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.\textsuperscript{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, backchanneling 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 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 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 $a_1 : T_1$, $a_2 : T_2(a_1)$, $\ldots$, $a_n : T(a_1, a_2, \ldots, a_{n-1})$, then the record on the left in (2) is of the record type on the right in (2):

\[
\begin{align*}
&\{l_1 = a_1\} \\
&\vdots \\
&\{l_n = a_n\}
&\\
&\{l_1 : T_1\} \\
&\vdots \\
&\{l_n : T(l_1, l_2, l_{n-1})\}
\end{align*}
\]

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 ‘$\&$’ is exemplified in (3):

\[
\begin{align*}
&\text{(3)} \ a. A = \begin{bmatrix} l_1 : T_1 \\
&l_2 : T_2(l_1) \end{bmatrix} \text{ and } B = \begin{bmatrix} l_1 : T_1 \end{bmatrix} \\
&\text{b. } A \& B = \begin{bmatrix} l_1 : T_1 \\
&l_2 : T_2(l_1) \\
&l_1 : T_3 \end{bmatrix}
\end{align*}
\]

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 $e_1 e_2 e_3$ 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:

\[
\begin{align*}
&\{t_0 = e_1\}, \text{ where time indices } t_i \text{ are in } \mathbb{N}. \\
&t_1 = e_2 \\
&t_3 = e_3
\end{align*}
\]

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 notationally enclosed in brackets. For instance, the first move of a chess player playing a Sicilian opening is represented in (5):
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

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 → 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:

```
(5) [x : Ind
  e : (e : chess-player(x))
  y : Ind
  l1 : Loc
  c1 : field-c7(l1)
  l2 : Loc
  c2 : field-c5(l2)
  cm : on(y,l1)
  e : (e : grab(x,y)) − [e : move-to(x,y,l2)] − [e : drop(x,y)].

(6) a. Lex(‘Beethoven’, NP)
   
   b. [s-event:
       e : beethoven
       spkr : Ind
       addr : Ind
       cm : addressing(spkr,addr,e)
       phon : /Beethoven/
       syn : cat=mp : Cat
       dtrs = ( ) : list(Sign)
       q-params : [refind : Ind
                   cm : named(refind, ‘Beethoven’)]
       cont=q-params.refind : Ind]

   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 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: Lex(T, C) ∧ [s-event : [e=u : T]].

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

(8) [e1 beethoven : Phon
     e : (e1 : start(e1)) − [e1 : end(e1)]]

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) [e1 beethoven : Phon
     e2 : Lex(‘Beethoven’, NP) ∧ [s-ev: [e=e1 : /Beethoven/]]
     e : (e1 : start(e1)) − [e1 : end(e1)]
     e2 : start(e2)) − [e2 : end(e2)]

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 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

(10) [e1 beethoven : Phon
      e2 : Lex(‘Beethoven’, NP) ∧ [s-ev: [e=e1 : /Beethoven/]]
      e1 : (rule=S−→NP VP : NP−→VP fnd=e2 : Sign)
      e : (e1 : start(e1)) − [e1 : end(e1)]
      e2 : start(e2)) − [e2 : end(e2)] − [e3 : end(e3)]]

      If the string chart processes an input of the category marked as required, a sentential parse will be achieved:
cont labels a structure of type (record type):

\[ \text{cat=vp} \land \text{dtrs=} \]

Projection also an anticipatory version of Sign(e) can be defined, 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 \( \sigma \) leading from the syntactic parse of \( u \) to \( s \), then

\[ \text{ProjSign(e)} \]

We will refer to the maximal sign processed by a chart for a speech event \( e \) as 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:

\[ (13) \]

Sign(e)

a. \( \text{sites-event : [e : Phon]} \land \text{sit-type : Lex(T, Sign)} \land \text{sit.e : T} \)

b. \( \text{sites-event : [e : (Phon)^{\text{\small{Phon}}}] } \land \text{sit-type : Sign} \land \text{syn : [dtrs=e : String(Sign)]} \)

Furthermore, given the mechanism of rule projection also an anticipatory version of Sign(e) can be defined, ProjSign(e) (‘projected sign’). The
A special kind of propositions are *locutionary propositions* (LocProp, Ginzburg, 2012, 172):

\[
\text{LocProp} := \left[ \begin{array}{c}
\text{sit} : \text{Sign} \\
\text{sit-type} : \text{RecType}
\end{array} \right]
\]

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):

\[
\text{preconds : qud=\{\} : \text{poset(Question)}}
\]

\[
\text{effects : turnUnderspec } \land \\
\text{a : Prop} \\
\text{R : IllocRel} \\
\text{moves.latest=R(spkr,addr,a) : IllocProp}
\]

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)\}.

![Multimodal chart parser](image)

Within HPSG grammars, speech-gesture integration rules have been defined based on multichart parsing (Alahverdzhiya et al., 2017; Lücking, 2016). 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\(_{\text{TR}}\) 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”.

\[
\text{sg-ensemble}
\]

\[
\text{s-dtr} \rightarrow \text{phon=aff.s-dtr.phon : Phon} \\
\text{cat=aff.s-dtr.cat : SynCat} \\
\text{cont=aff.s-dtr.cont : cvm}
\]

\[
\text{g-dtr} \rightarrow \text{phn=aff.g-dtr.phon : Phon} \\
\text{cat=aff.g-dtr.cat : SynCat} \\
\text{cont=aff.g-dtr.cont : cvm}
\]

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*TR* 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 handshake, 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.

$$\text{(19) a. Gest}$$

$$\begin{align*} 
\text{hand} & : \{\text{left, right}\} \\
\text{hs} & : \text{hShape} \\
\text{carrier} & : \begin{cases} \\
\text{wrist} : \text{Move} \\
\text{palm} : \text{Dir} \\
\text{boh} : \text{Dir} \\
\text{mov} : \text{Conc} \end{cases} \\
\text{sync} & : \begin{cases} \\
\text{s-loc} : \text{gSpace} \\
\text{e-loc} : \text{gSpace} \end{cases} \\
\text{rel} & : \text{gRel} \\
\end{align*}$$

$$\text{(19) b. Post-stroke hold}$$

$$\begin{align*} 
\text{hand} & : \{\text{left, right}\} \\
\text{hs} & : \text{hShape} \\
\text{carrier} & : \begin{cases} \\
\text{wrist} : \text{Move} \\
\text{palm} : \text{Dir} \\
\text{boh} : \text{Dir} \\
\text{mov} : \text{Conc} \end{cases} \\
\text{sync} & : \begin{cases} \\
\text{s-loc} : \text{gSpace} \\
\text{e-loc} : \text{gSpace} \end{cases} \\
\text{rel} & : \text{gRel} \\
\end{align*}$$

For example, moving the wrist rightwards, back (i.e., towards the body of the gesturer), and leftwards in a rectangular manner (‘line’) – $$\text{path} : \text{line}$$ $$\text{wrist=mr}^{\text{mb}^{\text{ml}}} : \text{Move}$$ – 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 fulfill 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 RE Ct- 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

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_0) and a gesture integration rule (e_{10}) have been triggered, but are still pending:

(24)

\[
\begin{align*}
e_1 &= \text{the : } \text{Phon} \\
e_2 &= \text{Lex('the', DET) } \land s\text{-event : } e=e_1 : /\text{the}/ \\
e_3 &= \left( \text{rule=}NP\rightarrow\text{DET N : DET} \land \text{N} \rightarrow \text{Sign} \land \text{rule=} \land \text{sign=} \right) \\
e_4 &= \text{house : } \text{Phon} \\
e_5 &= \text{Lex('house', N) } \land s\text{-event : } e=e_4 : /\text{house}/ \\
e_6 &= \text{prep : } \text{Phase} \\
e_7 &= \text{has : } \text{Phon} \\
e_8 &= \text{Lex('have', V) } \land s\text{-event : } e=e_7 : /\text{has}/ \\
& \quad \text{ rule=}\text{VP} \rightarrow \text{V NP} \\
& \quad \text{ req} = \text{NP : } \text{Sign} \\
& \quad \text{ e : required(req, rule)} \\
& \quad \text{rule=}sg\text{-ensemble} \rightarrow X[\text{accent,cvm}] \text{ stroke} \\
& \quad \text{ req} = \text{stroke : } \text{Phase} \\
& \quad \text{ req2=}X[\text{accent,cvm}] : \text{Sign} \\
& \quad \text{ e : required(req1,req2,rule)} \\
\end{align*}
\]

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:
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 cont(ent) fields of the HPSGTR 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 reason intrasentential 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) dgbr.pending =

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:

(28) dgbr-params:

‘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’, Brown 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:

\[
\begin{align*}
&v1 : \text{Ind} \\
&c_{v1} : \text{violin(v1)} \\
&v2 : \text{Ind} \\
&c_{v2} : \text{violin(v2)} \\
&va : \text{Ind} \\
&c_{va} : \text{viol(a)va} \\
&co : \text{Ind} \\
&c_{co} : \text{cello(co)} \\
&c_{\text{player}}(e,\{v1,v2,va,co\}) \\
&c_{\text{play}_1}(e) \\
&c_{\text{play}_2}(e) \\
&c_{\text{play}_3}(e) \\
&c_{\text{play}_4}(e) \\
&\epsilon : \{\epsilon_{v1} : \text{Silent}, \epsilon_{v2} : ([b4]), \epsilon_{va} : ([b4]), \epsilon_{co} : ([b4])\} \\
&\epsilon_{v1} : \text{Silent} \\
&\epsilon_{v2} : \text{Silent} \\
&\epsilon_{va} : \text{Silent} \\
&\epsilon_{co} : \text{Silent} \\
&\epsilon_{v1} : ([b4]) \\
&\epsilon_{v2} : ([b4]) \\
&\epsilon_{va} : ([b4]) \\
&\epsilon_{co} : ([b4])
\end{align*}
\]

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 \(\epsilon\) (‘silence’), so there is no assumption that signalling goes on all the time on every channel.

\[
\begin{align*}
&\text{ag} : \text{Ind} \\
&c_{\text{signal}} : \text{produce}(ag,e) \\
&mth : \text{Ind} \\
&c_{\text{mth}} : \text{mouth-of}(mth,ag) \\
&c_{\text{art}_\text{mth}} : \text{articulate}(mth,e,c_{\text{sp}}) \\
&\text{eye} : \text{Ind} \\
&c_{\text{eye}} : \text{eyes-of}(eye,ag) \\
&c_{\text{art}_\text{eye}} : \text{articulate}(eye,e,c_{\text{g}2}) \\
&l-hnd : \text{Ind} \\
&c_{\text{l-hnd}} : \text{left-hand-and-arm-of(hnd,ag)} \\
&c_{\text{art}_\text{l-hnd}} : \text{articulate}(l-hnd,e,c_{\text{g}3}) \\
&r-hnd : \text{Ind} \\
&c_{\text{r-hnd}} : \text{right-hand-and-arm-of(hnd,ag)} \\
&c_{\text{art}_\text{r-hnd}} : \text{articulate}(r-hnd,e,c_{\text{g}3}) \\
&\text{sync} : \text{Rel}(l-hnd,r-hnd) \\
&\text{hd} : \text{Ind} \\
&c_{\text{hd}} : \text{head-of(hd,ag)} \\
&c_{\text{art}_\text{hd}} : \text{articulate}(hd,e,c_{\text{hd}}) \\
&\epsilon : \{c_{\text{sp}} : \text{Phon}, c_{\text{g}2} : \text{VisSit}, c_{\text{g}3} : \text{Trajectory}, c_{\text{hd}} : \text{headMove}\}^+
\end{align*}
\]

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).

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

We refer to the speaker in this narrower sense as ‘spkr’ (spkr(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 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

\[ \text{Bar} = \left[ \begin{array}{l}
\text{ag}_1 : \text{Ind} \\
\text{ag}_2 : \text{Ind} \\
\text{participants} : \{\text{ag}_1, \text{ag}_2\} \\
\text{e}_{\text{diag}} : \text{produce}(\text{ag}_1, \text{ag}_2, e) \\
\text{e} : \text{String(Bar)}
\end{array} \right] \]

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.

\[ \text{Bar} = \left[ \begin{array}{l}
\text{c-utt} : \text{address}(\text{ag}_1, \text{ag}_2, \text{utt-time}) \\
\text{utt-time} : \text{Time} \\
\text{facts} : \text{set(Prop)} \\
\text{pending} : \text{list(LocProp)} \\
\text{moves} : \text{list(LocProp)} \\
\text{quad} : \text{poset(Quotation)} \\
\text{convrule} : \text{set(Rule)}
\end{array} \right] \]

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):

\[ \text{spkr} : \text{Ind} \\
\text{addr} : \text{Ind} \\
\text{e}_{\text{spkr}} : (\{\text{e} : \text{Phon}\}) \\
\text{e}_{\text{locprop}} : \text{Sign}(\text{e}_{\text{locprop}}, \text{prop}) \\
\text{e}_{\text{addr}} : (\{\text{e}_{\text{addr}} : \text{headMove} \text{cont} \text{Accept}(\text{addr,spkr,p}) : \text{IllocProp}\}) \]

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):

\( \text{(35) S: And I got a light for Christmas} \)

\( \text{A: [slight nod]} \)

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
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 (≈ ‘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) (≈ ‘I can’t believe that Peter doesn’t believe you’), or be a nonverbal expression of the negative particle n’t. 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 Subsection 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

5Head shake, illocutionary roles and attitudes are, among others, further discussed in Ginzburg & Lücking, manuscript in prep.
We would like to thank three anonymous reviewers\footnote{ANR-10-LABX-0083). It contributes to the IdEx Katya Alahverdzhieva, Alex Lascarides, and Dan Université de Paris – ANR-18-IDEX-0001. Investments d’Avenir of the program French National Research Agency (ANR) as part 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 \textit{Investissements d’Avenir} (reference: ANR-10-LABX-0083). It contributes to the IdEx Université de Paris – ANR-18-IDEX-0001.\footnote{ANR-18-IDEX-0001. Investments d’Avenir of the program French National Research Agency (ANR) as part of the program \textit{Investissements d’Avenir} (reference: ANR-10-LABX-0083). It contributes to the IdEx Université de Paris – ANR-18-IDEX-0001.}

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


Cornelia Ebert. 2014. The non-at-issue contributions of gestures. Workshop on Demonstration and Demonstratives, April 11-12 2014, Stuttgart.


Craig Martell, Chris Osborn, Jesse Friedman, and Paul Howard. 2002. FORM: A kinematic annotation


