

Towards the score of communication

Andy Lücking

Université de Paris

Laboratoire de Linguistique Formelle

Goethe-Universität Frankfurt

Text Technology Lab

luecking@em.uni-frankfurt.de

Jonathan Ginzburg

Université de Paris

Laboratoire de Linguistique Formelle

Laboratoire d'Excellence LabEx-EFL

Institut Universitaire de France

yonatan.ginzburg@u-paris.fr

Abstract

The exchange of verbal and non-verbal communication signals in face-to-face dialogue is complexly organised in several ways: each contribution is produced and processed incrementally, contributions may be consecutive (e.g. question-answer pairs) or overlapping (e.g. backchannelling), and the contributions themselves may be multimodal. Contributions nonetheless exhibit pairwise utterance coherence, and in two respects: across tiers and across discourse co-texts. For these reasons, we propose to distribute dialogue agents across different tiers and to ‘incrementalize’ the sequential notion of turns according to the model of music-inspired communication scores.

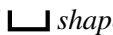
1 Motivation

It is a truism that natural language communication is multimodal, or as we will also say, proceeds on different *tiers*.¹ What is meant by this slogan is that dialogue agents in addition to speech exchange a great variety of non-verbal communication means like manual gestures, facial expressions, gaze, laughter or suprasegmental signals (the latter two are vocal but not verbal). The chief significance of nonverbal communication means usually resides in a *dialogue-oriented meaning* (Ginzburg and Poesio, 2016). For instance, backchannelling signals such as nodding or vocalisations such as ‘mhm’ influence the development of discourse (Bavelas et al., 2000). With *communication scores*, we aim to address these aspects of the fabric of communication by distributing dialogue agents across several communication tiers and allow tiers

¹We conceive a tier to be a layer of communication in the semiotic triad spanned by a transfer *medium*, an interpretive *code*, and a receiving sense *modality*. For instance, the phonetic events encoding natural language expressions, which are produced with the articulatory organ and spread through air and are perceived by the ears, constitute a tier, namely the vocal tier.

to be shared by different participants (i.e., giving up a strict sequential notion of turns). We hypothesise that the resulting (conceptual and formal) intricacies are regimented by a fundamental dialogical constraint of coherence.

2 Some data

The most striking feature of multimodal discourse is the *binding problem* (Feldman, 2012). Interpretation is often guided by the heuristics ‘if multiple signs occur simultaneously, take them as one’ (Enfield, 2009, 9). For instance, a drawing gesture co-occurring with a shape description is understood as expressing a single *idea unit* (McNeill, 1992): *the house [has a REctangular]  shape* (here the gesture temporally overlaps with the underlined portion of speech, the stretch of the gesture’s stroke is indicated by square brackets, the trajectory of the drawing gesture is sketched after the closing stroke bracket; capital letters indicate main stress). However, since simultaneity is not the only temporal pattern of multimodal communication, the simple heuristics has to be qualified. A point in case is post-stroke holds, as argued by Rieser (2015) – we will come back to such issues shortly.

A multimodal turn grabbing is exemplified in (1), which stems from the (German) SaGA dialogue V4, at 8:39 (Lücking et al., 2010). Here, the route-giver R wants to continue her route-giving and produced the definite article *die* ‘the [fem.]’. At this point, the follower F cuts in by index finger raising and exclaiming *Moment* ‘wait’ twice.

(1) R: [*die*]

F: [*moment*] *moment*



The turn grab was multimodal, but a rectangular gesture would be inappropriate in this case, a point-

ing or a ‘stop!’ gesture seems to be called for, if any. This might indicate a multimodal interface of certain discourse-related expressions to interactive gestures as has been argued for wrt. demonstrative expressions and deixis (Frege, 1918) and spatial expressions and iconic (Schegloff, 1984) gestures, among others.

Boarding the turn of the speaker doesn’t necessarily lead to a turn grab, rather, interlocutors can also produce *joint utterances* (Poncin and Rieser, 2006), undermining a strict notion of speaker and addressee even more than constant backchanneling.

With respect to the latter, an interesting asymmetry can be observed (one example is given in (35) below): it is possible that A is speaking and B is agreeing or disagreeing at the same time by nodding or head shaking, respectively. It is also possible that A is speaking and agreeing (with his-/herself) at the same time: namely when A speaks and nods. But it only seems to be possible that A is speaking and *disagreeing* (shaking head) at the same time under special circumstances, for instance, if A’s utterance contains a negative particle (*not*, *n’t*, *no*) or is part of a *hostile* (Gregoromichelaki et al., 2011) move (i.e., A is not committed to the content of the utterance). We take up some examples in Sec. 5.

We hypothesise that one can make sense out of this by a (multimodal extension of a) fundamental dialogical constraint, namely (*pairwise*) *utterance coherence* (Ginzburg, 2012, Sec. 6.7.1). Co-occurring signals as well as consecutive utterances constitute a kind of adjacency pair in the sense that they can be embedded in a dialogue gameboard in a *relevant* way. This embedding has two aspects: (i) there is a grammar rule that licenses the combination of the signals, and (ii) the resulting dialogue gameboard can be connected to the actual one by means of one or more conversational rules. In order to investigate the constraints that apply to largely simultaneous and multimodal interactions, one needs to ‘compartmentalize’ agents into articulators, as is outlined in the following.

3 Formal background

Communication scores (Sec. 4) rest on a couple of previous works which are briefly introduced subsequently.

3.1 TTR

Communication scores are formulated within *Type Theory with Records* (TTR, Cooper, 2005; Cooper and Ginzburg, 2015). 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 $a_1 : T_1$, $a_2 : T_2(a_1)$, ..., $a_n : T(a_1, a_2, \dots, a_{n-1})$, then the record on the left in (2) is of the record type on the right in (2):

$$(2) \quad \left[\begin{array}{c} l_1 = a_1 \\ \vdots \\ l_n = a_n \end{array} \right] : \left[\begin{array}{c} l_1 : T_1 \\ \vdots \\ l_n : T(l_1, l_2, l_{n-1}) \end{array} \right]$$

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 ‘ \wedge ’ is exemplified in (3):

$$(3) \quad \begin{array}{l} \text{a. } A = \left[\begin{array}{c} l_1 : T_1 \\ l_2 : T_2(l_1) \end{array} \right] \text{ and } B = \left[l_3 : T_3 \right] \\ \text{b. } A \wedge B = \left[\begin{array}{c} l_1 : T_1 \\ l_2 : T_2(l_1) \\ l_3 : T_3 \end{array} \right] \end{array}$$

For more (on) TTR see Cooper (2012).

3.2 Strings

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:

$$(4) \quad \left[\begin{array}{c} t_0 = e_1 \\ t_1 = e_2 \\ t_3 = e_2 \end{array} \right], \text{ where time indices } t_i \text{ are in } \mathbb{N}.$$

If $e_1 : T_1$, $e_2 : T_2$ and $e_3 : T_3$, then $e_1e_2e_3 : T_1 \hat{\wedge} T_2 \hat{\wedge} T_3$ – the type constructor ‘ $\hat{\wedge}$ ’ 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. For instance, the first move of a chess player playing a Sicilian opening is represented in (5):

(5)

$$\left[\begin{array}{l} x : Ind \\ c_x : chess-player(x) \\ y : Ind \\ c_y : pawn(y) \\ l_1 : Loc \\ c_{l_1} : field-c7(l_1) \\ l_2 : Loc \\ c_{l_2} : field-c5(l_2) \\ c_{on} : on(y, l_1) \\ e : ([e : grab(x, y)] \wedge [e : move-to(x, y, l_2)] \wedge [e : drop(x, y)]) \end{array} \right]$$

3.3 Incrementality

The feature-structure set-up of record types can be used in order to ‘import’ constraint-based grammars such as a *Head-driven Phrase Structure Grammar* (HPSG; Pollard and Sag, 1994). An emulation of HPSG within TTR is HPSG_{TTR} (Cooper, 2008; Ginzburg, 2012). Modelling the derivation of a constituent structure within a chart parser, an incremental version of HPSG_{TTR} can be obtained (Ginzburg et al., 2020). Incremental processing rests on perceptual classification: an acoustic event on the speech tier is classified as the realisation of the phonological part of a sign, that is, as an instance of a *lexical resource*. An example appropriate to Beethoven’s anniversary year 2020 is given in (6), which draws on the NP format of Lücking and Ginzburg (2019):

(6) a. Lex(‘Beethoven’, NP)

$$\text{b. } \left[\begin{array}{l} \left[\begin{array}{l} e : beethoven \\ \text{s-event} : \left[\begin{array}{l} \text{spkr} : Ind \\ \text{addr} : Ind \\ c_{sp} : addressing(\text{spkr}, \text{addr}, e) \end{array} \right] \end{array} \right] \\ \text{phon} : /Beethoven/ \\ \text{syn} : \left[\begin{array}{l} \text{cat=np} : Cat \\ \text{dtrs} = \langle \rangle : list(Sign) \end{array} \right] \\ \text{q-params} : \left[\begin{array}{l} \text{refind} : Ind \\ c_{nm} : named(\text{refind}, 'Beethoven') \end{array} \right] \\ \text{cont=q-params.refind} : Ind \end{array} \right]$$

We can proceed from a speech event to the associated sign information by dint of the phonological classification constraint in (7) (slightly modified from Ginzburg et al., 2020):

(7) If $\text{Lex}(T, C)$ is a lexical resource, then for any speech event u such that $u : T$, it is licensed to merge resource and speech event: $\text{Lex}(T, C) \wedge [\text{s-event} : [e=u : T]]$.

An example of an outcome of this sign classification is shown in (6b).

Since ‘Beethoven’ is a proper name (a full NP), it raises the expectation that a VP will follow. This expectation is backed by grammar which supplies a syntactic resource in form of the rule ‘ $S \rightarrow NP VP$ ’. String types (cf. Sec. 3.2) tie up these things as follows. The acoustic speech event ‘Beethoven’ gives rise to the following initialisation of a chart:

$$(8) \left[\begin{array}{l} e_1 = beethoven : Phon \\ e : ([e_1 : start(e_1)] \wedge [e_1 : end(e_1)]) \end{array} \right]$$

The start-end string in (8) corresponds to an edge in chart parsing (Earley, 1970), see also Fig. 1. Applying lexical resource classification adds sign information:

(9)

$$\left[\begin{array}{l} e_1 = beethoven : Phon \\ e_2 : \text{Lex}('Beethoven', NP) \wedge [\text{s-ev} : [e=e_1 : /Beethoven/]] \\ e : ([e_1 : start(e_1)] \wedge [e_1 : end(e_1)] \wedge [e_2 : start(e_2)] \wedge [e_2 : end(e_2)]) \end{array} \right]$$

Syntactic resources allow us to project hypotheses of possible continuations (what is required (req) for a rule to apply), given what has been found (fnd). Such a move has been developed for incremental processing in *Dynamic Syntax* (Gregoromichelaki et al., 2011). However, the closest precursor is Poesio and Traum (1997), where micro-conversational events generate (in the sense of Goldman, 1970) conversational moves. We adopt such an approach here to the HPSG_{TTR} format, following Ginzburg (2012, Sec. 8) and Ginzburg et al. (2018):

$$(10) \left[\begin{array}{l} e_1 = beethoven : Phon \\ e_2 : \text{Lex}('Beethoven', NP) \wedge [\text{s-ev} : [e=e_1 : /Beethoven/]] \\ e_3 : \left(\begin{array}{l} \text{rule} = S \rightarrow NP VP : NP \wedge VP \text{ fnd} = e_2 : Sign \\ \text{req} = VP : Sign \\ e : required(\text{req}, \text{rule}) \end{array} \right) \\ e : ([e_1 : start(e_1)] \wedge [e_1 : end(e_1)] \wedge [e_2 : start(e_2)] \wedge [e_2 : end(e_2)] \wedge [e_3 : start(e_3)] \wedge [e_3 : end(e_3)]) \end{array} \right]$$

If the string chart processes an input of the category marked as required, a sentential parse will be achieved:

(11)

$$\begin{array}{l}
e_1 = \text{beethoven} : \text{Phon} \\
e_2 : \text{Lex}(\text{'Beethoven'}, \text{NP}) \wedge \left[\text{s-ev} : \left[e=e_1 : / \text{Beethoven}/ \right] \right] \\
e_3 : \left(\begin{array}{l} \left[\begin{array}{l} \text{rule} = \text{S} \rightarrow \text{NP VP} : \text{NP} \wedge \text{VP} \\ \text{fnd} = e_2 : \text{Sign} \\ \text{req} = \text{VP} : \text{Sign} \\ e : \text{required}(\text{req}, \text{rule}) \end{array} \right] \wedge \left[\begin{array}{l} \text{fnd} = e_5 : \text{Sign} \\ \text{req} = \epsilon : \text{Sign} \\ e : \text{complete}(\text{rule}) \end{array} \right] \end{array} \right) \\
e_4 = \text{rocks} : \text{Phon} \\
e_5 : \text{Lex}(\text{'rock'}, \text{VP}) \wedge \left[\text{s-event} : \left[e=e_4 : / \text{rocks}/ \right] \right] \\
e : \left(\begin{array}{l} \left[\begin{array}{l} e_1 : \text{start}(e_1) \\ e_2 : \text{start}(e_2) \end{array} \right] \wedge \left[\begin{array}{l} e_1 : \text{end}(e_1) \\ e_2 : \text{end}(e_2) \\ e_3 : \text{start}(e_3) \\ e_4 : \text{start}(e_4) \\ e_5 : \text{start}(e_5) \end{array} \right] \wedge \left[\begin{array}{l} e_3 : \text{end}(e_3) \\ e_4 : \text{end}(e_4) \\ e_5 : \text{end}(e_5) \end{array} \right] \end{array} \right)
\end{array}$$

Semantic composition proceeds in parallel to chart construction as specified in the lexical resources and the syntactic rules. The result of a chart parse is a HSPG_{TTR} sign defined by the (completed) rule that covers the whole speech event. The chart in (11), for instance, derives the sentential (i.e. $\text{cat}=\text{vp}$ and $\text{dtrs}=\langle \rangle$) sign given in (12), where “cont” labels a structure of type *Prop(osition)*, the type of an Austinian (Austin, 1950) proposition, pairing a situation (record) and a situation type (record type):

(12)

$$\begin{array}{l}
\text{phon} : / \text{Beethoven}/ \wedge / \text{rocks}/ \\
\text{syn} : \left[\begin{array}{l} \text{cat} = \text{vp} : \text{Cat} \\ \text{dtrs} = \langle \rangle : \text{list}(\text{Sign}) \end{array} \right] \\
\text{q-params} : \left[\begin{array}{l} \text{refind} : \text{Ind} \\ c_{\text{nm}} : \text{named}(\text{refind}, \text{'Beethoven'}) \end{array} \right] \\
\text{cont} = \left[\begin{array}{l} \text{sit} = s_0 : \text{Rec} \\ \text{sit-type} = [\text{nucl} : \text{run}(\text{q-params.refind})] : \text{RecType} \end{array} \right]
\end{array}$$

We will refer to the maximal sign processed by a chart for a speech event e as $\text{Sign}(e)$. Since e can be either a simple acoustic event (corresponding to a lexical element), or a string of acoustic events, two cases have to be distinguished:

(13) $\text{Sign}(e)$

$$\begin{array}{l}
\text{a. } \left[\begin{array}{l} \text{sit} = \text{s-event} : [e : \text{Phon}] \\ \text{sit-type} : \text{Lex}(T, \text{Sign}) \wedge \text{sit.e} : T \end{array} \right] \\
\text{b. } \left[\begin{array}{l} \text{sit} = \text{s-event} : [e : (\text{Phon} \wedge \text{Phon})^+] \\ \text{sit-type} : \text{Sign} \wedge \left[\text{syn} : [\text{dtrs} = e : \text{String}(\text{Sign})] \right] \end{array} \right]
\end{array}$$

Furthermore, given the mechanism of rule projection also an anticipatory version of $\text{Sign}(e)$ can be defined, $\text{ProjSign}(e)$ (‘projected sign’). The

processing and production of natural language sentences in interaction is highly anticipatory, as is evinced by timing relations from turn-taking (Levinson and Torreira, 2015).

Given a non-sentential utterance u processed so far, and the series of “req” rules σ leading from the syntactic parse of u to S , then

(14) $\text{ProjSign}(e)$ is

$$\left[\begin{array}{l} \text{sit} = \text{s-event} : [u : \text{Phon}] \\ \text{sit-type} : \text{Sign} \wedge \left[\text{syn} : [\text{dtrs} = \sigma : \text{Sign}(\text{Sign})] \right] \end{array} \right]$$

$\text{ProjSign}(e)$ can further be refined in terms of the number, depth and probabilities of the syntax rules involved. The latter can be derived from *data-oriented parsing* (Bod and Scha, 1996) and implemented in a probabilistic version of TTR (Cooper et al., 2015). We leave such refinements and their empirical testing to future work.

We assume that $\text{ProjSign}(e)$ is part of a dialogue agent’s private share of its information state. $\text{ProjSign}(e)$ is constantly compared to $\text{Sign}(e)$ from the public information state via monitoring processes. $\text{ProjSign}(e)$ can lead to turn-overlapping backchannelling and joint utterances, among others; or to anticipatory errors, in case of which a special kind of correction will occur (*mhm, mhm ... Oh wait! No! That’s not what I expected*). So far, $\text{ProjSign}(e)$ is an ‘experimental feature’, but one that is needed not least for accounting for some puzzling, short timing relations and joint utterances. In any case, $\text{Sign}(e)$ and $\text{ProjSign}(e)$ connects incremental charts to information states and locutionary propositions utilized in KoS, where it can also lead to ‘turn projection’, as argued by Ginzburg et al. (2018) with regard to forward-looking disfluencies.

3.4 KoS

Within language use, the signs of a natural language are part of ‘mechanisms of interaction’ (Kempson et al., 2016). In order to analyze natural language interactions, we make use of *dialogue gameboards* (Ginzburg, 1994). A dialogue game board (DGB) is an information-state based structure for modeling *communicative interactions*. The DGB from KoS tracks the interlocutors (*spkr* and *addr* fields), a record of the dialogue history (*Moves*), dialogue moves that are in the process of grounding (*Pending*), the question(s) currently under discussion (*QUD*), the assumptions shared among the interlocutors (*Facts*) and the dialogue participant’s

view of the visual situation and attended entities (*VisualSit*). The TTR representation of a DGB is given in (15), where *LocProp* is the type of a *locutionary proposition* (see (16) below) and *poset* abbreviates ‘partially ordered set’.

$$(15) \left[\begin{array}{ll} \text{spkr} & : \text{Ind} \\ \text{addr} & : \text{Ind} \\ \text{utt-time} & : \text{Time} \\ \text{c-utt} & : \text{addressing}(\text{spkr}, \text{addr}, \text{utt-time}) \\ \text{facts} & : \text{set}(\text{Prop}) \\ \text{visualsit} & : \text{RecType} \\ \text{pending} & : \text{list}(\text{LocProp}) \\ \text{moves} & : \text{list}(\text{LocProp}) \\ \text{qud} & : \text{poset}(\text{Question}) \end{array} \right]$$

A special kind of propositions are *locutionary propositions* (*LocProp*, Ginzburg, 2012, 172):

$$(16) \text{LocProp} =_{\text{def}} \left[\begin{array}{ll} \text{sit} & : \text{Sign} \\ \text{sit-type} & : \text{RecType} \end{array} \right]$$

Locutionary propositions are the link between dialogue gameboards and incremental processing. In terms of chart parsing (Sec. 3.3), a *LocProp* is the classification of an acoustic speech event by means of a sign type. That is, *LocProp* corresponds to *Sign(e)* from (13). Locutionary propositions are sign objects required to explicate clarification potential and grounding (Ginzburg, 2012).

Dialogue dynamics is regimented by *conversational rules*. A conversational rule is specified in terms of its preconditions (preconds) and its effects. A simple example is free speech, where, given an empty QUD list, any dialogue participant can make a contribution, that is *turn(holder)Underspec(ified)* (Ginzburg, 2012):

$$(17) \left[\begin{array}{ll} \text{preconds} & : \left[\text{qud} = \langle \rangle : \text{poset}(\text{Question}) \right] \\ \text{effects} & : \left[\begin{array}{l} \text{turnUnderspec} \wedge \\ \left[\begin{array}{l} \text{a} : \text{Prop} \\ \text{R} : \text{IllocRel} \\ \text{moves.latest} = \text{R}(\text{spkr}, \text{addr}, \text{a}) : \text{IllocProp} \end{array} \right] \end{array} \right] \end{array} \right]$$

We will make use of free speech in the analysis of (35) in Fig. 2 below.

3.5 Multimodal charts

In order to process input on several tiers, multimodal chart parsing has been devised as an extension of unification-based grammar parsing (Johnston, 1998). A graphical illustration is given in Fig. 1, where speech and gesture input can be parsed into several *multicharts*, namely $\{(s, 0, 1), (g, 3, 4)\}$, $\{(s, 1, 2), (g, 3, 4)\}$, or $\{(s, 0, 2), (g, 3, 4)\}$.

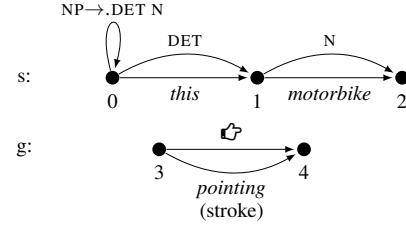


Figure 1: Multimodal chart parser

Within HPSG grammars, speech-gesture integration rules have been defined based on multichart parsing (Alahverdzhieva et al., 2017; Lücking, 2016).² In these approaches, the range of possible multicharts is further constrained by temporal and prosodic information. The original temporal constraint says that ‘the time of the speech [...] must either overlap with or start within 4 seconds of the time of the gesture’ (Johnston, 1998, 627). Since such temporal constraints are always a bit arbitrary, we propose a communication-based constraint instead. To this end we make use of the (still time-based) account of speech-gesture of Lücking (2013), respectively its HPSG_{TTR} reformulation (Lücking, 2016), since this account follows a perceptual classification approach as already used for speech interpretation (cf. Subsec. 3.3). Speech-gesture integration on this account is modelled in terms of a *speech-gesture ensemble* (Kendon, 2004), where a gesture (G-DTR) attaches to a phonetically marked *affiliate* (AFF; Schegloff, 1984) from speech (S-DTR), which is required to exhibit a feature called “CVM”.

$$(18) \begin{array}{c} \text{sg-ensemble} \\ \left[\begin{array}{l} \text{phon} = \text{s-dtr.phon} : \text{Phon} \\ \text{cat} = \text{s-dtr.cat} : \text{SynCat} \\ \text{cont} = \text{g-dtr.traj} \wedge \text{s-dtr.cont.cvm} \end{array} \right] \\ \swarrow \quad \searrow \\ \begin{array}{c} \text{s-dtr} \\ \left[\begin{array}{l} \text{phon.accent} : \text{Marked} \\ \text{cat} : \text{SynCat} \\ \text{cont} : \text{SemObj} \end{array} \right] \end{array} \quad \begin{array}{c} \text{g-dtr} \\ \left[\begin{array}{l} \text{aff} = \text{s-dtr} : \text{Sign} \\ \text{traj} : \text{Vec} \end{array} \right] \end{array} \end{array}$$

The underlying rationale of (18) is that a gesture movement 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). This gesture vector (hence


²For an approach to coverbal gesture integration based on rhetorical relations see Lascarides and Stone (2009). However, since this account presupposes a grammatical affiliation, it probably rests on a (variant of an) approach sketched here.

type *gesture-vec*) *exemplifies* a predicate from the restrictor list of the affiliated verbal expression by unifying into its *conceptual vector meaning* (CVM), which in turn is an abstract, vector-based representation of shapes, movements, orientations, or object axis, spelled out within the vector space algebra of Zwarts (2003). A translation procedure from gesture representation onto vector representations (and a HPSG_{TTR} version) is given in Lücking (2016). But what is a gesture representation? Drawing on work in gesture annotation, gestures are represented in terms of their kinematic features, giving rise to a ‘phonetic’ gesture representation. Within the spatial reference system of the *gesture space* (McNeill, 1992), the movements of the wrists and orientations of palms and backs of each hand are coded. Additional information concern the hand-shape, spatio-temporal extents, and, in case of a bimanual gestures, the relation between both hands (cf. Martell et al., 2002; Lücking et al., 2010).

Gestures are functionally distinguished in terms of the *representation technique* they perform (Müller, 1998; Streeck, 2008). A representation technique is a culture-based practise such as *drawing*, *sculpturing*, *modelling*, or *placing*. In analogy to lexical resources from incremental speech processing (Sec. 3.3), we construct *gesture resources* from kinematic manual gesture representations. An example resource for a drawing gesture is given in (19a), which expands to the structure in (19b). Note that the gesture resource in (19a) is *not* an emblem (i.e., a lexicalized form–meaning pair like *thumbs-up*). A drawing gesture does not have a fixed form side. This is more obvious in the CONT feature in (19b), where the movement features of the gesture token gets ‘vectorized’ – the iconic aspect of a drawing gesture. The content of a drawing gesture then is that the vector sequence described by the gesture’s wrist trajectory (Vec(carrier.wrist)) depicts the SHAPE attribute of the affiliated expression.

$$(19) \text{ a. } \text{Gest} \left(\begin{array}{l} \text{hand} : \{ \text{left}, \text{right} \} \\ \text{hs} : h\text{Shape} \\ \text{carrier} : \begin{array}{l} \text{wrist} : \text{Move} \\ \text{palm} : \text{Dir} \\ \text{boh} : \text{Dir} \\ \text{mov} : \text{Conc} \end{array} \\ \text{sync} : \begin{array}{l} \text{s-loc} : g\text{Space} \\ \text{e-loc} : g\text{Space} \end{array} \\ \text{rel} : g\text{Rel} \end{array} \right), \text{drawing})$$

$$\text{b. } \left[\begin{array}{l} \text{g-event} : \begin{array}{l} \text{e} : \begin{array}{l} \text{hand} : \{ \text{left}, \text{right} \} \\ \text{hs} : h\text{Shape} \\ \text{carrier} : \begin{array}{l} \text{wrist} : \text{Move} \\ \text{palm} : \text{Dir} \\ \text{boh} : \text{Dir} \\ \text{mov} : \text{Conc} \end{array} \\ \text{sync} : \begin{array}{l} \text{s-loc} : g\text{Space} \\ \text{e-loc} : g\text{Space} \end{array} \\ \text{rel} : g\text{Rel} \end{array} \\ \text{spkr} : \text{Ind} \\ \text{addr} : \text{Ind} \\ \text{c}_{\text{sp}} : \text{addressing}(\text{spkr}, \text{addr}, \text{e}) \end{array} \\ \text{mode} : \begin{array}{l} \text{act} = \text{draw} : g\text{Mode} \\ \text{cont} = \text{Vec}(\text{carrier.wrist}) \wedge \\ \text{aff.cont.cvm.shape} : \text{Exemplification} \end{array} \end{array} \right]$$





For example, moving the wrist rightwards, back (i.e., towards the body of the gesturer), and leftwards in a rectangular manner (‘line’) – $\left[\begin{array}{l} \text{path} : \text{line} \\ \text{wrist} = \text{mr} \wedge \text{mb} \wedge \text{ml} : \text{Move} \end{array} \right]$ – a U-shaped trajectory ‘’ is displayed. Note that an iconic gesture does not receive a direct translation into semantic predicates of the underlying formal semantics framework; it can *exemplify* such predicates, though. One could speculate that this difference underlies the non-at-issue status of most co-speech gestures observed by Ebert (2014).

A standard gesture movement is carried out in three *phases*, namely a *preparation phase* (where the hand is moved out of a rest position into gesture space), a *stroke* (where the gesture performs its actual meaningful trajectory), and a *retraction phase* (where the hand is moved from the stroke back into a rest position). The stroke may be ‘frozen’, amounting to a *post-stroke hold*. Such holds are particularly difficult for semantic modelling since they preserve the stroke’s meaning for later uptake while speech already progresses (Rieser, 2015). However, part of such difficulties is a time-based notion of affiliation. We propose to replace Johnston’s (1998) temporal constraint on speech-gesture ensembles by a relative one along the following lines:

(20) Speech-gesture integration within multicharts

- a. *Stroke*: a gesture’s stroke attaches to the closest verbal affiliate candidate;
- b. *Post-stroke hold*: during a post-stroke hold, a gesture attaches to any closest affiliate candidate.

An affiliate candidate of course has to fulfil phonetic and CVM requirements in addition. Note that an affiliate which constitutes an ensemble with the stroke phase of a gesture is not available to the hold phase any more. Note further that the relative constraint has a testable consequence: if a possible affiliate occurs in between a gesture and its actually ‘intended’ affiliate, then the possible one becomes the actual one instead – the *no intervening communication event hypothesis*. By this measures any of the following patterns are captured (assuming that the shape-related predicate *rectangular* provides the CVM interface):


- (21) a. A:  *the house has a RECTangular shape*
(gesture before speech)
- b. A: *the house has a RECTangular shape*

(speech before gesture)
- c. A: *the house has a RECTangular shape*

(speech overlaps gesture)
- d. A: *the house has a RECTangular shape*
B: 
(agent crossing speech–gesture integration)

However, the rectangular speech–gesture ensemble can be broken if another affiliate candidate intervenes, as in (22):


- (22) *the house has a RECT- no circular shape*


Since (22) involves a self-correction, a proper analysis would have much more to tell; however, corrections are not the topic of this paper and the examples hopefully already suffice to illustrate the relative patterning of speech and gesture integration within multimodal charts.

Holds (whose duration is indicated by asterisks) can affiliate to candidates not yet occupied by the stroke phase, which accounts, for instance, for verbal repetitions. In (23), the hold can be attached to the repeated attribution ‘rectangular shape’:

- (23) *the house has a RECTangular*
*****
shape – rectangular shape

Given the rule of thumb that the preparation phase precedes the verbal affiliate of the stroke

(McNeill, 1992), observing a preparation in multimodal parsing provides information for projecting an hypothesis of a multimodal integration rule – for instance, expecting a *sg-ensemble* headed by an NP in speech instead of a speech-only NP rule. Take, for instance, the chart parse of the example from Sec. 1: *the house [has a RECTangular]  shape*. The string chart in (24) 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:

(24)

e_1	= the : Phon
e_2	: Lex(‘the’, DET) \wedge [s-event : [e= e_1 : /the/]]
e_3	: ([rule=NP→DET N : DET \wedge N] \wedge [fnd= e_2 : Sign])
e_4	= house : Phon
e_5	: Lex(‘house’, N) \wedge [s-event : [e= e_4 : /house/]]
e_6	= prep : Phase
e_7	= has : Phon
e_8	: Lex(‘have’, V) \wedge [s-event : [e= e_7 : /has/]]
e_9	: ([rule=VP→V NP fnd= e_8 : Sign req=NP : Sign e : required(req,rule)])
e_{10}	: ([rule=sg-ensemble→X[accent,cvm] stroke fnd= e_6 : Phase req1=stroke : Phase req2=X[accent,cvm] : Sign e : required(req1,req2,rule)])
e	: ([e ₁ : start(e_1) e ₂ : start(e_2)] \wedge [e ₁ : end(e_1) e ₂ : end(e_2) e ₃ : start(e_3) e ₄ : start(e_4) e ₅ : start(e_5) e ₆ : start(e_6)] \wedge [e ₃ : end(e_3) e ₄ : end(e_4) e ₅ : end(e_5) e ₆ : end(e_6) e ₇ : start(e_7) e ₈ : start(e_8) e ₉ : start(e_9) e ₁₀ : start(e_{10})] \wedge [e ₇ : end(e_7) e ₈ : end(e_8)])

In the next steps, the indefinite article is processed and triggers an NP rule (we just consider a single NP hypothesis in e_{15} below). The gesture stroke is detected (e_{11}). The adjective *rectangular* is processed (e_{16} , e_{17}), which carries accent and CVM information required to complete the gesture integration rule (e_{10}). The string chart after these steps is as follows:

(25)

$$\begin{aligned}
e_1 &= \text{the} : \text{Phon} \\
e_2 &: \text{Lex}(\text{'the'}, \text{DET}) \wedge \left[\text{s-event} : \left[e=e_1 : / \text{the}/ \right] \right] \\
e_3 &: \left(\begin{array}{l} \text{rule}=\text{NP} \rightarrow \text{DET N} : \text{DET} \wedge \text{N} \\ \text{fnd}=e_2 : \text{Sign} \end{array} \right) \wedge \left[\text{fnd}=e_5 : \text{Sign} \right] \\
e_4 &= \text{house} : \text{Phon} \\
e_5 &: \text{Lex}(\text{'house'}, \text{N}) \wedge \left[\text{s-event} : \left[e=e_4 : / \text{house}/ \right] \right] \\
e_6 &= \text{prep} : \text{Phase} \\
e_7 &= \text{has} : \text{Phon} \\
e_8 &: \text{Lex}(\text{'have'}, \text{V}) \wedge \left[\text{s-event} : \left[e=e_7 : / \text{has}/ \right] \right] \\
e_9 &: \left(\begin{array}{l} \text{rule}=\text{VP} \rightarrow \text{V NP} \\ \text{fnd}=e_8 : \text{Sign} \\ \text{req}=\text{NP} : \text{Sign} \\ e : \text{required}(\text{req}, \text{rule}) \end{array} \right) \\
e_{10} &: \left(\begin{array}{l} \text{rule}=\text{sg-ensemble} \rightarrow \text{X}[\text{accent}, \text{cvm}] \text{stroke} \\ \text{fnd}=e_6 : \text{Phase} \\ \text{req1}=\text{stroke} : \text{Phase} \\ \text{req2}=\text{X}[\text{accent}, \text{cvm}] : \text{Sign} \\ e : \text{required}(\text{req1}, \text{req2}, \text{rule}) \end{array} \right) \\
&\quad \wedge \left(\begin{array}{l} \text{fnd1}=e_{11} : \text{Phase} \\ \text{fnd2}=e_{17} : \text{Sign} \end{array} \right) \\
e_{11} &= \left[\text{carrier} = \left[\begin{array}{l} \text{path} : \text{line} \\ \text{wrist}=\text{mr} \wedge \text{mb} \wedge \text{ml} : \text{Move} \end{array} \right] \right] : \text{Stroke} \\
e_{12} &: \text{Gest}(\text{Vec}(\text{carrier}), \text{draw}) \wedge \left[\text{g-ev} : \left[e=e_{11} : \text{L} \right] \right] \\
e_{13} &= \text{a} : \text{Phon} \\
e_{14} &: \text{Lex}(\text{'a'}, \text{DET}) \wedge \left[\text{s-event} : \left[e=e_{13} : / \text{a}/ \right] \right] \\
e_{15} &: \left(\begin{array}{l} \text{rule}=\text{NP} \rightarrow \text{DET A N} : \text{DET} \wedge \text{A} \wedge \text{N} \\ \text{fnd}=e_{14} \end{array} \right) \\
&\quad \wedge \left(\begin{array}{l} \text{fnd}=e_{17} : \text{Sign} \\ \text{req}=\text{N} : \text{Sign} \\ e : \text{required}(\text{req}, \text{rule}) \end{array} \right) \\
e_{16} &= \text{rectangular} : \text{Sign} \\
e_{17} &: \text{Lex}(\text{'rectangular'}, \text{A}) \wedge \left[\text{s-ev} : \left[e=e_{16} : / \text{rectangular}/ \right] \right] \\
e &: \left(\begin{array}{l} \left[\begin{array}{l} e_1 : \text{start}(e_1) \\ e_2 : \text{start}(e_2) \end{array} \right] \wedge \left[\begin{array}{l} 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{array} \right] \wedge \left[\begin{array}{l} e_7 : \text{start}(e_7) \\ e_8 : \text{start}(e_8) \\ e_9 : \text{start}(e_9) \\ e_{10} : \text{start}(e_{10}) \\ e_{11} : \text{start}(e_{11}) \\ e_{12} : \text{start}(e_{12}) \end{array} \right] \\ \left[\begin{array}{l} e_7 : \text{end}(e_7) \\ e_8 : \text{end}(e_8) \\ e_{13} : \text{start}(e_{13}) \\ e_{14} : \text{start}(e_{14}) \end{array} \right] \wedge \left[\begin{array}{l} e_{13} : \text{end}(e_{13}) \\ e_{14} : \text{end}(e_{14}) \\ e_{15} : \text{start}(e_{15}) \\ e_{16} : \text{start}(e_{16}) \\ e_{17} : \text{start}(e_{17}) \end{array} \right] \wedge \left[\begin{array}{l} e_{11} : \text{end}(e_{11}) \\ e_{12} : \text{end}(e_{12}) \\ e_{16} : \text{end}(e_{16}) \\ e_{17} : \text{end}(e_{17}) \\ e_{10} : \text{end}(e_{10}) \end{array} \right] \end{array} \right)
\end{aligned}$$

The rules e_{15} and e_9 are still open and will be completed once the object noun is processed: this completes e_{15} and gives rise to an NP which in turn can take the 'req' slot of the VP rule in e_9 . Along these lines, ultimately also the sentence hypothesis is confirmed (as are a couple of projective signs in between).

Along with the chart growing, the $\text{cont}(\text{ent})$ fields of the HPSG_{TTR} signs gets compositionally computed. To this end, a combination of functional application and unification (i.e. *merge* in TTR) is employed. An intransitive verb, for instance, is represented by means of a lambda abstract over an argument, as usual. The argument slot is eventually filled by a merge operation when the verb combines with the subject noun. By this means, 'the monitoring and update/clarification cycle is modified to happen *at the end of each word utterance event*' (emphasis in original, Ginzburg et al., 2018, 470). Such a move is necessary in order to analyse intra-sentential acknowledgements (Poesio and Traum, 1997):

(26) A: The house has a rectangular shape
B: mhm

The noun phrase utterance of A gives, via chart parsing, rise to the following NP semantics (where the sit-type value is of type *RecType*):

(27)

dgb.pending =

$$\begin{aligned}
&\left[\begin{array}{l} \text{s}=\text{u}_1 : \text{Sit} \\ \text{sit-type} = \left[\begin{array}{l} \text{dgb-params} : \left[\begin{array}{l} \text{refind} : \text{Ind} \\ \text{refset} : \text{Set}(\text{Ind}) \\ \text{compset} : \text{Set}(\text{Ind}) \\ \text{maxset} : \text{Set}(\text{Ind}) \\ \text{c1} : \text{house}(\text{maxset}) \\ \text{c2} : \text{union}(\text{refset}, \text{compset}, \text{maxset}) \end{array} \right] \\ \text{cont} : \left[\text{x}=\text{refind} : \text{Ind} \right] \end{array} \right] \end{array} \right]
\end{aligned}$$

The lexical entry for 'mhm' (slightly modified from Ginzburg, 2012, 286) allows for utterance fragments – that is, the maximal element in pending need not be a full *IllocProp* but can be of its super-type *RecType*:

$$\begin{aligned}
(28) \quad &\left[\begin{array}{l} \text{phon} : \text{mhm} \\ \text{cat}=\text{interjection} : \text{SynCat} \\ \text{dgb-params} : \left[\begin{array}{l} \text{ag1} : \text{Ind} \\ \text{ag2} : \text{Ind} \\ \text{MaxPending} : \text{RecType} \\ \text{ad} : \text{address}(\text{ag1}, \text{ag2}, \text{MaxPending}) \end{array} \right] \\ \text{cont} = \left[\text{c0} : \text{Understand}(\text{ag2}, \text{MaxPending}) \right] : \text{IllocProp} \end{array} \right]
\end{aligned}$$

‘mhm’ generates an acknowledgement move that shifts the acknowledged parts onto QUD and MOVES, ‘spreading incrementality’ over DGBs. An incremental backchannel account has also been developed for Dynamic Syntax (using TTR as incremental semantic framework, [Eshghi et al., 2015](#)): here, acknowledgement is modelled in terms of alignment of self- and other-pointers on parse trees. But why not just wait until the end of the turn? Apart from issues of anticipatory comprehension, we speculate that working memory burden also plays a role: once acknowledged, a sub-utterance need not to be maintained in the phonologic loop any more, releasing it from the need to be monitored. But we leave this for future research ([Ginzburg and Lücking, 2020](#)).

The so-called multimodal charts so far are strictly speaking *bimodal* charts. In order to integrate more than two modalities, a generalisation of bimodal charts, that is *proper* multimodal charts, has to be given. Technically, this can be done just by adding more layers to the multichart depicted in [Fig. 1](#). Conceptually we suggest to use scores.

3.6 Scores

The string charts in [Sec. 3.3](#) already exemplified a cascading structure in the sense that an acoustic event not only is an event in itself but also instantiates a ‘sign event’, whose components belong to linguistic knowledge and not to the directly observable world. Multimodal charts ([Sec. 3.5](#)) demonstrated that manual gesture events also can be processed. And there is no reason to stop here, given the variety of nonverbal signals (cf. the remarks at the beginning of [Sec. 1](#)). In this respect, communication events are like musical events, where a bunch of players (the articulators) jointly perform a piece ([Duranti, 1997](#); this analogy has also been drawn in phonetics, where the vocal articulators (lips, tongue, velum, glottis) are conceived as being organised on a ‘gestural score’, [Browman and Goldstein, 1990](#).³). For the coordinated action of playing the first four bars of the second movement of Beethoven’s first Razumovsky quartet, op. 59 no. 1, [Cooper \(2013\)](#) provides the following string event analysis, where we abbreviate his representations of the musical material of the first bar with ‘b1’, and so on:

³We treat the complex interplay of ‘sub-articulators’ in phonetics as a single articulator on a tier. This indicates that a more detailed, hierarchical account is possible in this respect.

$$(29) \left[\begin{array}{ll} v1 & : Ind \\ c_{v1} & : violin(v1) \\ v2 & : Ind \\ c_{v2} & : violin(v2) \\ va & : Ind \\ c_{va} & : viola(va) \\ co & : Ind \\ c_{co} & : cello(co) \\ c_{player} & : players(e, \{v1, v2, va, co\}) \\ c_{play_{v1}} & : play(v1, e.e_{v1}) \\ c_{play_{v2}} & : play(v2, e.e_{v2}) \\ c_{play_{va}} & : play(va, e.e_{va}) \\ c_{play_{co}} & : play(co, e.e_{co}) \\ e & : \left(\begin{array}{l} e_{v1} : Silent \\ e_{v2} : Silent \\ e_{va} : Silent \\ e_{co} : ([b1] \wedge [b2] \wedge [b3]) \end{array} \right) \sim \left(\begin{array}{l} e_{v1} : Silent \\ e_{v2} : ([b4]) \\ e_{va} : Silent \\ e_{co} : ([b4]) \end{array} \right) \end{array} \right]$$

We now adopt the notion of musical scores in order to model multimodal communication.

4 Communication scores

The communication event of an agent is partitioned into tiers, where each tier is tied to its articulator like the quartet is decomposed into the single string instruments. We focus on speech, gaze, head, and manual gesture here, further tiers can easily be added on that model. Note that the communication events on each tier include the empty event ε (‘silence’), so there is no assumption that signalling goes on all the time on every channel.

$$(30) \text{ Tiers} =_{\text{def}} \left[\begin{array}{ll} ag & : Ind \\ c_{\text{signal}} & : produce(ag, e) \\ mth & : Ind \\ c_{mth} & : mouth-of(mth, ag) \\ c_{\text{art}_{mth}} & : articulate(mth, e.e_{sp}) \\ eye & : Ind \\ c_{eye} & : eyes-of(eye, ag) \\ c_{\text{art}_{eye}} & : articulate(eye, e.e_{gz}) \\ l-hnd & : Ind \\ c_{l-hnd} & : left-hand-and-arm-of(hnd, ag) \\ c_{\text{art}_{l-hnd}} & : articulate(l-hnd, e.e_{gs}) \\ r-hnd & : Ind \\ c_{r-hnd} & : right-hand-and-arm-of(hnd, ag) \\ c_{\text{art}_{r-hnd}} & : articulate(r-hnd, e.e_{gs}) \\ c_{\text{sync}} & : Rel(l-hnd, r-hnd) \\ hd & : Ind \\ c_{hd} & : head-of(hd, ag) \\ c_{\text{art}_{hd}} & : articulate(hd, e.e_{hd}) \\ e & : \left(\begin{array}{l} e_{sp} : Phon \\ e_{gz} : VisSit \\ e_{gs} : Trajectory \\ e_{hd} : headMove \end{array} \right)^+ \end{array} \right]$$

Although (30) captures multimodal communication events, the speaker role is a distinguished one:

it is the speaker who produces verbal signals on the speech tier (i.e., *Phon* is not the empty string).

$$(31) \text{ speaker} = \left[\begin{array}{l} \text{ag} : \text{Ind} \\ \text{c}_{\text{signal}} : \text{produce}(\text{ag}, e) \\ e : ([e_{\text{sp}} \neq \varepsilon : \text{Phon}]) \end{array} \right]$$

We refer to the speaker in this narrower sense as ‘ $c_{\text{spkr}} : \text{spkr}(\text{ag})$ ’, or simply use the reserved label ‘*spkr*’ (conversely, ‘*addr*’ labels the agent not speaking according to (31)).

Multimodal communication events can be processed in terms of multicharts (Sec. 3.5). Communication scores combine multimodal scores to dialogue gameboards (Sec. 3.4). To this end, a DGB of the type in (15) is distributed over agent dimensions and a dialogue management dimension: utterance events (including the visual field, which is the content of the gaze tier) are tied to the respective contributors while the arguably more objective parts of a dialogue gameboard are tracked on the DGB dimension (for public vs. private partitions of the DGB see Ginzburg, 2012, Sec. 4.2). The blueprint of dyadic dialogue is that two agents jointly produce a discourse, a string of ‘bars’ (scaling up to multilogue is straightforward):

$$(32) \left[\begin{array}{l} \text{ag1} : \text{Ind} \\ \text{ag2} : \text{Ind} \\ \text{participants} : \{\text{ag1}, \text{ag2}\} \\ \text{cdiag} : \text{produce}(\text{ag1}, \text{ag2}, e) \\ e : \text{String}(\text{Bar}) \end{array} \right]$$

A bar is an incrementally unfolding unit of discourse which captures (pieces of) contributions made by an agent in isolation as well as overlapping signalling events – thus, it replaces the traditional notion of *turn*, which is too sequential for covering the signalling complexities of dialogue.

$$(33) \text{ Bar} =_{\text{def}} \left[\begin{array}{l} e_{\text{ag1}} : \left[\begin{array}{l} \text{c-utt} : \text{address}(\text{ag1}, \text{ag2}, \text{utt-time}) \\ \text{utt-time} : \text{Time} \end{array} \right] \wedge \text{Tiers} \\ e_{\text{ag2}} : \left[\begin{array}{l} \text{c-utt} : \text{address}(\text{ag2}, \text{ag1}, \text{utt-time}) \\ \text{utt-time} : \text{Time} \end{array} \right] \wedge \text{Tiers} \\ e_{\text{dgb}} : \left[\begin{array}{l} \text{facts} : \text{set}(\text{Prop}) \\ \text{pending} : \text{list}(\text{LocProp}) \\ \text{moves} : \text{list}(\text{LocProp}) \\ \text{qud} : \text{poset}(\text{Question}) \\ \text{convrule} : \text{set}(\text{Rule}) \end{array} \right] \end{array} \right]$$

Examples are given in Figs. 2 and 3.

Dialogue progression works as follows: The possibly multimodal contributions of dialogue agents *ag1* and *ag2* are incrementally processed as shown

in Sec. 3.5. The (multimodal) locutionary propositions manifest dialogue moves and are the link to DGB. Progress on the DGB dimension is then regimented by conversational rules. The progression of dialogues is now modelled in terms of scores of communication events, which not only integrates several modalities but also captures some overlap phenomena such as backchannelling. A lexical entry for backchannel nodding is as follows (that a couple of nonverbal signals constitute lexical resources has been worked out by Poggi, 2001 for various kinds of signals):

$$(34) \left[\begin{array}{l} \text{spkr} : \text{Ind} \\ \text{addr} : \text{Ind} \\ e_{\text{spkr}} : \left(\begin{array}{l} e_{\text{sp}} : \text{Phon} \\ e_{\text{locprop}} : \text{Sign}(e_{\text{sp}}) \\ p = e_{\text{locprop}}.\text{cont} : \text{Prop} \end{array} \right) \\ e_{\text{addr}} : \left(\begin{array}{l} e_{\text{hd}} = \text{nod} : \text{headMove} \\ \text{cont} = \text{Accept}(\text{addr}, \text{spkr}, p) : \text{IllocProp} \end{array} \right) \end{array} \right]$$

That is, nodding has a dialogue move-based meaning, contributing an illocutionary proposition. Being a lexical resource on their own, a nonverbal signal event *e* can also be the argument of a multimodal extension of the *Sign(e)* function from (13). Take, for example, the short exchange between a speaker (S) and an addressee (A) reported by Bavelas and Gerwing (2011):

$$(35) \begin{array}{ll} \text{S:} & \text{And I got a light for Christmas} \\ \text{A:} & [\text{slight nod}] \end{array}$$

An analysis of the backchannelling nodding from (35) is given in Fig. 2, where an empty initial dialogue context is assumed (and DGB typings are omitted for brevity). Research on backchannelling (Bavelas et al., 2000; Bavelas and Gerwing, 2011; Tolins and Fox Tree, 2014) agrees that backchannel signals influence the development of dialogue. We can make these findings more precise in comparing backchannelling from Fig. 2 with its (hypothetical) counterpart in Fig. 3 where nodding is a turn in itself and follows the speaker’s utterance. The crucial difference pertains to the DGB dimension: while without backchannelling (Fig. 3) it takes two steps to construct the final dialogue gameboard, backchannelling (Fig. 2) allows to get there in one swoop. Other examples from Sec. 2 can be modelled in basically the same way.

5 Head shake

In Sec. 2 we briefly pointed at utterance coherence, which (presumably not very surprisingly) is also

$$\left[\begin{array}{l} \text{participants} : \{A, B\} \\ c_{\text{diag}} : \text{produce}(A, B, e) \end{array} \right] \left(\left[\begin{array}{l} e : \left(e_{\text{dgb}} : \left[\begin{array}{l} \text{facts} = \{\}: \text{set}(\text{Prop}) \\ \text{pending} = \langle \rangle : \text{list}(\text{LocProp}) \\ \text{moves} = \langle \rangle : \text{list}(\text{LocProp}) \\ \text{qud} = \{\}: \text{poset}(\text{Question}) \\ \text{convrule} = \{\}: \text{set}(\text{ConvRule}) \end{array} \right] \right) \sim \left[\begin{array}{l} e_A : \left[\begin{array}{l} \text{spkr} = A : \text{Ind} \\ e_1 : \left[e_{\text{sp}} = \text{AIL} : \text{Phon} \end{array} \right] \end{array} \right] \sim \left[\begin{array}{l} e_A : \left[\begin{array}{l} \text{spkr} = A : \text{Ind} \\ e_2 : \left[e_{\text{sp}} = \text{FC} : \text{Phon} \right] \\ e_{\text{locprop}} : \text{Sign}(e_1 e_2) \\ p1 = e_{\text{locprop}}.\text{cont} : \text{Prop} \end{array} \right] \\ e_B : \left[\begin{array}{l} \text{ag2} = B : \text{Ind} \\ e_3 : \left[e_{\text{hd}} = \text{nod} : \text{Move} \right] \\ e_{\text{locprop}} : \text{Sign}(e_3) \\ \text{accept} = e_{\text{locprop}}.\text{cont} : \text{IllocProp} \end{array} \right] \\ e_{\text{dgb}} : \left[\begin{array}{l} \text{facts} = \{p1\} \\ \text{pending} = \langle \rangle \\ \text{moves} = \left\langle \begin{array}{l} \text{Accept}(B, A, p1), \\ \text{Assert}(A, B, p1) \end{array} \right\rangle \\ \text{qud} = \{\} \\ \text{convrule} = \{\text{FreeSpeech}, \text{Accept}, \text{Fact update}\} \end{array} \right] \end{array} \right) \right] \right)$$

Figure 2: Backchannelling. AIL abbreviates ‘And I got a light’, FC abbreviates ‘for Christmas’.

obeyed in tier-crossing discourse. We briefly want to illustrate coherence and the kinds of tier-crossing phenomena we want to deal with by means of three simple examples of head shake. For *intra*-speaker head shakes seem to be sensitive to first vs. third person subjects:

- (36) a. Peter believes you
head shaking
b. Peter doesn’t believe you
head shake
c.?I believe you
head shaking
d. I don’t believe you
head shaking

In (36a), the head shake signals the speaker’s attitude to the proposition expressed (\approx ‘I can’t believe that Peter believes you’). (36b) is ambiguous between an attitude and negation interpretation: the head shake can either be as in (36a) (\approx ‘I can’t believe that Peter doesn’t believe you’), or be a nonverbal expression of the negative particle *n’t*.⁴ Both readings – emphasising the *multi*-modal, score-based approach – can be dissociated if an emblematic index-finger shaking (meaning roughly “no”) is produced simultaneously (but see Subsec-3.5 on refinements with respect to timing) in addition to the head shake. In this case, index finger and

head shake are both interpreted as denials – that is, contributing to the illocutionary role –, not as attitude signals, leading to an inconsistency with (36a) but not (36b), since (b) but not (a) expresses a denial.

Curiously, interpretations change when the proposition expressed is a first person report (speaker and subject are identical), as in (36c) and (d). Now, we argue, head shake only contributes to the illocutionary role of the turn, namely signalling a *denying* move. However, denying needs to be witnessed on the propositional level: if there is no negative particle, there is nothing denied, leading to the inconsistency observed in (36c).⁵

6 Conclusion

The formal outline developed in this paper is an attempt to connect incremental natural language processing to multimodality and dialogue gameboards by following the model of music scores. These communication scores may pave the way for formal conversation analysis (i.e., bridging between CA and formal dialogue semantics). Part of future work is to spell out dialogical constraints and generalisations that hold across tiers and between contributions, including a multimodal extension of lexical resources. Notwithstanding multimodality, there

⁴For further uses of the head shake see Kendon (2002).

⁵Head shake, illocutionary roles and attitudes are, among others, further discussed in Ginzburg & Lücking, manuscript in prep.

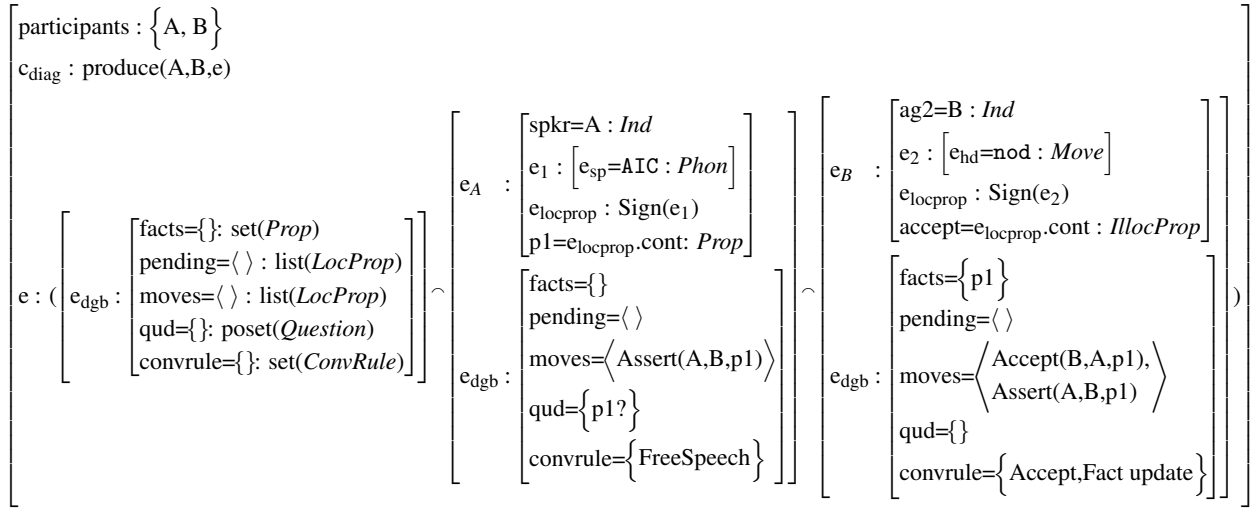


Figure 3: Without backchannelling. AIC abbreviates ‘And I got a light for Christmas’.

still seems to be a ‘leading voice’ among signals, usually speech.

Acknowledgments

We would like to thank three anonymous reviewers not only for their detailed and valuable comments, but also for their benevolent appreciation of the nascent nature of the work presented here. This work is 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.
- Janet B. Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952.
- Janet B. Bavelas and Jennifer Gerwing. 2011. The listener as addressee in face-to-face dialogue. *International Journal of Listening*, 25(3):178–198.
- Rens Bod and Remko Scha. 1996. Data-oriented language processing: An overview. Technical report, Department of Computational Linguistics, Institute for Logic, Language and Computation, University of Amsterdam.
- Catherine P. Browman and Louis Goldstein. 1990. Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, 18:299–320.
- Robin Cooper. 2005. [Austinian truth, attitudes and type theory](#). *Research on Language and Computation*, 3(2-3):333–362.
- Robin Cooper. 2008. Type theory with records and unification-based grammar. In Fritz Hamm and Stephan Kepser, editors, *Logics for Linguistic Structures*, number 201 in Trends in Linguistics: Studies and Monographs, pages 9–33. Mouton de Gruyter, Berlin and New York.
- Robin Cooper. 2012. Type theory and semantics in flux. In Ruth Kempson, Tim Fernando, and Nicholas Asher, editors, *Philosophy of Linguistics*, number 6 in Handbook of Philosophy of Science, pages 271–323. Elsevier, Oxford and Amsterdam.
- Robin Cooper. 2013. Type theory, interaction and the perception of linguistic and musical events. In Martin Orwin, Christine Howes, and Ruth Kempson, editors, *Language, Music and Interaction*, pages 67–90. College Publications.
- Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. 2015. Probabilistic type theory and natural language semantics. *Linguistic Issues in Language Technology – LiLT*, 10(4):1–43.
- 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. Wiley-Blackwell, Oxford, UK.
- 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.

- Alessandro Duranti. 1997. Polyphonic discourse: Overlapping in Samoan ceremonial greetings. *Text – Interdisciplinary Journal for the Study of Discourse*, 17(3):349–382.
- Jay Earley. 1970. [An efficient context-free parsing algorithm](#). *Communications of the ACM*, 13(2):94–102.
- Cornelia Ebert. 2014. The non-at-issue contributions of gestures. Workshop on Demonstration and Demonstratives, April 11–12 2014, Stuttgart.
- Nick J. Enfield. 2009. *The Anatomy of Meaning: Speech, Gesture, and Composite Utterances*. Number 13 in Language, Culture and Cognition. Cambridge University Press, Cambridge, UK.
- Arash Eshghi, Christine Howes, Eleni Gregoromichelaki, Julian Hough, and Matthew Purver. 2015. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 261–271.
- Jerome Feldman. 2012. [The neural binding problem\(s\)](#). *Cognitive neurodynamics*, 7(1):1–11.
- Tim Fernando. 2007. [Observing events and situations in time](#). *Linguistics and Philosophy*, 30:527–550.
- Tim Fernando. 2011. [Constructing situations and time](#). *Journal of Philosophical Logic*, 40(3):371–396.
- Gottlob Frege. 1918. Der Gedanke. *Beiträge zur Philosophie des deutschen Idealismus*, 1(2):58–77.
- Jonathan Ginzburg. 1994. An update semantics for dialogue. In *Proceedings of the first International Workshop on Computational Semantics*, Tilburg, The Netherlands. Katholieke Universiteit Brabant.
- Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford, UK.
- Jonathan Ginzburg, Robin Cooper, Julian Hough, and Schlangen David. 2018. Incrementality and clarification/slucing potential. In *Proceedings of Sinn und Bedeutung 21*, volume 1, pages 463–480.
- Jonathan Ginzburg, Robin Cooper, Julian Hough, and David Schlangen. 2020. 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 and Andy Lücking. 2020. Context as emotionally charged memory. In *Proceedings of The 24th Workshop on the Semantics and Pragmatics of Dialogue*, SemDial/WatchDial.
- Jonathan Ginzburg and Massimo Poesio. 2016. [Grammar is a system that characterizes talk in interaction](#). *Frontiers in Psychology*, 7:1938.
- Alvin I. Goldman. 1970. *Theory of Human Action*, legacy ed. 2015 edition. Number 1830 in Princeton Legacy Library. Princeton University Press, Princeton, NJ.
- Eleni Gregoromichelaki, Ruth Kempson, Matthew Purver, Gregory J. Mills, Ronnie Cann R., Wilfried Meyer-Viol, and Patrick G.T. Healey. 2011. [Incrementality and intention-recognition in utterance processing](#). *Dialogue and Discourse*, 2(1):199–233.
- Michael Johnston. 1998. Unification-based multimodal parsing. In *Proceedings of the 36th Annual Meeting on Association for Computational Linguistics – Volume I*, pages 624–630, Montreal, Quebec, Canada.
- Ruth Kempson, Ronnie Cann, Eleni Gregoromichelaki, and Stergios Chatzikyriakidis. 2016. [Language as mechanisms for interaction](#). *Theoretical Linguistics*, 42(3-4):203–276.
- Adam Kendon. 2002. Some uses of the head shake. *Gesture*, 2(2):147–182.
- Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press, Cambridge, MA.
- Alex Lascarides and Matthew Stone. 2009. [A formal semantic analysis of gesture](#). *Journal of Semantics*, 26(4):393–449.
- Stephen C. Levinson and Francisco Torreira. 2015. [Timing in turn-taking and its implications for processing models of language](#). *Frontiers in Psychology*, 6(731).
- Andy Lücking. 2013. *Ikonische Gesten. Grundzüge einer linguistischen Theorie*. De Gruyter, Berlin and Boston. 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, 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, Malta. 7th International Conference for Language Resources and Evaluation.
- Andy Lücking and Jonathan Ginzburg. 2019. [Not few but all quantifiers can be negated: towards a referentially transparent semantics of quantified noun phrases](#). In *Proceedings of the Amsterdam Colloquium 2019*, AC’19, pages 269–278.
- Craig Martell, Chris Osborn, Jesse Friedman, and Paul Howard. 2002. FORM: A kinematic annotation

- scheme and tool for gesture annotation. In *Proceedings of Multimodal Resources and Multimodal Systems Evaluation*, pages 15–22, Las Palmas, Spain. MITRE.
- David McNeill. 1992. *Hand and Mind – What Gestures Reveal about Thought*. Chicago University Press, Chicago.
- Cornelia Müller. 1998. *Redebegleitende Gesten. Kulturgeschichte – Theorie – Sprachvergleich*. Number 1 in *Körper – Kultur – Kommunikation*. Berlin Verlag, Berlin. Zugl. Diss. FU Berlin (1996).
- Massimo Poesio and David Traum. 1997. Conversational actions and discourse situations. *Computational Intelligence*, 13(3):309–347.
- Isabella Poggi. 2001. [The lexicon and the alphabet of gesture, gaze, and touch](#). In Angélica de Antonio, Ruth Aylett, and Daniel Ballin, editors, *Intelligent Virtual Agents*, number 2190 in *Lecture Notes in Computer Science*, pages 235–236. Springer, Berlin and Heidelberg.
- Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. CSLI Publications, Stanford, CA.
- Kristina Poncin and Hannes Rieser. 2006. [Multi-speaker utterances and co-ordination in task-oriented dialogue](#). *Journal of Pragmatics*, 38(5):718–744. Focus-on Issue: Linguistic Theory and Pragmatics.
- Hannes Rieser. 2015. When hands talk to mouth. Gesture and speech as autonomous communicating processes. In *Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue, SEMDIAL 2015: goDIAL*, pages 122–130, Gothenburg, Sweden.
- 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.
- Jürgen Streeck. 2008. [Depicting by gesture](#). *Gesture*, 8(3):285–301.
- Jackson Tolins and Jean E. Fox Tree. 2014. [Addressee backchannels steer narrative development](#). *Journal of Pragmatics*, 70:152–164.
- 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.