Requesting clarifications with speech and gestures

Jonathan Ginzburg
Université de Paris
Laboratoire de Linguistique Formelle
Institut Universitaire de France
yonatan.ginzburg@u-paris.fr

Andy Lücking
Université de Paris
Laboratoire de Linguistique Formelle
Goethe-Universität Frankfurt
luecking@em.uni-frankfurt.de

Abstract

In multimodal natural language interaction both speech and non-speech gestures are involved in the basic mechanism of grounding and repair. We discuss a couple of multimodal clarification requests and argue that gestures, as well as speech expressions, underlie comparable parallelism constraints. In order to make this precise, we slightly extend the formal dialogue framework KoS to cover also gestural counterparts of verbal locutionary propositions.

1 Introduction

Detailed taxonomies of verbal Clarification Requests (CRs) already exist (Purver et al., 2003; Rodriguez and Schlangen, 2004) and accounting for these motivate theories of grounding and clarification interaction such as (Schlangen, 2004; Purver, 2006; Ginzburg, 2012), which provide wide coverage thereof. Although there exists some corpus-based and experimental work on multimodal repair (Healey et al., 2015; Seo and Koshik, 2010; Hough et al., 2015), detailed taxonomies are yet to be developed, nor formal accounts thereof.

In this paper we consider how to account for the multimodal versions of one of the commonest types of clarification request dubbed reprise fragments by Purver et al. (2003). Clarification requests play an important role in semantic methodology (Purver and Ginzburg, 2004) and in the construction of dialogue systems (Purver et al., 2011). Ginzburg and Cooper (2004) argue in detail that these exhibit significant syntactic and phonological parallelism with their source, as exemplified in (1a); concretely their claim is that the *intended content* reading ('what do you mean by . . . ') requires segmental identity with the source. A similar condition *mutatis mutandis* seems to be the case for gestural ones (2): (2a,b) involve clarifying a body movement (the former from example (1), Healey et al., 2014, 26, emphasis added), the latter two concern laughter, either with respect to content or in the latter case clarifying the manner of laughter ((2e) is from Fig. 1 of Healey et al., 2014, 26):[1]

(1) a. (i) A: Do you fear him? B: Fear? (= What do you mean by 'fear' or Are you asking if I *fear* him) / #Afraid? (ii) A: Are you afraid of him? B: Afraid? (= What do you mean by "afraid"? or Are you asking if I am *afraid* of him) / #Fear?

b. A: Are you afraid of him? B: Afraid? (= What do you mean by "afraid"?)

(2) a. *B*: You have to move your legs like this [*moves right hand up and down in a wave-like manner*]. *A*: [*moves right hand up and down in a wave-like manner, raises eye-brows*]

b. . . . and that movement really cracks your back

c. What's that? You do *that* and someone pulls?

d. A: I hear you're busy ⟨laughter ⟩ [= little giggle]. B: ⟨laughter ⟩ ? (= low arousal laughter with rising contour). (attested example)

e. Was it kind of like [H:o?]=
[H:hhh]

Clarification requests also occur on larger time scales, as is evinced in Figs. 1 to 3. The example is taken from the *Speech and Gesture Alignment*

[1] We use the letters 'A' and 'B' to denote the participants. Paraphrases of reprise fragments are introduced by an equation symbol, emphasis is indicated by italics, impossible or infelicitous clarifications are marked by '#'.

Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR), pages 21–31
June 16, 2021. ©2021 Association for Computational Linguistics

corpus SaGA (Lücking et al., 2010), which is a multimodal corpus of route direction dialogues. The example is about a section of a route where the route follower has to enter a park and walk around a pond, but not completely, just to three quarters. The route section is described by the route giver in Fig. 1. It is put to clarification by the addressee (route follower) in Fig. 2. Abstracting over perspective, the *moving around* movement is more or less kept constant, but modelling the pond is changed from a gesture hold to a drawing on the back of the hand. The route giver subsequently corrects the clarification by a path drawing on the addressee's back of hand in Fig. 3.

We show how to extend existing notions of conversational context and representation of speech multimodally to account for such cases. The basic extensions to the formal framework introduced in the following section are (i) multi-tier partiturs for capturing signals on different channels, (ii) a classification of gesture events on the tiers, and (iii) an anaphoric multimodal clarification rule requesting feedback concerning a previous multimodal fragment.

2 Background

Our account is formulated within *Type Theory with Records* (TTR, Cooper, 2005; Cooper and Ginzburg, 2015). TTR is a formal semantics framework based in the proof-theoretic, intuitionistic mathematics of Martin-Löf (1984). The reason for using a formal framework is that it enables researchers working on semantic phenomena in a scientific, precise manner. This is possible since the interpretation of types and structures used can be fixed in models—for such a denotational interpretation of TTR see Cooper (2021).[2] Although traditionally mainly applied to the compositional semantics of sentences, semanticists working on dialogue soon developed conversation-oriented extensions (just think of the content of particles such as *Hi!* or *Yes* or highly normative patterns such as question–answering.) However, classifying (multimodal) natural language utterances is not always a binary affair (think, e.g., of vagueness). To this end, there are probabilistic interpretations of TTR (Cooper et al., 2015). Although we could render our discussion in probabilistic terms,[3] we refrain from doing so since this paper is not concerned with probabilistic phenomena as such and this keeps representations simpler. TTR integrates logical techniques such as the lambda calculus and the expressiveness of feature-structure like objects (namely records and record types). A typing *judgement* $a : T$ is true iff object a is of type T. Types constructed from n-ary predicates ($n > 0$) are *dependent* on the values assigned to the *labels* that appear as arguments. Thus, if $a_1 : T_1$, $a_2 : T_2(a_1)$, ..., $a_n : T(a_1, a_2, \ldots, a_{n-1})$, then the record on the left in (3) is of the record type on the right in (3):

(3) $$\begin{bmatrix} l_1 = a_1 \\ \vdots \quad \vdots \\ l_n = a_n \end{bmatrix} : \begin{bmatrix} l_1 : T_1 \\ \vdots \quad \vdots \\ l_n : T(l_1, l_2, l_{n-1}) \end{bmatrix}$$

The notation $[l = a : T]$ represents a *manifest field* (Coquand et al., 2003). It is a notational convention for a *singleton type* T_a, where for any $b, b : T_a$ iff $b = a$.

Merge types correspond to unification in feature-structure formalisms. A merge '$\dot\wedge$' is exemplified in (4):

(4) a. $A = \begin{bmatrix} l_1 : T_1 \\ l_2 : T_2(l_1) \end{bmatrix}$ and $B = \begin{bmatrix} l_3 : T_3 \end{bmatrix}$

b. $A \mathbin{\dot\wedge} B = \begin{bmatrix} l_1 : T_1 \\ l_2 : T_2(l_1) \\ l_3 : T_3 \end{bmatrix}$

Drawing on work of Fernando (2007, 2011), TTR comes with a string theory of events. For three events e_1, e_2 and e_3, the string $e_1e_2e_3$ represents a course of events, namely the succession of e_1, e_2 and e_3, in that order. The notation $e_1e_2e_3$ is an abbreviation for a time-indexed record:

(5) $\begin{bmatrix} t_0 = e_1 \\ t_1 = e_2 \\ t_3 = e_2 \end{bmatrix}$, where time indices t_i are in $\mathbb{N}$.

If $e_1 : T_1$, $e_2 : T_2$ and $e_3 : T3$, then $e_1e_2e_3 : T_1{}^\frown T_2{}^\frown T_3$—the type constructor '$^\frown$' builds string types out of types. In order to exploit feature structure expressiveness in string types, a string of record types can be build by the same means, but is notationally enclosed in brackets.

[2] The semantic status of natural language processing (NLP) remains unclear, to say the least, as recently pointed out by Bender and Koller (2020). However, theoretical work such as the one developed here can of course trigger NLP applications, where, e.g., theoretically derived labels are automatically annotated on large(r) scale data.

[3] With some repercussions for some versions on its own (Larsson, 2020).

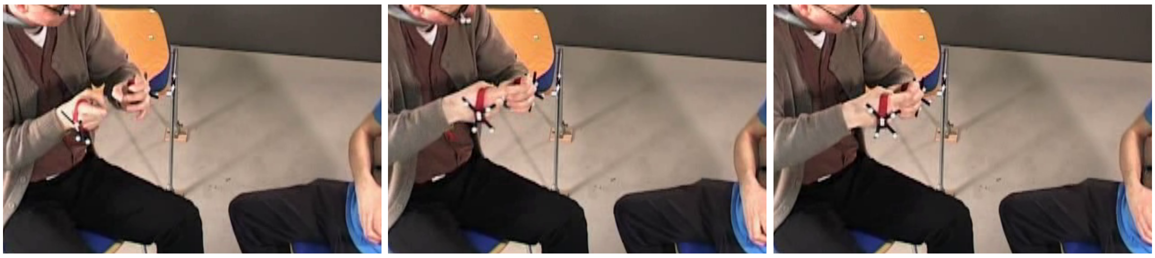


Figure 1: [Du fährst] ‘um den Teich herum’ ([You drive] *around the pond*): Index finger and thumb of left hand form a circle and right hand with stretched index finger is moved to three quarters around left hand.

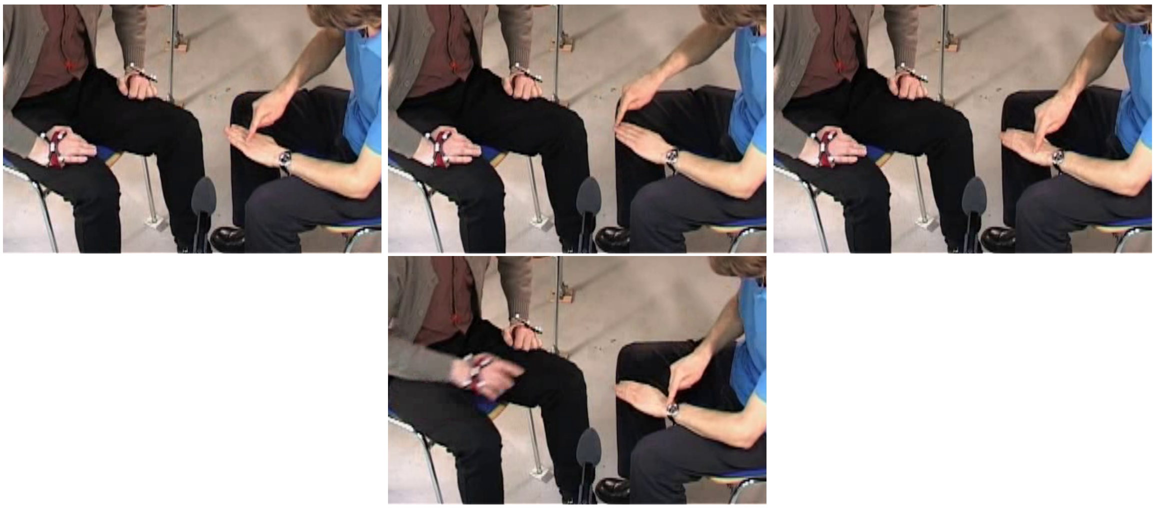


Figure 2: ‘Hier ist der Teich [Frame 1]. Ich komm’ auf den zu [Frames 2–3]. Und was heißt “rechts ab”? [Frame 4]’ (*Here is the pond* [Frame 1]. *I approach it* [Frames 2–3]. *And what do you mean ‘turn right?’* [Frame 4]): A circular index finger drawing gesture indicates the pond [Frame 1]. The index finger is first moved towards and then around the virtual pond [Frames 2–3]. A straight movement towards the wrist indicates *turning right* [Frame 4].

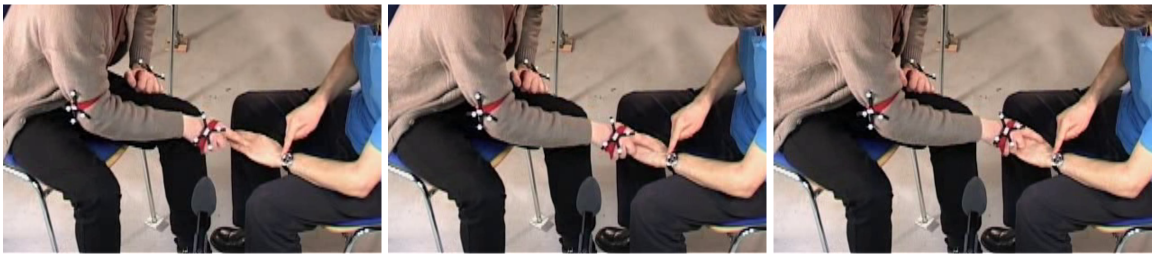


Figure 3: ‘Du fährst noch weiter rum.’ (*You drive around even more.*): Stretched index finger is moved around the virtual pond.

Making use of TTR, the simplest model of context, going back to Montague (1974) is one which specifies the existence of a speaker, addressing an addressee at a particular time. This can be captured in terms of the type in (6).

(6)
$$\left[\begin{array}{ll} \text{spkr} & : \textit{Ind} \\ \text{addr} & : \textit{Ind} \\ \text{u-time} & : \textit{Time} \\ \text{c}_{\text{utt}} & : \text{addr(spkr,addr,u-time)} \end{array}\right]$$

However, over recent decades it has become clearer how much more pervasive reference to context in interaction is. The visual situation is a key component in interaction from birth (see Tomasello, 1999, Chap. 3). Expectations due to illocutionary acts—one act (querying, assertion, greeting) giving rise to anticipation of an appropriate response (answer, acceptance, counter–greeting), also known as adjacency pairs (Schegloff, 2007). Extended interaction gives rise to shared assumptions or *presuppositions* (Stalnaker, 1978), whereas epistemic differences that remain to be resolved across participants—*questions under discussion* are a key notion in explaining coherence and various anaphoric processes (Ginzburg, 2012; Roberts, 1996). These considerations among several additional significant ones lead to positing a significantly richer structure to represent each participant's view of publicized context, the *dialogue gameboard* (DGB), whose basic make up is given in (7), following the recent version of the dialogue semantic framework called *KoS* including *mood* described by Ginzburg et al. (2020b):

(7) *DGBType* :=
$$\left[\begin{array}{ll} \text{spkr} & : \textit{Ind} \\ \text{addr} & : \textit{Ind} \\ \text{utt-time} & : \textit{Time} \\ \text{c-utt} & : \text{addressing(spkr,addr,utt-time)} \\ \text{facts} & : \textit{Set(Prop)} \\ \text{vis-sit} & = \left[\text{foa} : \textit{Ind} \vee \textit{Sit}\right] : \textit{RecType} \\ \text{pending} & : \textit{List(LocProp)} \\ \text{moves} & : \textit{List(IllocProp)} \\ \text{qud} & : \textit{poset(Question)} \\ \text{mood} & : \textit{Appraisal} \end{array}\right]$$

Here *facts* represents the shared assumptions of the interlocutors—identified with a set of propositions. *Vis-sit* represents the visual situation of an agent, including his or her focus of attention (*foa*), which can be an object (*Ind*), or a situation or event (*Sit*). The remaining fields concern locutionary and illocutionary interaction: Dialogue moves that are in the process of being grounded or under clarification are the elements of the *pending* list; already grounded moves are moved to the *moves* list. Within *moves* the first element has a special status given its use to capture adjacency pair coherence and it is referred to as *LatestMove*. The current question under discussion is tracked in the *qud* field, whose data type is a partially ordered set (*poset*). *Mood* tracks public displays of emotion, crucial for *inter alia* laughter and smiling (Ginzburg et al., 2020b).

The evolution of context in interaction is described in terms of *conversational rules*, mappings between two cognitive states, the *precond(ition)s* and the *effects*. Some examples of such rules are given in (8):

(8) a. Ask QUD-incrementation: given a question *q* and ASK(A,B,q) being the LatestMove, one can update QUD with *q* as MaxQUD.
$$\left[\begin{array}{ll} \text{pre :} & \left[\begin{array}{ll} \text{q} & : \text{Question} \\ \text{LatestMove = Ask(spkr,addr,q)} & : \text{IllocProp} \end{array}\right] \\ \text{effects :} & \left[\text{QUD} = \left\langle \text{q,pre.QUD} \right\rangle : \text{poset(Question)}\right] \end{array}\right]$$

b. Assert QUD-incrementation: a straightforward analogue for assertion of (8a): given a proposition *p* and ASSERT(A,B,p) being the LatestMove, one can update QUD with *p*? as MaxQUD.
$$\left[\begin{array}{ll} \text{pre :} & \left[\begin{array}{ll} \text{p} & : \text{Prop} \\ \text{LatestMove = Assert(spkr, addr, p)} & : \text{IllocProp} \end{array}\right] \\ \text{effects :} & \left[\text{QUD} = \left\langle \text{p?, pre.QUD} \right\rangle : \text{poset(Question)}\right] \end{array}\right]$$

c. QSPEC: this rule characterizes the contextual background of reactive queries and assertions—if *q* is MaxQUD, then subsequent to this either conversational participant may make a move constrained to be *q*-specific (i.e., either About or Influencing *q*).
$$\left[\begin{array}{ll} \text{pre :} & \left[\text{QUD} = \left\langle \text{q, Q} \right\rangle : \text{poset(Question)}\right] \\ \text{effects :} & \left[\begin{array}{l} \text{r : Question} \vee \text{Prop} \\ \text{R: IllocRel} \\ \text{LatestMove = R(spkr, addr, r) : IllocProp} \\ \text{c1 : Qspecific(r, q)} \end{array}\right] \end{array}\right]$$

As emphasized by Clark (1996) and by work in Conversation Analysis (CA; Schegloff et al., 1977) grounding and clarification interaction are important structuring processes in interaction. In Ginzburg (2012) these are modelled as a process triggered by awareness of an utterance event u and the attempt to instantiate the fields of an utterance type T_u emergent from parsing and resolving u. The pair of u and T_u is referred to as *locutionary proposition LocProp*. This is a special kind of (Austinian) proposition—records of type $\begin{bmatrix}\text{sit} & : \mathit{Rec}\\ \text{sit-type} & : \mathit{RecType}\end{bmatrix}$ (Austin, 1950; Barwise and Etchemendy, 1987)[4]—where *sit* is an utterance event and *sit-type* the type of a grammatical sign. This allows *inter alia* access to the individual constituents of an utterance. Purver (2004) and Ginzburg (2012) show how to account for the main classes of CRs using rule schemas of the form "if u is the interrogative utterance and $u0$ is a constituent of u, allow responses that are *co-propositional*[5] with the clarification question $\text{CQ}^i(u0)$ into QUD.", where '$\text{CQ}^i(u0)$' is one of the three types of clarification question (repetition, confirmation, intended content) specified with respect to $u0$.

Thus, the schema 'if u is an utterance spoken by A and $u0$ is a constituent of u, allow responses that are *co-propositional* with "What did A mean by u"' can be formulated as in (9): the issue $q0$, *what did A mean by u0*, for a constituent $u0$ of the maximally pending utterance, A its speaker, can become the maximal element of QUD, licensing follow up utterances that are CoPropositional with $q0$. Assuming a propositional function view of questions, CoPropositionality allows in propositions from the range of $\mathit{Range}(q0)$ and questions whose range intersects $\mathit{Range}(q0)$. Since CoPropositionality is reflexive, this means in particular that the inferred clarification question is a possible follow up utterance, as are confirmations and corrections, as exemplified in (10a–c).

(9) Parameter identification:

$$\begin{bmatrix}\text{pre} & : & \begin{bmatrix}\text{MaxPENDING} = \begin{bmatrix}\text{sit = u}\\ \text{sit-type =T}_u\end{bmatrix} : \text{LocProp}\\ \text{A = u.dgb-params.spkr : IND}\\ \text{u0 : sign}\\ \text{c1 : Member(u0,u.constits)}\end{bmatrix}\\ \text{effects} & : & \begin{bmatrix}\text{MaxQUD} = \lambda x\text{Mean(A,u0,x) : Question}\\ \text{LatestMove : LocProp}\\ \text{c1: CoPropositional(LatestMove.cont,MaxQUD)}\end{bmatrix}\end{bmatrix}$$

(10) a. $\lambda x.\mathit{Mean}(A,u0,x)$

b. $?\mathit{Mean}(A,u0,b)$ ('Did you mean Bo?')

c. $\mathit{Mean}(A,u0,c)$ ('You meant Chris.')

3 Partiturs

In order to utilize the information state update semantics of KoS for analysing multimodal discourse, we add extra structure to the utterance events by incorporating tiers. Tiers can be likened to different instruments on a musical score: a partitur.[6] We represent partiturs as *strings* of multimodal communication events, which is a temporally ordered sequence of types. One can think of strings in term of a flip-book: a dynamic event is cut into slices, and each slice is modeled as a record type. Such *string types* (Fernando, 2007; Cooper, 2021) are notated in round brackets:

(11)
$$\mathit{partitur} := \begin{bmatrix}\text{e} : (\begin{bmatrix}\text{e}_{\text{speech}} & : \mathit{Phon}\\ \text{e}_{\text{gesture}} & : \mathit{Trajectory}\\ \text{e}_{\text{gaze}} & = \mathit{vis\text{-}sit} : \mathit{RecType}\\ \text{e}_{\text{head}} & : \mathit{headMove}\\ \text{e}_{\text{face}} & : \mathit{faceExpr}\end{bmatrix})^+\end{bmatrix}$$

The progressive unfolding of sub-events on the various tiers in time gives rise to incremental production and perception. Formally, this is indicated by the Kleene plus ('$^+$'): the string type in (11) classifies events which consists of a sequence of multimodal communication signals. Hence, partiturs provides a formal means for describing cross-tier interaction.

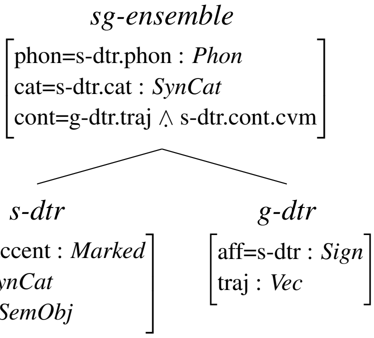
[4] On this view, a proposition p = $\begin{bmatrix}\text{sit} & = s\\ \text{sit-type} & = T\end{bmatrix}$ is true iff $s : T$—the situation s is of the type T.

[5] Here *CoPropositionality* for two questions means that, modulo their domain, the questions involve similar answers: for instance 'Whether Bo left', 'Who left', and 'Which student left' (assuming Bo is a student.) are all co-propositional.

[6] On a descriptive level, partiturs are akin to XML-encoded messages in the *Behavior Markup Language* (BML; Vilhjálmsson et al., 2007). But while BML is designed to define the generation of multimodal behavior in virtual agents, partiturs provide a platform for compositional multimodal chart parsing.

In order to model one sort of multimodal integration we make use of the account of speech-gesture of Lücking (2013), respectively its TTR reformulation (Lücking, 2016). Speech-gesture integration on this account is modelled in terms of a *speech-gesture ensemble* (Kendon, 2004), where a gesture (G-DTR) from tier e_{gesture} attaches to a phonetically marked *affiliate* (AFF; Schegloff, 1984) from speech (S-DTR, tier e_{speech}). Thus, multimodal integration of this sort is constrained by both temporal alignment and phonetic-kinematic interface (cf. also Alahverdzhieva et al., 2017). Semantic integration is formally governed by a imagistic feature called *conceptual vector meaning* ("CVM"). CVM draws on abstract motion perception from psychophysics (Johansson, 1973) and can in semantics formally spelled out in terms of vector-based representations of shapes, movements, orientations, or object axes within the vector space algebra of Zwarts (2003). The basic integration scheme is given in (12):

(12)



The underlying rationale of (12) is that both a gesture movement and a CVM value is a trajectory that is mathematically described as a sequence of vectors in three dimensions ($\mathbb{R}^3$; or $\mathbb{R}^4$ if the temporal dimension is explicitly built in). Drawing on work in gesture annotation, gestures are represented in terms of their kinematic features, giving rise to a 'phonetic' gesture representation. For example, moving the wrist rightwards, back (i.e., towards the body of the gesturer), and leftwards in a rectangular manner ('line')— $\left[\begin{array}{l}\text{path} : \textit{line} \\ \text{wrist=mr}^\frown\text{mb}^\frown\text{ml} : \textit{Move}\end{array}\right]$— a cornered, horseshoe-shaped trajectory '⊔' is displayed. Via a translation procedure from gesture representations onto vector representations, the abstract trajectory in (13) is obtained (Lücking, 2016).

(13)

$$\left[\begin{array}{l} \text{aff} = \left[\text{phon} : \left[\text{accent} : \textit{marked}\right]\right] : \textit{sign} \\ \text{traj} = \left[\begin{array}{l} \text{pt} : \left[\begin{array}{l} \mathbf{u}\perp\mathbf{v}\perp\mathbf{w} \\ \mathbf{u}(0)\neq\mathbf{w}(1)\end{array}\right] \\ \text{sh} : \left\{\text{rectangular, open}\right\}\end{array}\right] : \textit{Vec}\end{array}\right]$$

Spatial predicates also carry trajectory information as part of their CVM feature. The vector sequence from (13) is part of the lexical entry of the adjective *u-shaped* (it modifies a nominal, whose content is an individual).

(14)

$$\left[\begin{array}{l} \text{phon} : \left\langle \texttt{u-shaped}\right\rangle \\ \text{mod} : \left[\text{cat} : \left[\begin{array}{l}\text{head} : \textit{noun} \\ \text{cont} : \textit{Ind}\end{array}\right]\right] \\ \text{cont} = \left[\begin{array}{l} \text{cvm} = \left[\begin{array}{l} \text{pt} : \left[\begin{array}{l} \mathbf{u}\perp\mathbf{v}\perp\mathbf{w} \\ \mathbf{u}(0)\neq\mathbf{w}(1)\end{array}\right] \\ \text{sh} : \left\{\text{rectangular, open}\right\}\end{array}\right] : \textit{Vec} \\ \text{c}_{\text{shape}} : \text{shape(mod.cat.cont, cvm)}\end{array}\right] \\ \quad : \textit{RecType}\end{array}\right]$$

Since the gesture's trajectory and the adjective's CVM value are compatible, both can merge into a *sg-ensemble*.[7] Abstracting away from concrete movements to abstract vector representations seem to provide a format that is appropriate for gestural parallelism constraints, as will be discussed in Sec. 4.

An example involving the 'u-shape' gesture is used by Lücking and Ginzburg (2020): *the house [has a RECtangular]* ⊔ *shape*. The noun phrase *the house has a rectangular shape* is accompanied by a rectangular shape gesture which temporally overlaps the bracketed portion of speech. This tier-crossing utterance is incrementally processed by a multimodal chart parser (Earley, 1970; Johnston et al., 1997; Ginzburg et al., 2020a; Alahverdzhieva et al., 2017). The string chart in (15) represents the state after having processed *the house has* and the gesture's preparation phase. Due to this input, a VP rule (e_9) and a gesture integration rule (e_{10}) have been triggered, but are still pending:

[7] The example illustrates the gist of one form of multimodal integration. Much needs to be said, of course, for instance, on timing, affiliation, and more complicated ways of semantic integration—further details can be found in the references provided here.

(15)
$$\left[\begin{array}{lll}
e_1 & = & \texttt{the} : \textit{Phon} \\
e_2 & : & \text{Lex('the', DET)} \mathrel{\dot{\wedge}} \begin{bmatrix}\text{s-event} : \begin{bmatrix} \text{e}{=}e_1 : \text{/the/}\end{bmatrix}\end{bmatrix} \\
e_3 & : & \left(\begin{bmatrix}\text{rule}{=}\text{NP}{\rightarrow}\text{DET N} : \text{DET}{}^\frown\text{N} \\ \text{fnd}{=}e_2 : \textit{Sign}\end{bmatrix} {}^\frown \begin{bmatrix}\text{fnd}{=}e_5 : \textit{Sign}\end{bmatrix}\right) \\
e_4 & = & \texttt{house} : \textit{Phon} \\
e_5 & : & \text{Lex('house', N)} \mathrel{\dot{\wedge}} \begin{bmatrix}\text{s-event} : \begin{bmatrix} \text{e}{=}e_4 : \text{/house/}\end{bmatrix}\end{bmatrix} \\
e_6 & = & \texttt{prep} : \textit{Phase} \\
e_7 & = & \texttt{has} : \textit{Phon} \\
e_8 & : & \text{Lex('have', V)} \mathrel{\dot{\wedge}} \begin{bmatrix}\text{s-event} : \begin{bmatrix} \text{e}{=}e_7 : \text{/has/}\end{bmatrix}\end{bmatrix} \\
e_9 & : & \left(\begin{bmatrix}\text{rule}{=}\text{VP}{\rightarrow}\text{V NP} \\ \text{fnd}{=}e_8 : \textit{Sign} \\ \text{req}{=}\text{NP} : \textit{Sign} \\ \text{e} : \text{required(req,rule)}\end{bmatrix}\right) \\
e_{10} & : & \left(\begin{bmatrix}\text{rule}{=}\text{sg-ensemble}{\rightarrow}\text{X[accent,cvm] stroke} \\ \text{fnd}{=}e_6 : \textit{Phase} \\ \text{req1}{=}\text{stroke} : \textit{Phase} \\ \text{req2}{=}\text{X[accent,cvm]} : \textit{Sign} \\ \text{e} : \text{required(req1,req2,rule)}\end{bmatrix}\right) \\
\text{e} & : & \left(\begin{bmatrix} e_1 : \text{start}(e_1) \\ e_2 : \text{start}(e_2)\end{bmatrix} {}^\frown \begin{bmatrix} e_1 : \text{end}(e_1) \\ e_2 : \text{end}(e_2) \\ e_3 : \text{start}(e_3) \\ e_4 : \text{start}(e_4) \\ e_5 : \text{start}(e_5) \\ e_6 : \text{start}(e_6)\end{bmatrix} {}^\frown \begin{bmatrix} e_3 : \text{end}(e_3) \\ e_4 : \text{end}(e_4) \\ e_5 : \text{end}(e_5) \\ e_6 : \text{end}(e_6) \\ e_7 : \text{start}(e_7) \\ e_8 : \text{start}(e_8) \\ e_9 : \text{start}(e_9) \\ e_{10} : \text{start}(e_{10})\end{bmatrix}\right. \\
 & & \left. {}^\frown \begin{bmatrix} e_7 : \text{end}(e_7) \\ e_8 : \text{end}(e_8)\end{bmatrix}\right)
\end{array}\right]$$

Note that a multimodal ensemble—e_{10} in (15) and (14)—differs from phrasal constructions usually described by grammar: while the constituents of phrases are serialized (as captured in the string type 'e' in (15)), constituents of ensembles usually co-occur. In terms of locutionary propositions, the structure of an ensemble—consisting of a manual gesture and speech—is as in (16):

(16)
$$\left[\begin{array}{ll}
\text{mm-event} : & \begin{bmatrix} \text{u-time} : \textit{Time} \\ \text{spkr} \;\; : \textit{Ind} \\ \text{addr} \;\; : \textit{Ind} \\ e_{\text{sync}} \;\; = \begin{bmatrix} e_{\text{speech}} : \textit{Phon} \\ e_{\text{r-hand}} : \textit{Trajectory}\end{bmatrix} : \textit{Rec}\end{bmatrix} \\
\text{syn} & : \begin{bmatrix} \text{cat}{=}\text{mm-ensemble} : \textit{SynCat} \\ \text{drts}{=}\text{mm-event}.e_{\text{sync}} : \textit{Sign}^* \end{bmatrix} \\
\text{cont} & : \textit{SemObj}
\end{array}\right]$$

In contrast to the 'horizontal' chart parsing edges represented in terms of string types in the preceding incrementally growing partiturs, the daughters of multimodal ensembles are combined *via* 'vertical' edges. Such edges are defined in multichart parsers which have been developed exactly for the purpose of processing multimodal input (Johnston et al., 1997; Alahverdzhieva et al., 2017). We notate tier-crossing bindings on the level of utterance events (where an utterance comprises speech and gesture) in terms of the reserved label e_{sync}—such combined representations are object of at least one class of gestural clarifications.

4 Gestural Clarification: the case of reprise fragments

In this section we show how to modify an existing account of speech reprise fragments with minimal additions, though important empirical questions about the unity of this type of clarification request remain.

The analysis proposed by Ginzburg (2012) for this class of reprise fragments involves two components:

1. A construction *utt-ana-ph* that enables deixis to the repaired constituent under the constraint of segmental phonological parallelism. This is captured by identifying the phonological type of the clarification seeking utterance with that of the repaired constituent *rc.sit-type.phon*; whereas the content is identified with the speech event of the repaired constituent *rc.sit*. This makes crucial use of the fact that locutionary propositions store both type and token information:[8]

(17) *utt-ana-ph* =
$$\begin{bmatrix} \text{dgb-params} : \begin{bmatrix}\text{rc} : \textit{LocProp}\end{bmatrix} \\ \text{phontype} = \text{rc.sit-type.phon} : \textit{Type} \\ \text{phon} : \textit{phontype} \\ \text{cat} : \text{syncat} \\ \text{cont} = \text{rc.sit} : \textit{Rec} \end{bmatrix}$$

2. evocation of the clarification question 'what do you mean by u' accommodated via the clarification context update rule (9).

These two components get reified into a somewhat more general construction *qud-anaph-int-cl*:

[8] This construction, which arguably occurs already at the one word stage (Clark and Bernicot, 2008), is needed for other 'quotative' utterances such as

- A: Bo is coming. B: Who do you mean 'Bo'?
- D: I have a Geordie accident. J: 'accident' that's funny.

its content is identified with *max-qud*, whereas its sole constituent is a phrase of type *utt-ana-ph*:

(18) *qud-anaph-int-cl* =

$$\begin{bmatrix} \text{dgb-params}:\begin{bmatrix}\text{MAX-QUD}:\textit{Question}\end{bmatrix}\\ \text{cont} = \text{max-qud}:\textit{Question}\\ \text{hd-dtr: }\textit{utt-anaph-ph}\end{bmatrix}$$

This is exemplified in (19):

(19) a. Input utterance: A: Did Bo leave?

b. Context assuming the reference of ‘Bo’ cannot be fully resolved: MAX-QUD: ?x.mean(A,x,‘bo’) (*Who$_i$ is A referring to as ‘Bo’*);

c. Content of Bo? = MAX-QUD.question (=Who$_i$ is A referring to as ‘Bo’?)

Scaling up (18) multimodally involves two moves:

1. generalizing phonological segmental parallelism to multimodal parallelism

2. positing a lexical entry for frowns

With respect to the former task we need to generalize the condition *phontype = rc.sit-type.phon* in (17) so that it can apply to gestures, laughs and their combinations with speech. The most obvious generalization would be to require type identity with respect to form on all tiers. However, this will not work because in all cases small but important divergences actually need to apply. In the case of speech the identity is segmental identity, but not with respect to the speech contour (where the reprise is typically LH), whereas in the case of gesture reprises the face is required to involve a frown (in the FACS system Ekman and Friesen, 1978 a combination of A(ction)U(nits) 1 and 4 (Hager, 1985).). Indeed it seems like a repetition which involves total form identity such as repetition of an utterance that is already bearing an LH contour or repeating a frown cannot be understood as clarification requests—they cannot be understood as clarifying the clarification requests (which could be achieved by saying e.g., ‘What do you mean . . . ’):

(20) a. A: Will Bo be selected? B: Bo? (LH) A: # Bo? (LH)

b. A: Can you undertake this mission? B: (frowns) # A: (frowns).

In both cases, then, one needs to leave a channel free, presumably to express interrogative force. Hence, the most straightforward way to achieve this generalized parallelism condition is simply to specify the facial form as identity modulo specification of AUs 1 and 4 and the speech form as identity modulo intonation. An additional question is whether or not multimodal reprises require all channels to be reactivated, as exemplified in (21). We hypothesize that only the complete reprise can communicate a ‘what do you mean’ content, whereas the other reprises are understood as confirmations. However, clearly this requires experimental investigation.

(21) A: I don’t care + shrug. B: You don’t care + shrug + frown?/ You don’t care?/Shrug + frown

For now we will postulate a generalized *utt-ana-ph* type, building on (16)

(22) *mm-utt-ana-ph* =

$$\begin{bmatrix} \text{dgb-params}:\begin{bmatrix}\text{rc}:\textit{MMProp}\end{bmatrix}\\ \text{formtype}:\textit{Type}\\ \text{c1}:\text{quasi-identical(rc.syn,form-type)}\\ \text{syn}:\textit{formtype}\\ \text{cont} = \text{rc.mm-event.e}_{\text{sync}}:\textit{Rec}\end{bmatrix}$$

Why can a frown give rise to a clarification question in this context? We assume, following Ginzburg et al. (2020b), who in turn build on proposals of Scherer (1992); Wierzbicka (2000), that frowns communicate the emergence of a problem in interaction, more specifically involve the frownable giving rise to a question, which can indeed be spoken:[9]

(23)

$$\begin{bmatrix} \text{face}:\texttt{frownbrowtype}\\ \text{dgb-params}:\begin{bmatrix}\text{spkr}:\textit{Ind}\\ \text{addr}:\textit{Ind}\\ \text{t}:\textit{Time}\\ \text{c1}:\text{addressing(spkr,addr,t)}\\ \text{q}:\textit{Question}\\ \text{p}:\textit{Prop}\end{bmatrix}\\ \text{content} = \text{NegRaise(p,q,spkr)}:\textit{Prop}\end{bmatrix}$$

[9] This is backed by entries on *Eyebrow Raise* and, even stronger, *Eyebrow Cock* in the *Nonverbal Body Dictionary*, which are described as signalling surprise, excitement, or general disbelief (`http://bodylanguageproject.com/nonverbal-dictionary/`, accessed April 27, 2021). Eyebrows are also used as question markers in sign languages (e.g. Baker et al., 2016, 132). There different kind of eyebrow movement are correlated with different types of sentences (e.g., yes-no vs. wh; see Freitas et al., 2014, 183, Tab. 3 for a particular clear overview of eyebrow use in Brazilian sign language questioning).

How to package this to attain a construction akin to (18)? There seem to be two options: assume that there is a single reprise fragment construction with certain components that are optional. On this line all instances spoken and purely gestural involve frowning with an utterance anaphora constituent involving phonological or gestural parallelism. The other option is to assume two subtypes of such a construction, a spoken one which involves an LH tone sequence, and a gestural where the interrogative force is driven by the frown. Choosing between these options requires a detailed experimental study, which we leave for future work. For concreteness we offer in (24) a sketch of the former strategy:

(24)
$$\left[\begin{array}{l} \text{(phon : LH)} \\ \text{face : } \texttt{frownbrowtype} \\ \text{dgb-params : } \left[\text{MAX-QUD : } \textit{Question}\right] \\ \text{cont=max-qud :} \textit{Question} \\ \text{hd-dtr: } \textit{mm-utt-anaph-ph} \end{array}\right]$$

A precise semantic analysis along these lines of the discourse functions of gestures in multimodal interaction is attained (for a related work on the so-called *what are you talking about* face see Francis, 2020). Such analyses are needed in order to understand and model tier-crossing coherence in natural language processing, in both artificial and human agents. CRs are a key interactional competence in this respect.

5 Conclusions

Clarifications requests are an important dialogical resource for seeking mutual understanding and driving conversational interactions. However, in face-to-face dialogue CRs extend to the full range of verbal and non-verbal signals. We provided some data illustrating the phenomena at stake and introduced the basic ingredients to develop multimodal clarifications for linguistic theories.

This work fills in particular two explanatory gaps left by current multimodal studies, namely (i) projecting (non-emblematic) gestures to illocutionary acts, and (ii) connecting gestures to the basic dialogue dynamics of grounding and repair.

Acknowledgments

This work is partially supported by a public grant overseen by the French National Research Agency (ANR) as part of the program “Investissements d’Avenir” (reference: ANR-10-LABX-0083). It contributes to the IdEx Université de Paris – ANR-18-IDEX-0001

References

Katya Alahverdzhieva, Alex Lascarides, and Dan Flickinger. 2017. Aligning speech and co-speech gesture in a constraint-based grammar. *Journal of Language Modelling*, 5(3):421–464.

John L. Austin. 1950. Truth. In *Proceedings of the Aristotelian Society. Supplementary*, volume xxiv, pages 111–128. Reprinted in John L. Austin: *Philosophical Papers*. 2. ed. Oxford: Clarendon Press, 1970.

Anne Baker, Beppie van den Bogaerde, Roland Pfau, and Trude Schermer, editors. 2016. *The Linguistics of Sign Languages*. John Benjamins, Amsterdam.

Jon Barwise and John Etchemendy. 1987. *The Liar*. Oxford University Press, New York.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

Eve V. Clark and Josie Bernicot. 2008. Repetition as ratification: how parents and children place information in common ground. *Journal of child language*, 35(2):349–71.

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

Robin Cooper. 2005. Austinian truth, attitudes and type theory. *Research on Language and Computation*, 3(2-3):333–362.

Robin Cooper. 2021. From perception to communication: An analysis of meaning and action using a theory of types with records (TTR). `https://github.com/robincooper/ttl`. Unpublished book draft.

Robin Cooper, Simon Dobnik, Staffan Larsson, and Shalom Lappin. 2015. Probabilistic type theory and natural language semantics. *Linguistic Issues in Language Technology*, 10.

Robin Cooper and Jonathan Ginzburg. 2015. Type theory with records for natural language semantics. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*, 2 edition, chapter 12, pages 375–407. John Wiley & Sons.

Thierry Coquand, Randy Pollack, and Makoto Takeyama. 2003. A logical framework with dependently typed records. In *Typed Lambda Calculi and Applications. Proceedings of the 6th International Conference*, TLCA 2003, pages 105–119.

Jay Earley. 1970. An efficient context-free parsing algorithm. *Communications of the ACM*, 13(2):94–102.

Paul Ekman and Wallace V. Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA.

Tim Fernando. 2007. Observing events and situations in time. *Linguistics and Philosophy*, 30(5):527–550.

Tim Fernando. 2011. Constructing situations and time. *Journal of Philosophical Logic*, 40(3):371–396.

Naomi Francis. 2020. Objecting to discourse moves with gestures. Talk given at *Sinn und Bedeutung 25*. Special session: Gestures and Natural Language Semantics.

Fernando de Almeida Freitas, Sarajane Marques Peres, Clodoaldo Aparecido de Moraes Lima, and Felipe Venâncio Barbosa. 2014. Grammatical facial expressions recognition with machine learning. In *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference*, FLAIRS 2014, pages 180–185.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press.

Jonathan Ginzburg and Robin Cooper. 2004. Clarification, ellipsis, and the nature of contextual updates. *Linguistics and Philosophy*, 27(3):297–366.

Jonathan Ginzburg, Robin Cooper, Julian Hough, and David Schlangen. 2020a. Incrementality and HPSG: Why not? In Anne Abeillé and Olivier Bonami, editors, *Constraint-Based Syntax and Semantics: Papers in Honor of Danièle Godard*. CSLI Publications.

Jonathan Ginzburg, Chiara Mazzocconi, and Ye Tian. 2020b. Laughter as language. *Glossa*, 5(1).

Joseph C. Hager. 1985. A comparison of units for visually measuring facial actions. *Behavior Research Methods, Instruments, & Computers*, 17(4):450–468.

Patrick George Healey, Nicola Plant, Christine Howes, and Mary Lavelle. 2015. When words fail: Collaborative gestures during clarification dialogues. In *2015 AAAI Spring Symposium Series: Turn-Taking and Coordination in Human-Machine Interaction*, pages 23–29.

Patrick G.T. Healey, Matthew Purver, and Christine Howes. 2014. Divergence in dialogue. *PLOS ONE*, 9(6).

Julian Hough, Iwan de Kok, David Schlangen, and Stefan Kopp. 2015. Timing and grounding in motor skill coaching interaction: Consequences for the information state. In *Proceedings of SEMDIAL 2015*, goDIAL 2015, pages 86–94.

Gunnar Johansson. 1973. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211.

Michael Johnston, Philip R. Cohen, David McGee, Sharon L. Oviatt, James A. Pittman, and Ira Smith. 1997. Unification-based multimodal integration. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, pages 281–288, Madrid, Spain. European Chapter Meeting of the ACL, Association for Computational Linguistics.

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

Staffan Larsson. 2020. Extensions are indeterminate if intensions are classifiers. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue – Full Papers*, Virtually at Brandeis, Waltham, New Jersey. SEMDIAL.

Andy Lücking. 2013. *Ikonische Gesten. Grundzüge einer linguistischen Theorie*. De Gruyter, Berlin. Zugl. Diss. Univ. Bielefeld (2011).

Andy Lücking. 2016. Modeling co-verbal gesture perception in type theory with records. In *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, volume 8 of *Annals of Computer Science and Information Systems*, pages 383–392. IEEE.

Andy Lücking and Jonathan Ginzburg. 2020. Towards the score of communication. In *Proceedings of The 24th Workshop on the Semantics and Pragmatics of Dialogue*, SemDial/WatchDial.

Andy Lücking, Kirsten Bergmann, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2010. The Bielefeld speech and gesture alignment corpus (SaGA). In *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, LREC 2010, pages 92–98. 7th International Conference for Language Resources and Evaluation.

Per Martin-Löf. 1984. *Intuitionistic Type Theory*. Studies in Proof Theory. Bibliopolis, Napoli.

Richard Montague. 1974. Pragmatics. In Richmond Thomason, editor, *Formal Philosophy*. Yale UP, New Haven.

Matthew Purver. 2004. *The Theory and Use of Clarification in Dialogue*. Ph.D. thesis, King's College, London.

Matthew Purver. 2006. CLARIE: Handling clarification requests in a dialogue system. *Research on Language & Computation*, 4(2):259–288.

Matthew Purver, Arash Eshghi, and Julian Hough. 2011. Incremental semantic construction in a dialogue system. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS 2011, pages 365–369.

Matthew Purver and Jonathan Ginzburg. 2004. Clarifying noun phrase semantics. *Journal of Semantics*, 21(3):283–339.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. In Jan van Kuppevelt and Ronnie W. Smith, editors, *Current and New Directions in Discourse and Dialogue*, number 22 in Text, Speech and Language Technology book series, pages 235–255. Springer Netherlands, Dordrecht.

Craige Roberts. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, pages 91–136. Reprinted in Semantics and Pragmatics, 2012.

Kepa Rodriguez and David Schlangen. 2004. Form, intonation and function of clarification requests in german task-oriented spoken dialogues. In *Proceedings of Catalog'04, The 8th Workshop on the Semantics and Pragmatics of Dialogue*, Universitat Pompeu Fabra, Barcelona.

Emanuel A. Schegloff. 1984. On some gestures' relation to talk. In J. Maxwell Atkinson and John Heritage, editors, *Structures of Social Action. Studies in Conversational Analysis*, Studies in Emotion and Social Interaction, chapter 12, pages 266–296. Cambridge University Press, Cambridge, MA.

Emanuel A. Schegloff. 2007. *Sequence Organization in Interaction*. Cambridge University Press, Cambridge.

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organisation of repair in conversation. *Language*, 53(2):361–382.

Klaus R. Scherer. 1992. What does facial expression express? In *International Review of Studies of Emotion*, volume 2. John Wiley & Sons.

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 136–143.

Mi-Suk Seo and Irene Koshik. 2010. A conversation analytic study of gestures that engender repair in ESL conversational tutoring. *Journal of Pragmatics*, 42(8):2219–2239.

Robert C. Stalnaker. 1978. Assertion. In P. Cole, editor, *Syntax and Semantics, Volume 9*, pages 315–332. AP, New York.

Michael Tomasello. 1999. *The Cultural Origins of Human Cognition*. Harvard University Press, Cambridge, MA.

Hannes Vilhjálmsson, Nathan Cantelmo, Justine Cassell, Nicolas E. Chafai, Michael Kipp, Stefan Kopp, Maurizio Mancini, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Zsofi Ruttkay, Kristinn R. Thórisson, Herwin van Welbergen, and Rick J. van der Werf. 2007. The behavior markup language: Recent developments and challenges. In *Intelligent Virtual Agents*, pages 99–111, Berlin, Heidelberg. Springer Berlin Heidelberg.

Anna Wierzbicka. 2000. The semantics of human facial expressions. *Pragmatics & cognition*, 8(1):147–183.

Joost Zwarts. 2003. Vectors across spatial domains: From place to size, orientation, shape, and parts. In *Representing Direction in Language and Space*, number 1 in Explorations in Language and Space, chapter 3, pages 39–68. Oxford University Press, Oxford, NY.