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: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

---

1We 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 infelicitious clarifications are marked by ‘#’.
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).

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, 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 $a_1 : T_1$, $a_2 : T_2(a_1)$, $\ldots$, $a_n : T(a_1, a_2, \ldots, a_{n-1})$, then the record on the left in (3) is of the record type on the right in (3):

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

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 ‘$\land$’ is exemplified in (4):

$$(4) \begin{align}
a. & A = \begin{bmatrix} l_1 : T_1 \\
l_2 : T_2(l_1) \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} l_3 : T_3 \end{bmatrix} \\
b. & A \land B = \begin{bmatrix} l_1 : T_1 \\
l_2 : T_2(l_1) \\
l_3 : T_3 \end{bmatrix}
\end{align}$$

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 $e_1 e_2 e_3$ represents a course of events, namely the succession of $e_1$, $e_2$ and $e_3$, in that order. The notation $e_1 e_2 e_3$ is an abbreviation for a time-indexed record:

$$(5) \begin{bmatrix} t_0 = e_1 \\
t_1 = e_2 \\
t_2 = e_3 \end{bmatrix}, \text{where time indices } t_i \text{ are in } \mathbb{N}.$$  

If $e_1 : T_1$, $e_2 : T_2$ and $e_3 : T_3$, then $e_1 e_2 e_3 : T_1^* T_2^* T_3^*$—the type constructor ‘$\ast$’ 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 notionally 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) [spkr : Ind
 addr : Ind
 u-time : Time
 c-utt : addr(spkr, addr, u-time)]

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 game-board (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) DGBT :=
[spkr : Ind
 addr : Ind
 u-time : Time
 c-utt : addressing(spkr, addr, u-time)
 facts : Set(Prop)
 vis-sit : [foa : Ind ∨ Sir : RecType
 pending : List(LocProp)
 moves : List(IllocProp)
 qud : poset(Question)
 mood : Appraisal]

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 (Sir). 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 precondition(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.

\[
\begin{align*}
\text{pre} : & q, \text{Question} \\
\text{effects} : & \text{QUD} = \langle q, \text{pre.QUD} : \text{poset(Question)} \rangle
\end{align*}
\]

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.

\[
\begin{align*}
\text{pre} : & p, \text{Prop} \\
\text{effects} : & \text{QUD} = \langle p?, \text{pre.QUD} : \text{poset(Question)} \rangle
\end{align*}
\]

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).

\[
\begin{align*}
\text{pre} : & \text{QUD} = \langle q, Q \rangle : \text{poset(Question)} \\
\text{effects} : & \text{R} : \text{Question ∨ Prop} \\
& \text{R} : \text{IllocRel} \\
& \text{QUD} = \langle r, \text{IllocProp} \rangle \\
& \text{c1} : \text{Qspecific(r, q)}
\end{align*}
\]

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 \textit{locutionary proposition \( \text{LocProp} \)}. This is a special kind of (Austrian) proposition—records of type \( \text{Rec} \) (Austin, 1950; Barwise and Etchemendy, 1987)\(^4\)—where \( \text{sit} \) is an utterance event and \( \text{sit-type} \) the type of a grammatical sign. This allows \textit{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 \textit{co-propositional}\(^5\) with the clarification question \( \text{CQ}(u0) \) into QUD.”, where ‘\( \text{CQ}(u0) \)’ 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 \textit{co-propositional} with “What did A mean by \( u \)”, can be formulated as in (9): the issue \( q0 \), \textit{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 \textit{CoPropositional} with \( q0 \). Assuming a propositional function view of questions, \textit{CoPropositionality} allows in propositions from the range of \( \text{Range}(q0) \) and questions whose range intersects \( \text{Range}(q0) \). 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).

\( \lambda x.\text{Mean}(A,u0,x) \)
\( ?\text{Mean}(A,u0,b) \) (‘Did you mean Bo?’)
\( \text{Mean}(A,u0,c) \) (‘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 \textit{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 \textit{string types} (Fernando, 2007; Cooper, 2021) are notated in round brackets:

\( \text{partitur} := \left[ \begin{array}{c} \text{\( \varepsilon_{\text{speech}} \) : Phon} \\ \text{\( \varepsilon_{\text{gesture}} \) : Trajectory} \\ \text{\( \varepsilon_{\text{gaze}} \) = vis-sit : \text{RecType}} \\ \text{\( \varepsilon_{\text{head}} \) : headMove} \\ \text{\( \varepsilon_{\text{face}} \) : faceExpr} \end{array} \right] \)

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.

\( \left[ \begin{array}{c} \text{\( \varepsilon_{\text{speech}} \) : Phon} \\ \text{\( \varepsilon_{\text{gesture}} \) : Trajectory} \\ \text{\( \varepsilon_{\text{gaze}} \) = vis-sit : \text{RecType}} \end{array} \right]^+ \)

\( \left[ \begin{array}{c} \text{\( \varepsilon_{\text{speech}} \) : Phon} \\ \text{\( \varepsilon_{\text{gesture}} \) : Trajectory} \\ \text{\( \varepsilon_{\text{gaze}} \) = vis-sit : \text{RecType}} \end{array} \right] \)

\( \left[ \begin{array}{c} \text{\( \varepsilon_{\text{speech}} \) : Phon} \\ \text{\( \varepsilon_{\text{gesture}} \) : Trajectory} \\ \text{\( \varepsilon_{\text{gaze}} \) = vis-sit : \text{RecType}} \end{array} \right]^+ \)

---

\(^4\)On this view, a proposition \( p = \left[ \begin{array}{c} \text{sit} \\ \text{sit-type} \end{array} \right] \) is true iff \( \text{sit} = s \) and \( \text{sit-type} = T \), the situation \( s \) is of the type \( T \).

\(^5\)Here \textit{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 \textit{Behavior Markup Language} (BML; Vilhjálmsdóttir 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\text{gesture} attaches to a phonetically marked \text{affiliate} (Aff; Schegloff, 1984) from speech (S-DTR, tier e\text{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 \textit{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):

\begin{equation}
\text{sg-ensemble} \begin{cases}
\text{phon}=\text{d-tr}.\text{phon} : \text{Phon} \\
\text{cat}=\text{d-tr}.\text{cat} : \text{SynCat} \\
\text{cont}=\text{d-tr}.\text{traj} \land \text{d-tr}.\text{cont.cvm}
\end{cases}
\end{equation}

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')—

\begin{aligned}
\text{path} : \text{line} \\
\text{wrist}=\text{mr} \quad \text{mb}^\text{oc} : \text{Move}
\end{aligned}

a cornered, horseshoe-shaped trajectory \text{U} is displayed. Via a translation procedure from gesture representations onto vector representations, the abstract trajectory in (13) is obtained (Lücking, 2016).

\begin{equation}
\text{traj} = \begin{bmatrix}
\text{pt} : \{ \text{u} \perp \text{v} \perp \text{w} \} \\
\text{sh} : \{ \text{rectangular, open} \}
\end{bmatrix} : \text{Vec}
\end{equation}

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 \textit{u-shaped} (it modifies a nominal, whose content is an individual).

\begin{equation}
\text{cont} = \begin{bmatrix}
\text{cvm} = \begin{bmatrix}
\text{pt} : \{ \text{u} \perp \text{v} \perp \text{w} \} \\
\text{sh} : \{ \text{rectangular, open} \}
\end{bmatrix} : \text{Vec} \\
\text{c}_\text{shape} = \text{shape}(\text{mod.cat.cont.cvm})
\end{bmatrix}
\end{equation}

Since the gesture’s trajectory and the adjective’s CVM value are compatible, both can merge into a \textit{sg-ensemble}.

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 \textit{house [has a RECtangular]} \text{U} shape. The noun phrase \textit{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 \textit{the house has} and the gesture’s preparation phase. Due to this input, a VP rule (e₉) and a gesture integration rule (e₁₀) have been triggered, but are still pending:

\footnote{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.}
multimodal ensembles are combined via co-occur. In terms of locutionary propositions, the and (14)—differs from phrasal constructions usu-
(15)

In contrast to the ‘horizontal’ chart parsing edges (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: 

(17) $utt$-ana-$ph$ =

```
dgb-params : [rc : LocProp] phonotype = rc$\cdot$sit-type$\cdot$phon : Type phon : phonotype cat : syncat cont = rc$\cdot$sit : Rec
```

2. E-voculation of the clarification question ‘what do you mean by $u$’ accommodated via the clarification context update rule (9).

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

```
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:

\[
\text{(18)} \quad \text{qud-anaph-int-cl} = \\
\begin{bmatrix}
\text{dgb-params : } \{\text{MAX-QUD : Question}\} \\
\text{cont} = \text{max-qud : Question} \\
\text{hd-dtr : utt-anaph-ph}
\end{bmatrix}
\]

This is exemplified in (19):

\[
\text{(19)}
\begin{align*}
\text{a. Input utterance: A: Did Bo leave?} \\
\text{b. Context assuming the reference of ‘Bo’ cannot be fully resolved: MAX-QUD: } \forall x.\text{mean}(A,x,\text{‘bo’}) (\text{Who is A referring to as ‘Bo’?}) \\
\text{c. Content of Bo? = MAX-QUD.question (Who is A referring to as ‘Bo’?)}
\end{align*}
\]

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 \text{A(ction)U(nits)}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 . . . ’):

\[
\text{(20)}
\begin{align*}
\text{a. A: Will Bo be selected? B: Bo? (LH)} \\
\text{A: # Bo? (LH)} \\
\text{b. A: Can you undertake this mission? B: (frowns) # A: (frowns).}
\end{align*}
\]

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.

\[
\text{(21)}
\begin{align*}
\text{A: I don’t care + shrug. B: You don’t care + shrug + frown?/ You don’t care?/Shrug + frown}
\end{align*}
\]

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

\[
\text{(22)} \quad \text{mm-utt-anaph} = \\
\begin{bmatrix}
\text{dgb-params : } \{\text{rc : MMProp}\} \\
\text{formtype} : \text{Type} \\
\text{c1 : quasi-identical(rc.syn,form-type)} \\
\text{syn} : \text{formtype} \\
\text{cont} = \text{rc.mm-event.}e\text{sync : Rec}
\end{bmatrix}
\]

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:

\[
\text{(23)}
\begin{bmatrix}
\text{face : frownbrowtype} \\
\text{dgb-params : } \{\text{spkr : Ind} \\
\text{addr} : \text{Ind} \\
\text{t : Time} \\
\text{c1 : addressing(spkr,addr,t)} \\
\text{q : Question} \\
\text{p : Prop} \\
\text{content} = \text{NegRaise(p,q,spkr) : Prop}
\end{bmatrix}
\]

\footnote{This is backed by entries on \text{Eyebrow Raise} and, even stronger, \text{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) \[
\begin{array}{l}
\text{phon} : \text{LH} \\
\text{face} : \text{frownbrowtype} \\
\text{dgb-params} : [\text{MAX-QUD} : \text{Question}] \\
\text{cont-max-qud} : \text{Question} \\
\text{hd-dtr} : \text{mm-att-anaph-ph}
\end{array}
\]

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


