


ARTICLE

Leading voices: dialogue semantics, cognitive science and the polyphonic structure of multimodal interaction

Andy Lücking^{1,2*}  and Jonathan Ginzburg¹

¹Laboratoire de Linguistique Formelle (LLF), Université Paris Cité, CNRS – UMR 7110, Paris, France; ²Text Technology Lab, Goethe University Frankfurt, Frankfurt am Main, Germany

*Corresponding author. Email: andy.luecking@u-paris.fr

(Received 16 May 2022; Revised 05 October 2022; Accepted 13 October 2022)

Abstract

The neurocognition of multimodal interaction – the embedded, embodied, predictive processing of vocal and non-vocal communicative behaviour – has developed into an important subfield of cognitive science. It leaves a glaring lacuna, however, namely the dearth of a precise investigation of the meanings of the verbal and non-verbal communication signals that constitute multimodal interaction. Cognitively construable dialogue semantics provides a detailed and context-aware notion of meaning, and thereby contributes content-based identity conditions needed for distinguishing syntactically or form-based defined multimodal constituents. We exemplify this by means of two novel empirical examples: *dissociated* uses of negative polarity utterances and head shaking, and attentional clarification requests addressing speaker/hearer roles. On this view, interlocutors are described as co-active agents, thereby motivating a replacement of *sequential turn organisation* as a basic organising principle with notions of *leading* and *accompanying voices*. The *Multimodal Serialisation Hypothesis* is formulated: multimodal natural language processing is driven in part by a notion of *vertical relevance* – relevance of utterances occurring simultaneously – which we suggest supervenes on sequential (‘horizontal’) relevance – relevance of utterances succeeding each other temporally.

Keywords: dialogue semantics; multimodal interaction; turn taking; overlap; clarification requests

1. Introduction

Let us face it: it is all about meaning. A phoneme is the smallest meaning-distinguishing sound, a morpheme a meaning-carrying form. Most distinctions even in syntax – long regarded the core of linguistics – are based on semantic considerations. Now, investigating meanings poses a perplexing problem: we cannot directly encounter them or point at them or count them, and talking about meaning itself requires meaning. There are different ways to proceed in this situation. In psycholinguistics, for instance, experimental studies are used, where meaning is observed

indirectly by observable features of language users' processing of stimuli sentences. A quite different approach has been developed in philosophy and formal semantics: here, the act of interpretation is objectified in terms of mathematical models, that is, 'small worlds' which are used as items within which semantic representations of natural language expressions are evaluated. Both approaches exemplify research programmes that target distinct *levels* of meaning: This has recently been discussed in terms of Marrian (Marr, 1982) implementation versus computation (resp. neural activity vs. behaviour; Krakauer et al., 2017), and in terms of cognitive architectures complementing algorithmic representational models (Cooper & Peebles, 2015), among others. With regard to language, there has been a long-standing collaboration: answers to *What?* questions are provided by formal grammar and theoretical linguistics, *How?* questions are addressed in psycholinguistics. Yet, this cooperation cooled down for a while (Ferreira, 2005). There are several reasons for the disenchantment. With regard to meaning proper – that is, *semantics* – we think that theoretical linguistics 'underaccomplishes' the obligation to provide cognitively potent models of meaning given mainstream formal semantics' sentence-oriented approach. The reason is this: consider a toy world that consists of three individuals, *a* (Ayдын), *n* (Nuria), and *x* (Xinying). A mainstream model-theoretic approach to semantics maps natural language expressions onto terms of a formal language (mostly predicate logic), which in turn are interpreted in terms of the individuals of a world (*denotation* or *reference*). The meaning of a one-place predicate like *sleep*, for instance, is the set [*sleep'*] assigned to the formal translation *sleep'* of the verb, and in our toy model (let us assume) is {*a*, *x*} (i.e., Ayдын and Xinying sleep). The meaning of the sentence *Ayдын sleeps* is compositionally derived as *sleep'(a)* and is true iff (abbreviates *if and only if*) $a \in [\textit{sleep}']$. However, the formulae used in traditional formal semantics (e.g., *sleep'(a)*) are *dispensable*: they eventually get reduced to the basic notions of truth and reference (*sleep'(a)*, e.g., is true in our toy model) and therefore have no cognitive bearing. Hence, while being formally precise, it is unclear whether an approach of such a kind succeeds to 'formulate the computational problem under consideration', as Cooper and Peebles (2015, p. 245) put it.

Nonetheless, over the past 30 years, theoretical linguistics *has* developed a different sort of a formal model of meaning, namely *dynamic update semantics* – most notably *Discourse Representation Theory* (Kamp & Reyle, 1993) – where the construction of semantic representations *is* constitutive of meaning (Kamp, 1979, p. 409) and has cognitive (Hamm et al., 2006) and neuroscientific (Brogaard, 2019) interpretations (see also Garnham, 2010). The sentence *Ayдын sleeps* is processed within a dynamic update semantics in such a way that a file¹ for an object *x* (due to the proper name Ayдын) is *opened* (if new) or *continued* (if known). We emphasise this detail since it reveals a dynamic shift in the notion of meaning: the meaning of an utterance updates a previous context and returns an updated context. Hence, reducing the meaning of an assertive sentence to truth and reference is replaced (or at least complemented) by its *context change potential*. The (new or continued) file is then populated with conditions that *x* is named Ayдын (if not already known) and *x* is

¹The metaphor of files and file changing is due to Heim (1982); in cognitive science, the closely related notion of *mental files* is used (Perner et al., 2015) – see also their re-emergence in the philosophy of mind (Recanati, 2012).

sleeping.² A dynamic update semantics rooted in *spoken language* – also known as *dialogue semantics* – is KoS (Ginzburg, 1994, 2012). KoS [not an acronym] is formulated by means of *types* from a Type Theory with Records (TTR; Cooper, 2013; Cooper & Ginzburg, 2015) instead of terms and expressions from an interpreted language like predicate logic. There is a straightforward model-theoretic, denotational construal of types much in the spirit of classical formal semantics (Cooper, n.d.), but one can also think of types as symbolic but embodied structures which are rooted in perception (as Cooper, n.d., points out), which label instances of linguistic processing (Connell, 2019; Frankland & Greene, 2020), and are associated with motor and perception activation (Bickhard, 2008; Hummel, 2011; Meteyard et al., 2012). Indeed, types can also be construed neurally (Cooper, 2019).

Why promote dialogue semantics and to a cognitive science audience? Cognitive science has come to acknowledge that multimodal interaction is the ‘central ecological niche’ of sentence processing (Holler & Levinson, 2019, p. 639). The dominant view on interaction and coordination in cognitive science is a *systemic view*: interlocutors are observed and construed as a complex system – there is work on systemic coupling on neural, behavioural, and attentional, goal-predicting levels (Fusaroli et al., 2014; Hasson et al., 2012; Pickering & Garrod, 2004, 2013; Sebanz & Knoblich, 2009).

However, while a systemic view certainly provides important insights into the neural and cognitive underpinnings of alignment and communication within its ecological niche, we argue that significant lacunae remain unless this is complemented by analyses of the verbal and nonverbal signals (and their interactions) that constitute multimodal communication: the multimodal, interactive turn in cognitive science induces a renewed need for a precise formulation of the computational *What?* problem. Simplifying to a necessary degree, Fig. 1 summarises the semantic position

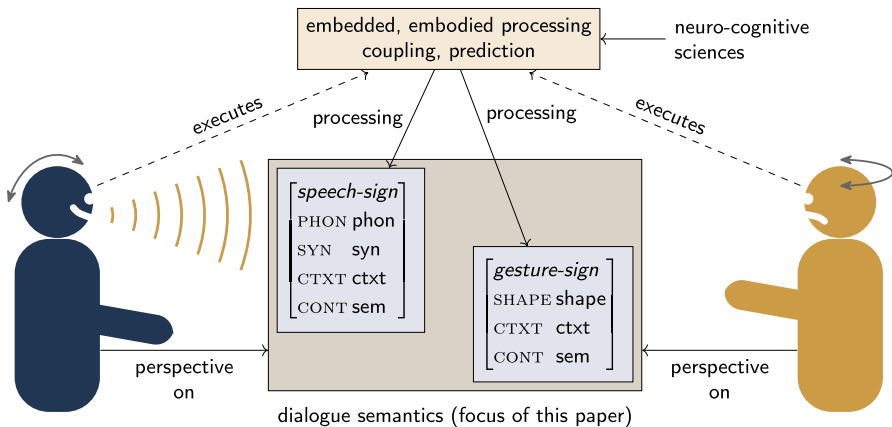


Fig. 1. A dialogue-semantics perspective for completing the systemic understanding of multimodal discourse.

²This is the minimal information that is received from the sentence. One can also add that Aydın very likely is human since it is a common first or family name, and, in recent memory-oriented approaches, that the semantic value for the proper name is to be found in long-term memory (Cooper, n.d.; Ginzburg & Lücking, 2020).

within the multimodal discourse landscape. We focus on contents (CONT) here and demonstrate throughout how contents depend to a very large extent on a fine-grained structured context (CTXT).

In particular, we argue that a dialogue-semantics perspective makes at least three crucial contributions:

- Dialogue semantics provides a formal notion of *content* that is needed in order to define different kinds of cross-modal signals. From gesture studies, we have the notion of *multimodal ensembles* (Kendon, 2004) – utterances including speech–gesture composites – and from psycholinguistics, *multimodal gestalts* (Holler & Levinson, 2019) – recurrent, statistically significant multimodal actions, signals or features which are interlinked by a (common) communicative intent or meaning.³ However, recurrent ensembles or gestalts often occur with a simplification in form (Lücking et al., 2008). This raises the issue of how formally different gestalt or ensemble *tokens* are assigned to a common *type* instead of to different ones. Moreover, how to account for communicative signals, features or utterances which do *not* belong to a unique ensemble or gestalt? We argue mainly based on data from head shaking (Sections 2.1 and 4.2) that an explicit semantic analysis is needed to provide the required identity conditions and, among others, tell apart multimodal behaviour that, with respect to its perceptual forms, deceptively looks like a unified composite utterance.
- In line with research on attentional mechanisms (Mundy & Newell, 2007), we discuss (non-)attending to interlocutors as new attentional data and argue that it can be used to explain – as far as we know – hitherto unstudied occurrences of specific types of other-repair in discourse targeting the speaker and hearer roles.
- Timing and coherence within multimodal interaction is a subject *sui generis* for both cognitive science and dialogue semantics: dialogue agents are co-active during more or less the whole time of interaction – see also the analysis of Mondada (2016).⁴ Accordingly, the notion of turn should be replaced by the notion of *leading voice*. Moreover, this applies even to spoken contributions, where despite the entrenched assumption of *one speaker per turn*, assumed in Conversation Analysis to be one of the essential and universal structuring notions of conversation (Levinson & Torreira, 2015), overlap is in certain situations and with inter-subject variation an acceptable option (Bennett, 1978; Falk, 1980; Hilton, 2018; Tannen, 1984; Yuan et al., 2006). In this respect, multimodal interaction is akin to a polyphonic musical piece.⁵ Just like

³In fact, there is information-theoretic evidence for such gestalts at least on the level of manual co-speech gestures (Mehler & Lücking, 2012). The notion of ‘local gestalts’ used by Mondada (2014) seems to be a generalisation of the notion of ensembles, but to lack the statistical import gained from recurrence.

⁴A more conservative view seems to be embraced by Streeck and Hartge (1992), who analyse mid-turn gestures to ‘contextualise “next speech units”’, including a preparation of potential transition places (p. 137). This view is reinforced in Streeck (2009, Ch. 8).

⁵Thinking of conversational interaction in musical terms has been proposed by Thompson (1993), whereas Clark (1996, Ch. 2, p. 50) mentions string quartets as a ‘mostly nonlinguistic joint activity’. In fact, string quartets were originally inspired by the eighteenth-century French salon tradition (Hanning, 1989). Duranti’s (1997) paper documents what he calls ‘polyphony’ (‘normative overlap’) in Samoan ceremonial greetings. Based on a convergent effect of joint musical improvisation on the alignment of body movements and periodicity across speech turns, it has recently been argued that both music and linguistic interaction

polyphonic music is organised by harmonic or contrapuntal composition techniques, polyphonic interaction is driven partly by dialogical *relevance* or coherence.⁶ Note that the terms ‘leading’ and ‘accompanying voices’ also give rise to a subjective interpretation: a speaker or gesturer may have the impression to hold the leading voice in a conversation regardless of observational evidence.

Section 2 illustrates the above-mentioned challenges posed by multimodal phenomena. Section 3 then sketches a formal theory of multimodal interaction that involves: (i) semantic representations which can (and should) be construed as cognitive information state representations, (ii) *partiturs* (multimodal input representations) and (iii) lexical entries and conversational update rules that capture dialogical relevance enabling incremental and predictive processing. The machinery is applied to analyse the sample observations in Section 4. The formal theory may appear complex to those exposed to it for a first time or who do not endorse formal approaches, but its expressive granularity has been developed in light of many diverse dialogical phenomena, as explained in Section 3.2. In particular, it facilitates formulating our ultimate upshot in Section 5 in an explicit way: our claim, the *multimodal serialisation hypothesis*, is that vertical relevance – relevance of utterances occurring simultaneously – supervenes⁷ on horizontal relevance – relevance of utterances succeeding each other temporally. Hence, multimodality compresses interaction temporally, but is not richer in terms of semantic expressivity. In other words, and with certain caveats we will spell out, simultaneous interaction, though more efficient and perhaps more emotionally engaging and aesthetically pleasing, can always be serialised without loss of semantic information. This is a rather strong claim, and it needs to be refined right away. On the one hand, there are multimodal signals which simply cannot be separated – for instance, you cannot separate spoken utterances from their intonation: they are coarticulated. This is, however, not just due to a common channel: speech–laughter is transmitted via the acoustic channel but can be separated into speech and laughter. On the other hand, serialising multimodal input gives rise to different possible orderings. We do not claim that every order of the elements from multimodal input when put into a sequence is equivalent, quite the contrary: we provide evidence for the opposite below. But in accordance with the claim, one of the possible orderings *is* semantically equivalent to the original multimodal input. Simultaneity and sequentiality in multimodal interaction can always become manifest in two ways: (i) across interlocutors and (ii) within one

belong to a common human communicative facility (Daltrozzo & Schön, 2009; Robledo et al., 2021). However, despite the fact that we use the term *leading voice* in the very title, we use it here solely as a metaphor for depicting the *structure* of multimodal communication. In particular, we do not derive strong implications for the organisation of dialogue (or music) from it; in fact, other comparisons such as *contrapuntal structure* serve similar purposes, as we discuss below.

⁶These two terms are frequently used interchangeably; we use the former for consistency with earlier work in the framework utilised in this paper, KoS. Coherence has been emphasised as a fundamental principle of the alignment of manual co-speech gesture and speech by Lascarides and Stone (2009).

⁷Supervenience is a non-reductionist but asymmetric mode of dependence (see, e.g. Kim, 1984), which, with respect to the multimodal serialisation hypothesis, can be paraphrased as follows: any difference in the set of properties of vertical coherence is accounted for by some difference in the set properties of horizontal coherence, but not the other way round. In this sense does vertical relevance depend on horizontal relevance but does not get ontologically reduced to it.

interlocutor. The multimodal serialisation hypothesis intentionally generalises over both manifestations (in fact, the empirical phenomena discussed in the following involve both kinds). Given these qualifications, the expressivity claim is a hypothesis that has to be explored in multimodal communication studies by cognitive science, theoretical linguistics, gestures studies and related disciplines.

2. Observations

2.1. Head shake

Eight uses of the head shake are documented by Kendon (2002). The most well-known (Kendon's use I) is a non-verbal expression of the answer particle 'No'. Thus, a head shake can be used in order to answer a polar question:

- (1) a. A: (i) Do you want some coffee? / (ii) You do not want some coffee?
 b. B: head shake

Depending on whether A produced a negative or a positive propositional kernel in the question, B's head shake is either a denial of the positive proposition or a confirmation of the negative one (which is not discussed by Kendon, 2002). In uses such as those documented in (1), the head shake conveys a proposition. However, *the proposition expressed by the head shake is in part determined by the context in which it occurs* – in (1), it can be one of two *contradictory* dialogue moves: a *denial* or a *confirmation* one. Hence, what is needed for instances such as (1) is a notion of contextually aware content. We provide such a content in Section 4.2.

Having a context-aware semantic representation of denial at our disposal, it makes predictions for head shakes in other contexts as well. Consider (2):

- (2) a. I do not believe you.
 head shake
 b.? I do believe you.
 head shake

While (2a) is fully coherent, (2b) (at least without additional context – examples of which we provide in Section 4.2) has a contradictory flavour: the head's denial is not matched in speech. Hence, in order to discuss apparently simple uses of head shakes, one already has to draw on a precisely formulated, contextually aware notion of contents.

2.2. Co-activity and communicative breakdown

A well-known pattern of co-activity in spoken discourse is the interplay of monitoring and feedback signals. For instance, backchannelling signals such as nodding or vocalisations such as 'mhm' influence the development of discourse (Bavelas & Gerwing, 2011). The absence of monitoring or feedback signals leads to communicative breakdown since it raises the question whether one is still engaged in the joint interaction. Suppose *A* and *B* are sitting on a window seat in a café. *A* is telling *B* about a near-accident she witnessed on Main Street the day before. While *A* has been talking, *B* has been continuously staring out of the window. Thus, *A* lacks attentional gaze signals, which in turn raises doubts about *B*'s conversational involvement. Accordingly, *A* will try to clarify *B*'s addressee role:

(3) *Hey, are you with me?*

A's clarification request or (other-initiated) repair⁸ is a natural response in the sketched situation since it is triggered by a neurocognitive social attention mechanism (Nummenmaa & Calder, 2009) in response to a violation of a behavioural norm. However, seen from a turn-based view, (3) is not easy to explain: A is speaking and B is listening, so the all-important roles of hearer and speaker are clearly filled – and now it is B's turn. Crucially, (3) could equally be made by B if s/he gets the impression A is rambling incoherently.

2.3. Summary

The upshot of the few phenomena we have discussed above is that multimodal interaction is:

- driven by a richly structured and fine-grained context,
- which is distinct but aligned across the participants,
- where the participants typically monitor each other's co-activity.

In the following, we introduce a theoretical dialogue framework which can capture these observations.

3. Polyphonic interaction: cognitive-formal tools

3.1. Partiturs

A prerequisite for an analysis of multimodal interaction is a systematic means for telling apart the manifold verbal and non-verbal signals. We employ tiers in this respect, where a tier is built following the model of phonetics and phonology. Phonetics comprises the triple of *articulatory phonetics*, *acoustic phonetics* and *auditory phonetics*. Signalling of other communication means can be construed in an analogous way. For instance, *facial muscles – facial display – vision* defines the tier for facial expression. Tiers give rise to a uniform approach to linguistic analysis which ultimately rests on perceptual classification (cf. Cooper, n.d.), which we formulate in terms of TTR (a *Type Theory with Records*; Cooper, n.d.; Cooper & Ginzburg, 2015). Classification in TTR is expressed as a *judgement*: in general, object *o* is of type *T*. With regard to spoken utterances, a *record r* (situation) providing a sound event

(construed as an Individual) – $r = \begin{bmatrix} s_{\text{event}} & = a \\ c & = s1 \end{bmatrix}$ – is correctly classified by a *record*

type T = $\begin{bmatrix} s_{\text{event}} & : \text{Ind} \\ c & : \text{Sign}(s_{\text{event}}) \end{bmatrix}$ (i.e., $r : T$), iff the object labelled 's_{event}' (in this case, a soundwave) belongs to the phonological repertoire of the language in question.⁹

⁸We assume these two latter terms are synonymous; the former often used in the dialogue community, the latter among Conversation Analysis researchers.

⁹Sign is modelled in terms of phonology–syntax–semantics structures developed in *Head-Driven Phrase Structure Grammar* (Pollard & Sag, 1994). We abstract over a speaker's knowledge of a language and the language system where it does not do any harm, as an anonymous reviewer of *Language and Cognition*

Tiers can be likened to different instruments on a musical score: a *partitur*.¹⁰ Building on Cooper (2015), we represent partiturs as *strings* of multimodal communication events *e*, which are temporally ordered sequences of types. One can think of strings in terms of a flip-book: a dynamic event is cut into slices, and each slice is modelled as a record type. Such *string types* (Cooper, n.d.; Fernando, 2007) are notated in round brackets and typed in an obvious manner, where *RecType* is the general type of a record type:

$$(4) \quad \textit{partitur} = e : \left(\left(\begin{array}{l} e_{\text{speech}} : \textit{Phon} \\ e_{\text{gesture}} : \textit{Trajectory} \\ e_{\text{gaze}} : \textit{RecType} \\ e_{\text{head}} : \textit{headMove} \\ e_{\text{face}} : \textit{faceExpr} \end{array} \right) \right)^+$$

The progressive unfolding of subevents on the various tiers in time gives rise to incremental production and perception. Formally, this is indicated by the Kleene plus ('+'). (4) shows the *type* of multi-tier signalling, and it remains silent concerning potential inherent rhythms of the individual tiers. In fact, it has been argued that different kinds of gestures exhibit a specific 'rhythmic pulse' (Tuite, 1993), as does speech, which lead to tier-specific temporal production cycles that may jointly peak in synchronised intervals (Loehr, 2007). The temporal relationship between signals on different tiers is therefore specified in a relative way, following the example set by the *Behaviour Markup Language* (Vilhjálmsón et al., 2007). It should be noted that the subevents on partiturs can be made as detailed as needed – from phonetic features to complete sentences or turns. A reasonable fine-grained temporal resolution of partiturs seems to be the level of syllables. Arguably, syllables constitute *coherent events* as do tones in a melodic phrase and movement elements in locomotion, and to which attentional processes are rhythmically attuned in the sense of Jones and Boltz (1989). See Lücking and Ginzburg (2020) for more details on parsing on partiturs. We will make crucial use of record-type representations along these lines in the following.

3.2. Cognitive states in dialogue semantics

We model cognitive states by means of dialogue agent-specific *Total Cognitive States* (TCS) of KoS (Ginzburg, 1994, 2012; Larsson, 2002; Purver, 2006). A TCS has two partitions, namely a *private* and a *public* one. A TCS is formally represented in (5). In a dialogue between A and B, there are both, A.TCS and B.TCS.¹¹

observed. A speaker who is not aware of a certain word form (sound) will, however, not be able to provide a witness for a sign type containing that form as value of the phon feature. This, in turn, can trigger clarification interaction.

¹⁰We use the Italian word *partitur* (and its English plural variant) since in semantics the term *score* is already taken due to the work of Lewis (1979).

¹¹We restrict attention here to two-person dialogue; for discussion on the differences between two-person and multi-party dialogue and how to extend an account of the former to the latter, see Ginzburg (2012, Sect. 8.1).

$$(5) \quad \text{TCS} := \left[\begin{array}{l} \text{public} : \text{DGBType} \\ \text{private} : \text{Private} \end{array} \right]$$

(The symbol “:=” indicates a definition relation.)

Now, trivially, communication events take place in some context. The simplest model of context, going back to Montague (1974), is one which specifies the existence of a speaker, addressing an addressee at a particular time. This can be captured in terms of the type in (6), which classifies situations (records) that involve the obvious entities and actions.

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

However, over the last four decades it has become clearer how much more pervasive reference to context in interaction is. Indeed, arguably, this traditional formulation gets things backwards in that it seems to imply that ‘context’ is some distinct component one refers to. In fact, as will become clear, following Barwise and Perry (1983), we take utterances – multimodal events – to be the basic units interlocutors assign *contents* to given their current cognitive states and from this generalise to obtain utterance types, the meanings/characters semanticists postulate.

The visual situation is a key component in interaction from birth (see Tomasello, 1999, Ch. 3).¹² Expectations arise due to illocutionary acts – one act (querying, assertion and greeting) giving rise to anticipation of an appropriate response (answer, acceptance and counter-greeting), also known as adjacency pairs (Schegloff, 2007). Extended interaction gives rise to shared assumptions or *presuppositions* (Stalnaker, 1978), whereas uncertainties about mutual understanding that remain to be resolved across participants – *questions under discussion* – are a key notion in explaining coherence and various anaphoric processes (Ginzburg, 1994, 2012; Roberts, 1996). These considerations among several additional significant ones lead to positing a significantly richer structure to represent each participant’s view of publicised context, the *dialogue gameboard* (DGB), whose basic make up is given in (7):

¹²The importance of vision in the establishment of joint attention is affirmed by studies on the development of joint attention in congenitally blind infants (Bigelow, 2003). Blind children must rely on non-visual attention-getting strategies such as hearing or touching. As a consequence, they not only develop joint attention at later stages than sighted children, but also depend on their interlocutors to establish a common focus of attention – at least until the symbolic competence of speech is developed to a sufficient degree (Bigelow, 2003). Furthermore, it has been found in event-related potential studies that congenitally blind subjects (but not sighted ones) recruit posterior cortical areas for the processing of information relevant for an auditory attention task, and in a temporally ordered manner (Liotti et al., 1998). The authors of the study speculate that the observed topographical changes might be due to a ‘reorganisation in primary visual cortex’ (p. 1011). With respect to the Vis-Sit in KoS, this can be seen as evidence that at least some of the visual information is replaced by information from other tiers of the partitur. Hence, a corresponding formal model can in principle be devised, accounting for interactions with congenitally blind interlocutors, an issue brought up by an anonymous reviewer of *Language and Cognition*.

(7)

$DGBType :=$	spkr : <i>Ind</i> addr : <i>Ind</i> utt-time : <i>Time</i> c-utt : addressing(spkr, addr, utt-time) facts : <i>Set(Prop)</i> vis-sit = [foa : <i>Ind</i> \vee <i>Rec</i>] : <i>RecType</i> pending : <i>List(LocProp)</i> moves : <i>List(IllocProp)</i> qud : <i>poset(Question)</i> mood : <i>Appraisal</i>
--------------	---

It should be emphasised (again) that there is not a single DGB covering a dialogical episode, but a DGB for each participant. Participants' DGBs are usually coupled, that is, develop in parallel. Participant specific DGBs, however, allow to incorporate misunderstandings, negotiation, coordination and the like in a straightforward manner in KoS. In any case, *facts* represents the shared assumptions of the interlocutors – identified with a set of propositions. In line with TTR's general conception of (linguistic) classification as type assignment – record types regiment records – propositions are construed as typing relationships between records (situations) and record types (situation types), that is, as Austinian propositions (Austin, 1950; Barwise & Etchemendy, 1987). More formally, propositions are records of type $\left[\begin{array}{l} \text{sit} \quad : \text{Rec} \\ \text{sit-type} : \text{RecType} \end{array} \right]$.¹³ The ontology of dialogue (Ginzburg, 2012) knows two special sorts of Austinian proposition: grammar types classifying phonetic events (*Loc(utionary)Prop(ositions)*) and speech acts classifying utterances (*Illoc(utionary) Prop(ositions)*). Both types are part and parcel of locutionary and illocutionary interaction: dialogue moves that are in the process of being grounded or under clarification are the elements of the *pending* list; already grounded moves (roughly, moves that are not contested, or agreed-upon moves) are moved to the *moves* list. Within *moves*, the first element has a special status given its use to capture adjacency pair coherence and it is referred to as *LatestMove*. The current question under discussion is tracked in the *qud* field, whose data type is a partially ordered set (*poset*). *Vis-sit* represents the visual situation of an agent, including his or her visual focus of attention (*foa*), which, if any (attention may be directed towards something non-visual, even non-perceptual¹⁴), can be an object (*Ind*), or a situation or event (which in TTR are modelled as records, i.e., entities of type *Rec*). *Mood* tracks a participant's public displays of emotion (i.e., externally observable appraisal indicators such as intonation or facial expressions, which often do but need not coincide with the participant's internal emotional state), crucial for *inter alia* laughter, smiling

¹³On this view, a proposition $p = \left[\begin{array}{l} \text{sit} \quad = s \\ \text{sit-type} = T \end{array} \right]$ is true iff $s : T$ – the situation s is of the type T . Note that an incongruous situation type (inquired about by an anonymous reviewer) lacks any witnessing situations and therefore in model-theoretic terms has an 'empty' extension.

¹⁴As is arguably the case in remembering and imagination (Irish, 2020; Werning, 2020).

and sighing (Ginzburg et al., 2020b), and, as we shall see, head shaking as well. The DGB structure in (7) might seem like an overly rich notion for interlocutors to keep track of. Ginzburg and Lücking (2020) show how the DGB type can be recast as a Baddeley-style (Baddeley, 2012) multicomponent working memory model interfacing with long-term memory.

Given that our signs (lexical entries/phrasal rules) are construed as *types for interaction*, they refer directly to the DGB via the field *dgb-params*. For instance, the linguistic meaning of the head shake from (1) in Section 2.1 patterns with the lexical entry for ‘No’ when used as an answer particle to a polar question (a.k.a. a ‘yes–no’ question) and, following Tian and Ginzburg (2016), is given in (8).

$$(8) \quad \left[\begin{array}{l} \text{phon: no / shape: head shake} \\ \\ \text{dgb-params: } \left[\begin{array}{l} \text{spkr: Ind} \\ \text{addr: Ind} \\ \text{u-time: Time} \\ \text{c1: addressing(spkr, addr, u-time)} \\ \text{p: Prop} \\ \text{MaxQUD = p? : PolarQuestion} \end{array} \right] \\ \\ \text{content = Assert(spkr, addr, u-time, NoSem(p)) : IllocProp} \end{array} \right]$$

When used in the context of a polar question with content p (the current question under discussion – MaxQUD – is $p?$), saying ‘No’ and/or shaking the head asserts a ‘No semantics’ applied to p . $NoSem(p)$ in turn is sensitive to the polarity of the proposition to which it applies (cf. the discussion of the head shake in Section 2.1). To this end, positive (*PosProp*) and negative (*NegProp*) propositions have to be distinguished. If a negative particle (*not, no, n’t, never* and *nothing*) is part of the constituents of a proposition $\neg p$, then $\neg p$ is of type *NegProp* ($\neg p$: *NegProp*). The corresponding positive proposition, the one with the negative particle removed, is p (p : *PosProp*). With this distinction at hand, $NoSem$ works as follows:

$$(9) \quad NoSem(p) = \begin{cases} \neg p & \text{if } p : PosProp \\ p & \text{if } p : NegProp \end{cases}$$

(Note that the result of ‘ $NoSem(p)$ ’ is always of type *NegProp* – p : *NegProp* means that $p = \neg q$, which $NoSem$ leaves unchanged according to the second condition in (9).) (8) and (9) provide a precise characterisation of answer particle uses of negation and head shake and therefore make testable predictions concerning meaning in context.

The evolution of context in interaction is described in terms of *conversational rules*, mappings between two cognitive states, the *precond(ition)s* and the *effects*. Two rules are given in (10): a DGB that satisfies *preconds* can be updated by *effects*.

- (10) a. *Assert QUD-incrementation*: given a proposition p and $Assert(A,B,p)$ being the LatestMove, one can update QUD with $p?$ as MaxQUD.

$$\left[\begin{array}{l} \text{preconds : } \left[\begin{array}{l} p \\ \text{LatestMove} = \text{Assert}(\text{spkr}, \text{addr}, p) : \text{IllocProp} \end{array} \right] \\ \text{effects : } [\text{QUD} = \langle p?, \text{pre.QUD} \rangle : \text{poset}(\text{Question})] \end{array} \right]$$

Example: the claim $p =$ ‘Carlsen will retain his title.’ is asserted by the speaker. This leads to the question $p? =$ ‘Will Carlsen retain his title?’ becoming the topmost question under discussion, waiting for the addressee to be accepted (in that case, p will be added to the set of propositions making up facts; see (7)) or discussed.

- b. QSPEC: this rule – a formalisation of Grice’s maxim of relevance – characterises the contextual background of reactive queries and assertions: if q is MaxQUD, then subsequent to this either conversational participant may make a move constrained to be *specific to q* (i.e., either About or Influencing q ; for a formal characterisation of Qspecific, see Ginzburg (5).

$$\left[\begin{array}{l} \text{preconds : } [\text{QUD} = \langle q, Q \rangle : \text{poset}(\text{Question})] \\ \text{effects : } \left[\begin{array}{l} r : \text{Question} \vee \text{Prop} \\ R : \text{IllocRel} \\ \text{LatestMove} = R(\text{spkr}, \text{addr}, r) : \text{IllocProp} \\ c1 : \text{Qspecific}(r, q) \end{array} \right] \end{array} \right]$$

Example: the question $r? =$ ‘Where is my box of chocolates?’ has been posed by the speaker, that is, $r?$ is MaxQUD. Now, both the assertion $p =$ ‘In the cupboard.’ ($\text{LatestMove} = \text{Assert}(\text{spkr}, \text{addr}, p)$) and the question $q? =$ ‘Where were you snacking from it last?’ ($\text{LatestMove} = \text{Ask}(\text{spkr}, \text{addr}, q?)$) are q -specific with respect to $r?$, whereas a question such as $w? =$ ‘Have you already seen the new movie?’ ($\text{LatestMove} = \text{Ask}(\text{spkr}, \text{addr}, w?)$) is not. The latter may therefore lead to other pragmatic interpretations such as trying to change topics.

Within the dialogue update model of KoS, following Ginzburg et al. (2020a), QUD gets modified *incrementally*, that is, at a word-by-word latency (or even higher).¹⁵ Technically, this can be implemented by adopting the predictive principle of incremental interpretation in (11) on top of partitural parsing (see Section 3.1). This says that if one projects that the currently pending utterance (the preconditions in (11)) will continue in a certain way (pending.sit-type.proj in (11)), then one can actually use this prediction to update one’s DGB, concretely to update LatestMove with the projected move; this will, in turn, by application of the existing conversational rules, trigger an update of QUD:¹⁶

¹⁵Ginzburg et al. (2020a) are motivated by data showing unfinished utterances can trigger updates driving, e.g., elliptical phenomena like sluicing: *He could bring the ball down, but opts to cushion a header towards ... well, who exactly? Nobody there.* (From a live match blog)

¹⁶Since there are more and less likely hypotheses concerning the continuation of an ongoing utterance, utterance projection should ultimately be formulated in a probabilistic manner using a probabilistic version

$$(11) \quad \text{Utterance Projection} := \left[\begin{array}{l} \text{preconds : [pending.sit-type.proj = a : Type]} \\ \text{effects : } \left[\begin{array}{l} \text{e1 : Sign} \\ \text{LatestMove = } \left[\begin{array}{l} \text{sit = e1} \\ \text{sit-type = a} \end{array} \right] : \text{LocProp} \end{array} \right] \end{array} \right]$$

We will make use of utterance projection in analysing head shakes synchronous to speech in Section 4.2 and in Section 5 when explicating vertical relevance. Such projective rules implement predictive processing in interactions and therefore provide a computational underpinning of a central cognitive mechanism (Litwin & Miłkowski, 2020).

4. Polyphonic interaction: cognitive–formal analyses

The formal tools from Section 3 are used to provide precise analyses of the observations from Section 2: attention and communicative breakdown (Section 4.1) and the semantics of head shake (Section 4.2).

4.1. Conversational engagement

In two-person conversation, the values of *spkr* and *addr* of a DGB are rarely in question, apart from initially (*Who is speaking? Are you addressing me?*), but the need to verify that the addressing condition holds is what we take to drive attention monitoring. We conceive the two states of being engaged or disengaged in conversation as two hypotheses in a probabilistic Bayesian framework. Relevant data for the (dis-)engagement hypotheses can be found in gaze, which is an excellent predictor of conversational attention (Nummenmaa & Calder, 2009; Vertegaal et al., 2001). The quoted sources as well as the discussion in the following concern unobstructed face-to-face dialogue, that is, dialogue where participants stand or sit opposing each other and can freely talk. The findings and the assumptions derived below do not carry over to ‘obstructed’ discourse situations *simpliciter*, for instance, when interlocutors are talking while carrying a piano.

Within cognitive DGB modelling, the Vis-Sit field already provides an appropriate data structure for gaze. Mutual gaze can be formulated as a perspectival default condition on partiturs.¹⁷ Of course, there is no claim that mutual gazing occurs continuously. Indeed, continuous gaze is often viewed as being rude or encroaching. In fact, mutual gaze tends to be short, often less than a second (Kendon, 1967).

Gaze is not the only attentional signalling system, however. Dialogue agents regularly provide verbal and non-verbal feedback signals (Bavelas & Gerwing, 2011). Among the verbal *reactive tokens* (Clancy et al., 1996) the majority are backchannels. As with gaze, a lack of backchannelling will result in communicative

of TTR (Cooper et al., 2015). Instead of a single *effect*, a range of probabilistically ranked predictions is acknowledged, as is common in statistical natural language parsing (e.g., Demberg et al., 2013). Incremental and predictive processing underlies grammatical framework such as *Dynamic Syntax* from the outset (Gregoromichelaki et al., 2013; Kempson et al., 2001).

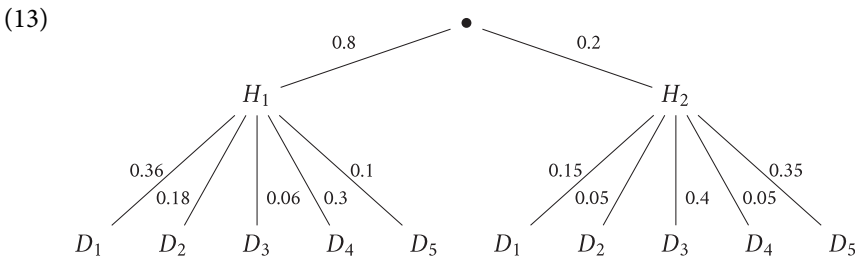
¹⁷Conditions or rules are *perspectival* if they are applicable only to particular dialogue participants; see Ginzburg et al. (2020b, Sect. 4.1.2) for a first use of ‘participant sensitive’ conversational rules.

breakdown. In sum, there is ample evidence that gazing and backchannelling provide important datapoints for tracking (mutual) attention. We combine both into a probabilistic framework along the following lines:

(12) **Bayes' attention hypotheses and data**

- a. $\mathcal{H} = \{H_1 = \text{being engaged}, H_2 = \text{being disengaged}\}$
- b. $\mathcal{D} = \{D_1 = \text{ind.gaze}, D_2 = \text{mutual gaze}, D_3 = \text{gaze away}, D_4 = \text{backchannel}, D_5 = \text{nobackchannel}\}$

We assume that gazing provides slightly more attentional evidence than backchannelling by a proportion of 0.6 to 0.4. We derive the prior probabilities for gaze under H_1 from Argyle (1988, p. 159), and the priors for gaze under H_2 are stipulated, as are the backchannel probabilities. Furthermore, we assume that engagement is the probabilistic default case of interaction with a plausibility of 0.8 to 0.2:



If one of the kinds of gaze from \mathcal{D} is observed, the posterior probability can be calculated from the probability tree in (13) by means of a Bayesian update according to Bayes' theorem ($P(H|D) = \frac{P(D|H)P(H)}{P(D)}$). Let us illustrate an update triggered by an observation of individual gaze, D_1 . Compared to the prior probabilities of the engagement and disengagement hypotheses, D_1 leads to an increase of the probability of H_1 at the expense of H_2 . The corresponding numerical values are collected in Table 1.

The change of the posteriors in comparison to the priors show that the already more probable engagement hypothesis gains further plausibility (increasing from 0.8 to 0.9). Hence, observing individual gaze, D_1 , supports (the public display of) mutual attention. Bayesian updates apply iteratively: In this way, only a mixture of data observations of different kinds leads to an oscillation of H_1 within a certain probability interval. This leads us to a testable hypothesis, namely that the extrema of the oscillation interval constitute thresholds of mutual attention. If the engagement posterior takes a value below the minimum of the interval, it triggers attention clarification: *Are you with me?* Values that exceed the maximum lead to irritation: *Why are you staring at me?*

Table 1. Bayesian update table

Hypothesis	Prior	Likelihood	Bayes numerator	Posterior
\mathcal{H}	$P(\mathcal{H})$	$P(D_1 \mathcal{H})$	$P(D_1 \mathcal{H})P(\mathcal{H})$	$P(\mathcal{H} D_1)$
H_1	0.8	0.36	0.288	0.906
H_2	0.2	0.15	0.03	0.094

4.2. Head shake and noetics

In Section 2, the exchange re-given in (14) was introduced as an obstacle for the *NoSemantics* of head shakes introduced in Section 3.

- (14) a. I do not believe you.
 head shake
 b.? I believe you.
 head shake

If we make the (rather consensual) assumption, that the outcomes of utterances are predicted as soon as possible (see Section 3, in particular example (11)), then an explanation of (14) is straightforward: A's utterance in (14a) provides a negative proposition, $\neg believe(A,B)$, which by *NoSem* the head shake affirms. On the other hand, (14b) provides a positive proposition, $believe(A,B)$, which by the same lexical entry the head shake negates, hence a contradiction ensues.

The contradiction in (14b) can be ameliorated, however:

- (15) (Context: Claims that B stole 500€)
 a. B: They say I stole the money. But I did not.
 b. A: I believe you.
 head shake

In this case, one can understand A as verbally expressing his belief in B's protestation of innocence, whereas the head shake affirms the negative proposition B makes, $\neg stole(B,500€)$ (when related to the second sentence uttered by B), or expresses that A is upset about what 'they' did (when related to B's initial uttered sentence). In either case, this requires us to assume that the head shake can be disassociated from speech that is simultaneous with it, an assumption argued for in some detail with respect to speech laughter by Mazzocconi et al. (2020/22). Such observations are of great importance for a multimodal theory. This is because it has been claimed that multi-tier interpretation is guided by the heuristic 'if multiple signs occur simultaneously, take them as one' (Enfield, 2009, p. 9). Such heuristics have to be refined in consideration of the above evidence.¹⁸

Examples like the head shake in (15b) – which can be glossed 'I disapprove of p ' – are therefore subsumed to a 'negative appraisal use' of negation (Tian & Ginzburg, 2016) by Lücking and Ginzburg (2021), and analysed as a *noetic* act expressing a speaker's attitude towards the content of his or her speech *via* DGB's Mood field.¹⁹ Note, finally, that A's response in (15) can be serialised as head shake followed by

¹⁸As pointed out by an anonymous reviewer, Enfield's heuristics can be understood more loosely along the lines of 'if multiple signs occur simultaneously, interpret them *in relation to one another*'. Since Enfield does not provide a semantics, there remains some leeway for interpretation. The semantic and pragmatic synchrony rules stated by McNeill (1992) are more explicit in this respect ('[...] speech and gesture, present the same meanings at the same time', p. 27; '[...] if gestures and speech co-occur they perform the same pragmatic functions', p. 29).

¹⁹The term 'noetic' is inspired by William James (James, 1981, Ch. XXV), who emphasised, for instance, that '[i]n instinctive reactions and emotional expressions thus shade imperceptibly into each other' (p. 1058). In this sense, noetics describe how feelings, sentiments, sensations, memories, emotions and unconscious acts bear on and are transmitted through a feedback loop of thinking and knowledge (Krader, 2010). We believe

speech (i.e., head shake + ‘I believe you’). However, the sequence ‘I believe you’ + head shake seems to be a bit odd, illustrating a remark concerning the multimodal serialisation hypothesis we made in Section 1, namely that sequential orderings need not be equivalent. Such temporal effects need to be explored further in future studies.

5. Upshot: from ‘horizontal’ to ‘vertical’ relevance in multimodal dialogue

In uni-modal interaction (best exemplified perhaps by chat conducted sequentially between users across a network), conversation is constrained by *relevance* or *coherence* between *successive* participant moves (and ultimately across longer stretches). For reasons related to our metaphor with musical notion (cf. *partiturs*), we call this notion *horizontal* relevance.

Some examples for relevant (indicated by ‘✓’) responses to a query and to an assertion are given in (16a,b) and irrelevant ones (indicated by ‘#’) to both in (16c).

- (16) a. A: Is that chair new? B: ✓Yes / ✓It’s a Louis XIV replica / ✓New?
 b. A: Jill arrived late last night. B: ✓She did not. / ✓Why? / ✓Jill? / ✓To spite us.
 c. B: # Tomorrow / # Please insert your card / # The train.

For conversation, the query/response relation is the one studied in greatest detail (Berninger & Garvey, 1981; Ginzburg et al., 2022; Stivers & Enfield, 2010). The basic characterisation of this relationship given in Ginzburg et al. (2022) is that the class of responses to a question q_1 can be partitioned into three classes.

- (17) a. q(uestion)-specific: responses directly about or subquestions of q_1 ;
 b. MetaCommunicative: responses directly about or subquestions of a question defined in part from the *utterance* of q_1 ;
 c. Evasion: responses directly about or subquestions of a question that is distinct from q_1 and arises from some other component of the context:
 1. Ignore (address the situation, but not the question; e.g., *Anon: on the Sunday before you killed the animals, you did not in fact feed them. Why was that? Harry: Only water* (BNC));
 2. Change the topic (e.g., *Nicola: Come on, let us get dressed. Which pants are you wearing? Oliver: What’s he got on his mouth?* (BNC));
 3. Motive (‘Why do you ask?’);
 4. Difficult to provide a response (‘I do not know’).

A formal account of horizontal relevance in terms of conversational rules is given in Ginzburg (2012, Sects. 4.4.5 and 6.7.1). The basic idea is that an utterance u is relevant in the current context iff u can be integrated as the (situational component of the) LatestMove via some conversational rule.

that emphasising the inherent integration of appraisal and content, among others, is a useful way of conceiving attitudes in conversations.

Table 2. Vertical relevance: possible content relationships between overlapping utterances across two speakers

u_A .cont	u_B .cont	Relationship	Examples
p	$\neg p$	Negation	B head shake/speech (Kendon, 2002)
		Disbelief	B laugh (Ginzburg et al., 2020b)
p	p	Agreement	B head nod/speech (Hadar et al., 1985) B low arousal laugh
p	$prob(p) < \theta$	Doubt	B head tilt (Heylen, 2008)
	Understand(B, u_A)	Acknowledgement	B mild nod (Hadar et al., 1985)
	\neg Understand (B, u_A)	Clarification request	B frown and head back/speech: what? (Poggi, 2001)
p	find_disgusting (B, p)	Negation, disgust	B 'Not face' with action units AU9, AU10 and AU17 (Benitez-Quiroz et al., 2016), also other faces discussed
	disengaged(B, u_a)	Incapacity, powerlessness, indetermination, indifference, obviousness	B shrug + rotating forearms outwards with hand in 'palm up' position + mouth closed, lips pulled downwards (potentially combined with eyebrow raising and head tilt; Debras, 2017)
	presupposes \neg WishDiscuss (B, u_A)	Topic-changing, interruptions	Simultaneous speech (Bennett, 1978; Hilton, 2018)
$\left[\begin{array}{l} \text{sit} = s \\ \text{sit-type} = T_1 \end{array} \right]$	$\left[\begin{array}{l} \text{sit} = s \\ \text{sit-type} = T_2 \end{array} \right]$	Shared situation assessments	(Falk, 1980; Goodwin & Goodwin, 1992)
Rel(A,B)	CounterRel(B,A)	Chordal greetings and partings	(Schegloff, 2000)

But how does the sequential notion of horizontal relevance relate to simultaneous interaction on partiturs, that is, to vertical relevance (to stick to the basic metaphor)? We believe that vertical relevance is supervenient on horizontal relevance. To the best of our knowledge, a careful study, either experimental or corpus-based, of vertical dialogical relevance has yet to be undertaken, apart from one subclass of cases involving speech, known as *overlaps* and *interruptions*, to which we return in our discussion below. We offer an initial, partial and impressionistic characterisation of the notion of vertical relevance in Table 2.

Table 2 offers a selection of signals/contents that a non-leading voice *B* can express simultaneously relative to a leading voice *A* (speaking in terms of turn-replacements, not in terms of subjectively assumed importance; cf. Section 1). Note that two cases

can be distinguished. The first case involves a single speaker for whom certain signals from the multimodal utterances may take the leading voice over other ones. A natural leading voice is speech (de Ruiter, 2004). Co-leading or accompanying roles of non-verbal signals can be assigned in relation to speech. In this respect, at-issue (\approx co-leading) and non-at-issue (\approx accompanying) uses of co-verbal manual gestures have been distinguished (Ebert, 2014).

The second case concerns the distribution of voices among several interlocutors. Inhabiting a leading or an accompanying role is rooted in processes of utterance projection (11) and incremental QUD construction, as we discuss in more formal detail below. We assume that the interlocutor who is responsible for publicly constructing the initial QUD – a process which (by the first case above) can be multimodal or even nonverbal itself – has/is the leading voice. We think that the classic notion of turn holder dissolves into the notion of leading voice. Accompanying voices are characterised by monitoring the incremental QUD construction and commenting on it – in ways exemplified in Table 2. In the most trivial case, this consists in providing backchannelling, but it may also involve the joint production of an utterance (in which case, it could be argued that the accompanying voice becomes a co-leading voice).

The final class we mention is one that has been, in certain respects, much studied, namely simultaneous *speech*. This is a somewhat controversial area because whereas the ‘normativity’ of one speaker using speech and another producing a non-verbal signal is not in question, the normativity of the corresponding case where both participants use speech is very much in question. This is so given the notion of *turn* and the rule-based system which interlocutors are postulated to follow in the highly influential account of Sacks et al. (1974). This system is based on the assumption that normatively at any given time there should be a single speaker; deviations are ‘performance errors’, either unintentional *overlaps* or one interlocutor *interrupting*, attempting to gain the floor. The set-up we have provided does not *predict* any sharp contrast between non-speech/speech overlap and speech/speech overlap, although this could in principle be enforced by introducing conversational rules privileging the speech tier. Nonetheless, we do not think such a strategy is promising. Rather, there are other explanatory factors which conspire to suppress pervasive overlap. In a study of the multilingual *CallHome* corpus, Yuan et al. (2007) note that overlapping varies across languages, with significantly more (non-backchannel) overlaps in Japanese than in the other languages they study (Arabic, English, German, Mandarin and Spanish); they also find that males and females make more overlaps when talking to females than to males, and similarly find more overlaps when talking with familiars than with strangers. Tannen (1984) argues for the existence of distinct conversational styles, including a *high-involvement* style that favours a fast delivery pace, cooperative overlaps and minimal gaps contrasting with a dichotomous *high-considerateness* style. Hilton (2018) conducted a study which found statistically significant correlations between a subject’s conversational style preference and their assessment of the acceptability of overlaps. All this argues against viewing *avoidance of overlapping* as a fundamental, systematic organising principle.

Can we say anything systematic based on subject matter about cases where overlap seems to be acceptable? There is no dearth of evidence for such cases going back to Bennett (1978), Falk (1980), Goodwin and Goodwin (1992) and indeed Schegloff (2000), who while defending the basic intuition underlying Sacks et al. (1974) list various cases of acceptable overlaps. We mention several subclasses: the first involves

what we dub, following Goodwin and Goodwin (1992), *shared situation assessments*. Examples of this are given in (18a,b); in all three cases, a single situation is being described. A second class, noted by Schegloff (2000), is symmetric moves like greetings, partings and congratulations (“we won!” “Yay!” etc.). A third class is exemplified by the attested (18c) – cases where the same question is being addressed; additional instances of this, noted by Schegloff (2000), are utterances involving self-addressed questions (Tian et al., 2017) and ‘split utterances’ – utterances started by A and completed by B (Goodwin, 1979; Gregoromichelaki et al., 2011; Lerner, 1988; Poesio & Rieser, 2010).

- (18) a. B: y’know where they are separate and they do differently things and we are doing this
and $\left[\begin{array}{l} \text{there’s a y’know we operate in a vacuum} \\ \text{C: Mhm, yeah you choose the party you want.} \end{array} \right]$ And you choose what you want.
(Bennett, 1978, example (2))²⁰
- b. O: You do not seem too enthusiastic about it
 $\left[\begin{array}{l} \text{J: well it was a great trip yeah except that} \\ \text{R: it was a good trip yeah} \end{array} \right]$ it was yeah it was a foggy day and we ...
(Falk, 1980, example II)
- c. Paul: Tell y- Tell Debbie about the dog on the ((smile intonation begins)) golf course t’day
Eileen: eh $\left[\begin{array}{l} \text{hnh ha} \\ \text{Paul: hih hih} \end{array} \right]$ $\left[\begin{array}{l} \text{has! ha!} \\ \text{Heh Heh! *hh hh *h} \end{array} \right]$
Eileen: Paul en I got ta the first green, (0.6)
Eileen: *hh An this beautiful, ((swallow))
Paul: I $\left[\begin{array}{l} \text{rish Setter ((reverently))} \\ \text{Eileen: Irish Setter} \end{array} \right]$
Debbie: Ah::,
Eileen: Came tear $\left[\begin{array}{l} \text{in upon ta the first =} \\ \text{Paul: Oh it was beautiful} \end{array} \right]$
Eileen: =gree(h)n an tried ta steal Pau(h)l’s go(h)lf ball. *hh
(Goodwin & Goodwin, 1992, example (1))
- d. M: How old was he? $\left[\begin{array}{l} \text{D: Not very old} \\ \text{J: Very old} \end{array} \right]$
D: No, not that old.

Our assumption throughout has been that vertical relevance supervenes on horizontal relevance – what we labelled earlier the *multimodal serialisation hypothesis*. We adopt this assumption since, at least on the basis of Table 2, all polyphonic utterances seem to have sequential manifestations which give rise to equivalent

²⁰I would like to think of discourse as not so much an exchange but as a shared world that is built up through various modes of mutual response over the course of time in particular interaction.’ (Bennett, 1978, p. 574).

contents; such cases, nonetheless, do lead to distinct DGBs since the partiturs in the two cases are distinct. On the other hand, we believe that there exist sequential adjacency pairs that do not have polyphonic manifestations which give rise to equivalent contents: turn-assigning moves, such as those arising by using the assignee's name or via gaze, do not have a polyphonic equivalent.

Assuming supervenience to hold, we derive vertical relevance from conversational rules by applying *incrementalisation* – in other words, given two conversational rules CR_1 and CR_2 that can apply in sequence where A holds the turn as a consequence of CR_1 and this is exchanged in CR_2 , if by means of incremental interpretation B finds herself in a DGB applicable to CR_2 before the move taking place CR_1 is complete, an overlap arises. To make this concrete, A asserting p and B discussing whether p is the case can be explicated in terms of the sequence of *Assert QUD-incrementation* and *QSPEC* (see (10)). Incrementalising this involves B using *Assert QUD-incrementation* before A completed their utterance, which then satisfies the preconditions of *QSPEC*. In such a case, as discussed above, A is the 'leading voice' and B is an 'accompanying voice'.

All this means that to the extent that the conversational rules underlying horizontal relevance ensure the coherence of dialogue, the same applies to dialogue with polyphonic utterances. Given this, incrementalising conversational rules provides a detailed model for coherence-driven, predictive processing in natural language interaction. In particular, it makes the testable prediction that accompanying behaviour commenting on a leading voice (examples of which are collected in Table 2) is *expected* to occur before the leading voice finished its contribution on its own.

6. Conclusions

We have outlined a unified framework for describing multimodal dialogical interaction. We show how minor adjustments to an existing dialogue framework, KoS, which provides richly structured cognitive states and conversational rules along with (i) *partiturs*, representations of multimodal events and (ii) an incremental semantic framework are needed to analyse multimodal phenomena.

- We demonstrate the existence of noetic head shakes whose contents are dissociated from simultaneous speech. Such dissociation has been demonstrated in previous work for laughter.
- We offer a testable, quantitative account of mutual gaze repair and backchannelling driven by monitoring of participant roles – *not enough* leading to clarification requests, *too much* leading to complaints.
- We have argued that *no overlap* is not a defensible norm in multimodal interaction, including in cases where the two tiers involve speech. The intrinsically sequential notion of *turn* should be replaced by a notion such as *leading/accompanying voice*, which is driven by *vertical coherence*.

On the more basic level of theory design, the observations we discussed all exemplify the need for analytic semantic tools within the systemic landscape of cognitive science. We argued that a dynamic dialogue semantics incarnates a cognitively potent, formally precise linguistic framework for fertilising cross-talk between the disciplines.

As is frequently pointed out but cannot be overemphasized, an important goal of formalization in linguistics is to enable subsequent researchers to see the defects of an analysis as clearly as its merits; only then can progress be made efficiently. (Dowty, 1979, p. 322)

The issues of timing and coherence as captured in terms such as *leading voice* and *vertical relevance* have been identified as specific topics within multimodal dialogue semantics.

Acknowledgements. We wish to thank Judith Holler, two anonymous reviewers for *Language and Cognition*, Robin Cooper, Mark Liberman, Chiara Mazzocconi, and Hannes Rieser, for comments on earlier versions of this paper. Portions of this paper have been presented at the 2021 Dialogue, Memory, and Emotion workshop in Paris, at seminars in Bochum, Saarbrücken, at the Padova Summer School on Innovative Tools in the Study of Language, and at the 2022 ESSLLI summer school in Galway. We also wish to thank audiences there for their comments.

Funding statement. This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the programme 'Investissements d'Avenir' (reference: ANR-10-LABX-0083). It contributes to the IdEx Université Paris Cité – ANR-18-IDEX-0001.

References

- Argyle, M. (1988). *Bodily communication* (2nd ed.). Routledge.
- Austin, J. L. (1950). Truth. In *Proceedings of the Aristotelian society. Supplementary, Reprinted in John L. Austin: Philosophical papers* (2nd ed., Vol. XXIV, pp. 111–128). Clarendon Press.
- Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, 63, 1–29. <https://doi.org/10.1146/annurev-psych-120710-100422>
- Barwise, J., & Etchemendy, J. (1987). *The Liar*. Oxford University Press.
- Barwise, J., & Perry, J. (1983). *Situations and attitudes*. MIT Press.
- Bavelas, J. B., & Gerwing, J. (2011). The listener as addressee in face-to-face dialogue. *International Journal of Listening*, 25(3), 178–198. <https://doi.org/10.1080/10904018.2010.508675>
- Benitez-Quiroz, C. F., Wilbur, R. B., & Martinez, A. M. (2016). The not face: A grammaticalization of facial expressions of emotion. *Cognition*, 150, 77–84. <https://doi.org/10.1016/j.cognition.2016.02.004>
- Bennett, A. (1978). Interruptions and the interpretation of conversation. *Annual Meeting of the Berkeley Linguistics Society*, 4, 557–575.
- Berninger, G., & Garvey, C. (1981). Relevant replies to questioners: Answers versus evasions. *Journal of Psycholinguistic Research*, 10(4), 403–420.
- Bickhard, M. H. (2008). Is embodiment necessary? In C. Paco & T. Gomila (Eds.), *Handbook of cognitive science: An embodied approach, perspectives on cognitive science, chapter 2* (pp. 29–40). Elsevier.
- Bigelow, A. E. (2003). The development of joint attention in blind infants. *Development and Psychopathology*, 15(2), 259–275. <https://doi.org/10.1017/s0954579403000142>
- Brogaard, B. (2019). What can neuroscience tell us about reference? In B. Abbott & J. Gundel (Eds.), *The Oxford handbook of reference* (pp. 365–383). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199687305.013.17>
- Clancy, P. M., Thompson, S. A., Suzuki, R., & Tao, H. (1996). The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of Pragmatics*, 26(3), 355–387. [https://doi.org/10.1016/0378-2166\(95\)00036-4](https://doi.org/10.1016/0378-2166(95)00036-4)
- Clark, H. (1996). *Using language*. Cambridge University Press.
- Connell, L. (2019). What have labels ever done for us? The linguistic shortcut in conceptual processing. *Language, Cognition and Neuroscience*, 34(10), 1308–1318. <https://doi.org/10.1080/23273798.2018.1471512>
- Cooper, R. (2015). Type theory, interaction and the perception of linguistic and musical events. In M. Orwin, C. Howes, & R. Kempson (Eds.), *Language, Music and Interaction* (pp. 67–90). College Publications.
- Cooper, R. (2019). Representing types as neural events. *Journal of Logic, Language and Information*, 28(2), 131–155.

- Cooper, R. (2013). *From perception to communication: An analysis of meaning and action using a theory of types with records (TTR)*. Oxford University Press (in press).
- Cooper, R., Dobnik, S., Larsson, S., & Lappin, S. (2015). Probabilistic type theory and natural language semantics. *Linguistic Issues in Language Technology*, 10(4), 1–43. <https://doi.org/10.33011/lilt.v10i.1357>
- Cooper, R., & Ginzburg, J. (2015). Type theory with records for natural language semantics. In S. Lappin & C. Fox (Eds.), *The handbook of contemporary semantic theory* (chapter 12, 2nd ed., pp. 375–407). Wiley-Blackwell.
- Cooper, R. P., & Peebles, D. (2015). Beyond single-level accounts: The role of cognitive architectures in cognitive scientific explanation. *Topics in Cognitive Science*, 7(2), 243–258. <https://doi.org/10.1111/tops.12132>
- Daltrozzo, J., & Schön, D. (2009). Conceptual processing in music as revealed by N400 effects on words and musical targets. *Journal of Cognitive Neuroscience*, 21(10), 1882–1892. <https://doi.org/10.1162/jocn.2009.21113>
- de Ruyter, J. P. (2004). On the primacy of language in multimodal communication. In *Proceedings of the workshop on multimodal corpora* (pp. 38–41). European Language Resources Association (CD-ROM).
- Debras, C. (2017). The shrug: Forms and meanings of a compound enactment. *Gesture*, 16(1), 1–34. <https://doi.org/10.1075/gest.16.1.01deb>
- Demberg, V., Keller, F., & Koller, A. (2013). Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, 39(4), 1025–1066.
- Dowty, D. R. (1979). *Word meaning and Montague grammar*. Reidel.
- Duranti, A. (1997). Polyphonic discourse: Overlapping in Samoan ceremonial greetings. *Text – Interdisciplinary Journal for the Study of Discourse*, 17(3), 349–382.
- Ebert, C. (2014). The non-at-issue contributions of gestures. In *Workshop on demonstration and demonstratives*. University of Stuttgart.
- Enfield, N. J. (2009). *The anatomy of meaning: Speech, gesture, and composite utterances*. Language, Culture and Cognition, Vol. 13. Cambridge University Press.
- Falk, J. (1980). The conversational duet. In B.R. Caron, M. A. B. Hoffman, M. Silva, J. Van Oosten, D. K. Alford, K. A. Hunold, M. Macaulay & J. Manley-Buser (Eds.), *Annual meeting of the Berkeley Linguistics Society* (Vol. 6, pp. 507–514). Berkeley, CA: Berkeley Linguistics Society.
- Fernando, T. (2007). Observing events and situations in time. *Linguistics and Philosophy*, 30(5), 527–550. <https://doi.org/10.1007/s10988-008-9026-1>
- Ferreira, F. (2005). Psycholinguistics, formal grammars, and cognitive science. *The Linguistic Review*, 22(2–4), 365–380. <https://doi.org/10.1515/tlir.2005.22.2-4.365>
- Frankland, S. M., & Greene, J. D. (2020). Concepts and compositionality: In search of the brain's language of thought. *Annual Review of Psychology*, 71(1), 273–303. <https://doi.org/10.1146/annurev-psych-122216-011829>
- Fusaroli, R., Gangopadhyay, N., & Tylén, K. (2014). The dialogically extended mind: Language as skillful intersubjective engagement. *Cognitive Systems Research*, 29–30, 31–39. <https://doi.org/10.1016/j.cogsys.2013.06.002>
- Garnham, A. (2010). Models of processing: discourse. *WIREs Cognitive Science*, 1(6), 845–853. <https://doi.org/10.1002/wcs.69>
- Ginzburg, J. (1994). An update semantics for dialogue. In H. Bunt (Ed.), *Proceedings of the 1st international workshop on computational semantics*. Tilburg University.
- Ginzburg, J. (2012). *The interactive stance: Meaning for conversation*. Oxford University Press.
- Ginzburg, J., Cooper, R., Hough, J., & Schlangen, D. (2020a). Incrementality and HPSG: Why not? In A. Abeillé & O. Bonami (Eds.), *Constraint-based syntax and semantics: Papers in honor of Danièle Godard*. CSLI Publications.
- Ginzburg, J., & Lücking, A. (2020). On laughter and forgetting and reconversing: A neurologically-inspired model of conversational context. In *Proceedings of the 24th workshop on the semantics and pragmatics of dialogue, SemDial/WatchDial*. Brandeis University.
- Ginzburg, J., Mazzocco, C., & Tian, Y. (2020b). Laughter as language. *Glossa*, 5(1), 104. <https://doi.org/10.5334/gjgl.1152>
- Ginzburg, J., Yusupujiang, Z., Li, C., Ren, K., Kucharska, A., & Łupkowski, P. (2022). Characterizing the response space of questions: Data and theory. *Dialogue and Discourse* (forthcoming).

- Goodwin, C. (1979). The interactive construction of a sentence in natural conversation. In G. Psathas (Ed.), *Everyday language: Studies in ethnomethodology* (pp. 97–121). Irvington Publishers.
- Goodwin, C., & Goodwin, M. H. (1992). Assessments and the construction of context. In P. Auer & A. Di Luzio (Eds.), *Rethinking context: Language as an interactive phenomenon* (Vol. 11, pp. 147–190). Amsterdam: John Benjamins.
- Gregoromichelaki, E., Cann, R., & Kempson, R. (2013). On coordination in dialogue: Sub-sentential speech and its implications. In L. Goldstein (Ed.), *Brevity* (chapter 3, pp. 53–73). Oxford University Press.
- Gregoromichelaki, E., Kempson, R., Purver, M., Mills, G. J., Ronnie Cann, R., Meyer-Viol, W., & Patrick, G. T. H. (2011). Incrementality and intention-recognition in utterance processing. *Dialogue and Discourse*, 2(1), 199–233. <https://doi.org/10.5087/dad.2011.109>
- Hadar, U., Steiner, T. J., & Rose, F. C. (1985). Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4), 214–228.
- Hamm, F., Kamp, H., & Van Lambalgen, M. (2006). There is no opposition between formal and cognitive semantics. *Theoretical Linguistics*, 32(1), 1–40.
- Hanning, B. R. (1989). Conversation and musical style in the late eighteenth-century Parisian Salon. *Eighteenth-Century Studies*, 22(4), 512–528.
- Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: A mechanism for creating and sharing a social world. *Trends in Cognitive Sciences*, 16(2), 114–121. <https://doi.org/10.1016/j.tics.2011.12.007>
- Heim, I. (1982). *The semantics of definite and indefinite noun phrases*. PhD thesis. University of Massachusetts Amherst.
- Heylen, D. (2008). Listening heads. In *Modeling communication with robots and virtual humans* (pp. 241–259). Springer.
- Hilton, K. (2018). *What does an interruption sound like?* PhD thesis. Stanford University.
- Holler, J., & Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8):639–652. <https://doi.org/10.1016/j.tics.2019.05.006>
- Hummel, J. E. (2011). Getting symbols out of a neural architecture. *Connection Science*, 23(2), 109–118. <https://doi.org/10.1080/09540091.2011.569880>
- Irish, M. (2020). On the interaction between episodic and semantic representations – constructing a unified account of imagination. In A. Abraham (Ed.), *The Cambridge handbook of the imagination*. (pp. 447–465). Cambridge Handbooks in Psychology. Cambridge University Press. <https://doi.org/10.1017/9781108580298.027>
- James, W. (1981). *The principles of psychology*. Harvard University Press.
- Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review*, 96(3), 459–491. <https://doi.org/10.1037/0033-295X.96.3.459>
- Kamp, H. (1979). Events, instants and temporal reference. In R. Bäuerle, U. Egli, & A. von Stechow (Eds.), *Semantics from different points of view* (pp. 376–417). Springer Series in Language and Communication, Vol. 6. Springer.
- Kamp, H., & Reyle, U. (1993). *From discourse to logic*. Kluwer Academic Publishers.
- Kempson, R., Meyer-Viol, W., & Gabbay, D. M. (2001). *Dynamic syntax*. Blackwell Publishers.
- Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26(1), 22–63. [https://doi.org/10.1016/0001-6918\(67\)90005-4](https://doi.org/10.1016/0001-6918(67)90005-4)
- Kendon, A. (2002). Some uses of the head shake. *Gesture*, 2(2), 147–182.
- Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.
- Kim, J. (1984). Concepts of supervenience. *Philosophy and Phenomenological Research*, 45(2), 153–176.
- Krader, L. (2010). *Noetics: The science of thinking and knowing*. Peter Lang.
- Krakauer, J. W., Ghazanfar, A. A., Gomez-Marín, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, 93(3):480–490. <https://doi.org/10.1016/j.neuron.2016.12.041>
- Larsson, S. (2002). *Issue based dialogue management*. PhD thesis. Gothenburg University.
- Lascarides, A., & Stone, M. (2009). Discourse coherence and gesture interpretation. *Gesture*, 9(2), 147–180.
- Lerner, G. H. (1988). *Collaborative turn sequences: Sentence construction and social action*. PhD thesis. University of California.
- Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6, 731.

- Lewis, D. (1979). Scorekeeping in a language game. In R. Bäuerle, U. Egli, & A. von Stechow (Eds.), *Semantics from different points of view* (pp. 172–187). Springer Series in Language and Communication, Vol. 6. Springer.
- Liotti, M., Ryder, K., & Woldorff, M. G. (1998). Auditory attention in the congenitally blind: Where, when and what gets reorganized? *NeuroReport*, 9(6), 1007–1012.
- Litwin, P., & Milkowski, M. (2020). Unification by fiat: Arrested development of predictive processing. *Cognitive Science*, 44, e12867. <https://doi.org/10.1111/cogs.12867>
- Loehr, D. (2007). Aspects of rhythm in gesture in speech. *Gesture*, 7(2), 179–214.
- Lücking, A., & Ginzburg, J. (2020). Towards the score of communication. In *Proceedings of the 24th workshop on the semantics and pragmatics of dialogue, SemDial/WatchDial*. Brandeis University.
- Lücking, A., & Ginzburg, J. (2021). Saying and shaking ‘no’. In *Proceedings of the 28th international conference on head-driven phrase structure grammar, HPSG 2021*. University Library.
- Lücking, A., Mehler, A., & Menke, P. (2008) Taking fingerprints of speech-and-gesture ensembles: Approaching empirical evidence of intrapersonal alignment in multimodal communication. In *Proceedings of the 12th workshop on the semantics and pragmatics of dialogue, LonDial’08* (pp. 157–164). King’s College London.
- Marr, D. (1982). *Vision*. Freeman.
- Mazzocconi, C., Tian, Y., & Ginzburg, J. (2020/22) What is your laughter doing there: A taxonomy of the pragmatic functions of laughter. *IEEE Transactions of Affective Computing*, 13(3), 1301–1321 (Published online 2020).
- McNeill, D. (1992). *Hand and mind – What gestures reveal about thought*. Chicago University Press.
- Mehler, A., & Lücking, A. (2012). Pathways of alignment between gesture and speech: Assessing information transmission in multimodal ensembles. In G. Giorgolo & K. Alahverdzhieva (Eds.), *Proceedings of the international workshop on formal and computational approaches to multimodal communication under the auspices of ESSLLI*.
- Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, 48(7), 788–804. <https://doi.org/10.1016/j.cortex.2010.11.002>
- Mondada, L. (2014). The local constitution of multimodal resources for social interaction. *Journal of Pragmatics*, 65, 137–156. <https://doi.org/10.1016/j.pragma.2014.04.004>
- Mondada, L. (2016). Challenges of multimodality: Language and the body in social interaction. *Journal of Sociolinguistics*, 20(3), 336–366. https://doi.org/10.1111/josl.1_12177
- Montague, R. (1974). Pragmatics. In R. Thomason (Ed.), *Formal philosophy*. Yale University Press.
- Mundy, P., & Newell, L. (2007). Attention, joint attention, and social cognition. *Current Directions in Psychological Science*, 16(5), 269–274. <https://doi.org/10.1111/j.1467-8721.2007.00518.x>
- Nummenmaa, L., & Calder, A. J. (2009). Neural mechanisms of social attention. *Trends in Cognitive Sciences*, 13(3), 135–143. <https://doi.org/10.1016/j.tics.2008.12.006>
- Perner, J., Huemer, M., & Leahy, B. (2015) Mental files and belief: A cognitive theory of how children represent belief and its intensionality. *Cognition*, 145(Suppl C), 77–88. <https://doi.org/10.1016/j.cognition.2015.08.006>
- Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2), 169–190.
- Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4), 329–347. <https://doi.org/10.1017/S0140525X12001495>
- Poesio, M., & Rieser, H. (2010). Completions, continuations, and coordination in dialogue. *Dialogue and Discourse*, 1(1), 1–89.
- Poggi, I. (2001) Mind markers. In *The semantics and pragmatics of everyday gestures*. Verlag Arno Spitz.
- Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. CSLI Publications.
- Purver, M. (2006). CLARIE: Handling clarification requests in a dialogue system. *Research on Language & Computation*, 4(2), 259–288.
- Recanati, F. (2012). *Mental files*. Oxford University Press.
- Roberts, C. (1996) Information structure in discourse: Towards an integrated formal theory of pragmatics. In *OSU working papers in linguistics* (Vol. 49, pp. 91–136). Department of Linguistics, The Ohio State University.

- Robledo, J. P., Hawkins, S., Cornejo, C., Cross, I., Party, D., & Hurtado, E. (2021). Musical improvisation enhances interpersonal coordination in subsequent conversation: Motor and speech evidence. *PLoS One*, 16(4), e0250166. <https://doi.org/10.1371/journal.pone.0250166>
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735.
- Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29, 1–63.
- Schegloff, E. A. (2007). *Sequence organization in interaction*. Cambridge University Press.
- Sebanz, N., & Knoblich, G. (2009). Prediction in joint action: What, when, and where. *Topics in Cognitive Science*, 1(2), 353–367. <https://doi.org/10.1111/j.1756-8765.2009.01024.x>
- Stalnaker, R. C. (1978). Assertion. In P. Cole (Ed.), *Syntax and semantics* (Vol. 9, pp. 315–332). Academic Press.
- Stivers, T., & Enfield, N. J. (2010). A coding scheme for question–response sequences in conversation. *Journal of Pragmatics*, 42(10), 2620–2626.
- Streeck, J. (2009) *Gesturecraft*. Gesture Studies, Vol. 2. John Benjamins.
- Streeck, J., & Hartge, U. (1992). Previews: Gestures at the transition place. In P. Auer & A. Di Luzio (Eds.), *The contextualization of language* (pp. 135–157). John Benjamins.
- Tannen, D. (1984). *Conversational style: Analyzing talk among friends*. Oxford University Press.
- Thompson, H. S. (1993). Conversation as musical interaction. HCRC Edinburgh unpublished lecture.
- Tian, Y., & Ginzburg, J. (2016) No I am: What are you saying “No” to? In *Sinn und Bedeutung 21*. The University of Edinburgh.
- Tian, Y., Maruyama, T., & Ginzburg, J. (2017). Self addressed questions and filled pauses: A cross-linguistic investigation. *Journal of Psycholinguistic Research*, 46(4), 905–922.
- Tomasello, M. (1999). *The cultural origins of human cognition*. Harvard University Press.
- Tuite, K. (1993). The production of gesture. *Semiotica*, 93(1/2), 83–105.
- Vertegaal, R., Slagter, R., van der Veer, G., & Nijholt, A. (2001). Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *Proceedings of SIGCHI 2001, CHI '01* (pp. 301–308). Association for Computing Machinery. <https://doi.org/10.1145/365024.365119>
- Vilhjálmsón, H., Cantelmo, N., Cassell, J., Chafai, N. E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A. N., Pelachaud, C., Ruttkay, Z., Thórisson, K. R., van Welbergen, H., & van der Werf, R. J. (2007). The behavior markup language: Recent developments and challenges. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, & D. Pelé (Eds.), *Intelligent virtual agents* (pp. 99–111). Springer.
- Werning, M. (2020). Predicting the past from minimal traces: Episodic memory and its distinction from imagination and preservation. *Review of Philosophy and Psychology*, 11, 301–333. <https://doi.org/10.1007/s13164-020-00471-z>
- Yuan, J., Liberman, M., & Cieri, C. (2006). Towards an integrated understanding of speaking rate in conversation. In *Proceedings of INTERSPEECH* (pp. 541–544). Pittsburgh, Pennsylvania: International Speech Communication Association.
- Yuan, J., Liberman, M., & Cieri, C. (2007). Towards an integrated understanding of speech overlaps in conversation. In *ICPhS XVI*. The International Congress of Phonetic Sciences.

Cite this article: Lücking, A. & Ginzburg, J. (2023). Leading voices: dialogue semantics, cognitive science and the polyphonic structure of multimodal interaction *Language and Cognition* 15: 148–172. <https://doi.org/10.1017/langcog.2022.30>