogue. Talk given at Bielefeld Univ., February 2009

Appendices

Appendix A: Gesture Types and Description of Partial Ontology

Due to limited space gesture types and ontology descriptions are only partially characterised.

TwoHandedPrismSegment 1	
R.G.Left.HandShapeShape	*loose B spread*
R.G.Left.HandPalmDirection	*PDN/PTR*
R.G.Left.BackOfHandDirection	*BAB*
R.G.Left.Practice	*grasping-indexing*
R.G.Left.Perspective	*speaker*
R.G.Right.HandShapeShape	*loose B spread*
R.G.Right.HandPalmDirection	*PDN/PTL*
R.G.Right.BackOfHandDirection	*BAB*
R.G.Right.Practice	*grasping-indexing*
R.G.Right.Perspective	*speaker*
R.Two-handed-configuration	*TT*
R.Movement-relative-to-other-hand	0

Partial Ontology TwoHandedPrismSegment 1	
R.G.Left.HandShapeShape-loose B spread	*side(ls, p)*
R.G.Left.HandPalmDirection-PDN/PTR	*left(ls, Router)*
R.G.Right.HandShapeShape-loose B spread	*side(rs, p)*
R.G.Right.HandPalmDirection-PDN/PTL	*right(rs, Router)*
R.Two-handed-configuration-TT	*location(loc, p)*

OneHanded-U-shape	
R.G.Right.HandShapeShape	*G*
R.G.Right.PalmDirection	*PDN/PTL>PDN/PTB>PDN*
R.G.Right.BackOfHandDirection	*BAB/BTL>BAB/BDN>BAB/BDN/BTL*
R.G.Right.PathOfWristLocation	*ARC*
R.G.Right.WristLocationMovementDirectio	*MR>MF>ML*
R.G.Right.Extent	*MEDIUM*
R.G.Right.Practice	*drawing*
R.G.Right.Pespective	*speaker*

Partial Ontology OneHanded-U-shape	
R.G.Right.PathOfWristLocation-ARC	*U-shape(us)*
R.G.Right.WristLocation MovementDirection-MR>MF>ML	*straight-line(lr, us)* ∧ *bent-line(lb, us)* ∧ *straight-line(ll, us)*

TwoHandedPrismSegment 2	
R.G.Left.HandShapeShape	*B spread*
R.G.Left.HandPalmDirection	*PTR*
R.G.Left.BackOfHandDirection	*BAB/BUP > BAB*
R.G.Left.PathOfWristLocation	*LINE*
R.G.Left.WristLocation MovementDirection	*MF*
R.G.Left.Practice	*shaping-modelling*
R.G.Left.Perspective	*speaker*
R.G.Right.HandShapeShape	*B spread*
R.G.Right.HandPalmDirection	*PTL*
R.G.Right.BackOfHandDirection	*BAB/BUP > BAB*
R.G.Right.PathOfWristLocation	*LINE*
R.G.Right.WristLocation MovementDirection	*MF*
R.G.Right.Practice	*shaping-modelling*
R.G.Right.Perspective	*speaker*
R.Two-handed-configuration	*PF*
R.Movement-relative-to-other-hand	*SYNC*

Partial Ontology TwoHandedPrismSegment 2	
R.G.Left.HandShapeShape-B spread	*hight(hl, ls)*
R.G.Left.HandPalmDirection-PTR	*leftside(ls, p)* ∧ *prism(p)*
R.G.Left.PathOfWristLocation-LINE	*length(lel, ls)*
R.G.Left.WristLocation MovementDirection-MF	*frontside(fs, p)*
R.G.Left.Perspective-speaker	*left(ls, speaker)*
R.G.Right.HandShapeShape-B spread	*hight(hr, rs)*
R.G.Right.HandPalmDirection-PTL	*rightside(rs, p)* ∧ *prism(p)*
R.G.Right.PathOfWristLocation-LINE	*length(ler, rs)*
R.G.Right.WristLocation MovementDirection-MF	*frontside(fs, p)*
R.G.Right.Perspective-speaker	*right(rs, speaker)*
R.Two-handed-configuration-PF	*distance(d, ls, lr)*
R.Movement-relative-to-other-hand-SYNC	*lel = ler*

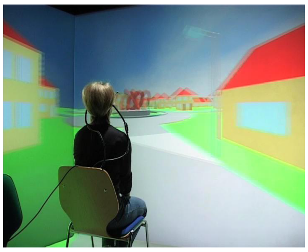
TwoHanded-U-shapedPrism	
R.G.Left.HandShapeShape	*small C*
R.G.Left.HandPalmDirection	*PAB*
R.G.Left.BackOfHandDirection	*BAB/BTR*
R.G.Left.PathOfWristLocation	*LINE*
R.G.Left.WristLocation MovementDirection	*MF*
R.G.Left.Practice	*shaping*
R.G.Left.Perspective	*speaker*
R.G.Right.HandShapeShape	*small C*
R.G.Right.HandPalmDirection	*PAB/PTL> PTL>PTB/PTL*
R.G.Right.BackOfHandDirection	*BAB/BTR> BAB>BAB/BTL*
R.G.Right.PathOfWristLocation	*LINE>LINE*
R.G.Right.WristLocation MovementDirection	*MF>ML*
R.G.Right.Practice	*shaping*
R.G.Right.Perspective speaker	
R.Two-handed-configuration	*BHA*
R.Movement-relative-to-other-hand	*SYNC*

Partial Ontology TwoHanded-U-shapedPrism	
R.G.Left.HandShapeShape-small C	*section(sectl, leftp)*
R.G.Left.PathOfWristLocation-LINE	*leftside(lefts, leftp)*
R.G.Left.WristLocation MovementDirection -MF	*length(ll, lefts)*
R.G.Left.Perspective-speaker	*left(lefts, speaker)*
R.G.Right.HandShapeShape-small	*section(sectr, rightp)*
R.G.Right.PathOfWristLocation-LINE>LINE	*rightside(rights, rightp)* ∧ *frontside(fronts, rightp)*
R.G.Right.WristLocation>ML MovementDirection-MF	*bent(rightp)* ∧ *meet(lefts, rights)*
R.G.Right.Perspective-speaker	*right(rights, speaker)*
R.Movement-relative-to-other-hand-SYNC	*parallel(lefts, rights)* ∧ *distance(d, lefts, rights)*

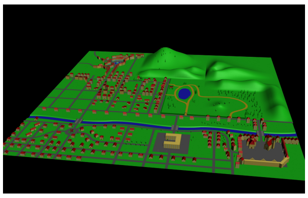
TwoHandedPrismSegment 3	
R.G.Left.HandShapeShape	*C*
R.G.Right.HandPalmDirection	*PDN/PTR>* *PAB/PUP*
R.G.Left.BackOfHandDirection	*BAB>* *BTL/BUP*
R.G.Left.PathOfWristLocation	*ARC*
R.G.Left.WristLocationMovementDirection	*ML>MB*
R.G.Left.Practice	*shaping*
R.G.Left.Perspective	*speaker*
R.G.Right.HandShapeShape	*C*
R.G.Right.HandPalmDirection	*PDN/PTL>* *PAB/PUP*
R.G.Right.BackOfHandDirection	*BAB>* *BTR/BUP*
R.G.Right.PathOfWristLocation	*ARC*
R.G.Right.WristLocationMovementDirection	*MR>MB*
R.G.Right.Practice	*shaping*
R.G.Right.Perspective	*speaker*
R.Two-handed-configuration	*BHA*
R.Movement-relative-to-other-hand	*Mirror-Sagittal*

Partial Ontology TwoHandedPrismSegment 3	
R.G.Left.HandShapeShape	*section(sect, p)*
R.G.Left.PathOfWristLocation	*leftpart(lftp, p)*
R.G.Left.WristLocationMovementDirection	*lengthl(lftp)*
R.G.Left.Perspective	*left(leftp, speaker)*
R.G.Right.HandShapeShape	*section(sect, p)*
R.G.Right.PathOfWristLocation	*rightpart(rtp, p)*
R.G.Right.WristLocationMovementDirection	*lengthr(rtp)*
R.G.Right.Perspective	*speaker*
R.Two-handed-configuration	$lftp = rtp$
R.Movement-relative-to-other-hand	$p = lftp \oplus rtp$

Appendix B: Figure 1



(a) The Router on his trip.



(b) The site traversed by the Router. The U-shaped building is the town hall

(c) Fig. 1c shows the town hall as described and gestured by the Router.

(d) Two-Handed-Prism-Segment-1

(e) One-Handed-U-Shape

(f) Two-Handed-U-Shaped-Prism-Segment

(g) Two-Handed-Prism-Segment-2A

(h) Two-Handed-Prism-Segment-2B

(i) Two-Handed-Prism-Segment-3

Figure 1: The SAGA Setting