Trends in Cognitive Sciences

CellPress REVIEWS

Opinion

Multimodal Language Processing in Human Communication

Judith Holler[1,2,*] and Stephen C. Levinson[1,3]

The natural ecology of human language is face-to-face interaction comprising the exchange of a plethora of multimodal signals. Trying to understand the psycholinguistic processing of language in its natural niche raises new issues, first and foremost the binding of multiple, temporally offset signals under tight time constraints posed by a turn-taking system. This might be expected to overload and slow our cognitive system, but the reverse is in fact the case. We propose cognitive mechanisms that may explain this phenomenon and call for a multimodal, situated psycholinguistic framework to unravel the full complexities of human language processing.

A Binding Problem at the Core of Language

Language as it is used in its central ecological niche – that is, in face-to-face interaction – is embedded in multimodal displays by both speaker and addressee. This is the niche in which it is learned, in which it evolved, and where the bulk of language usage occurs. Communication in this niche involves a complex orchestration of multiple **articulators** (see Glossary) and **modalities**: messages are auditory as well as visual, as they are spread across speech, nonspeech vocalizations, and the head, face, hands, arms, and torso. From the point of view of the recipient, this ought in principle to raise two serious computational challenges. First, not all bodily or facial movements are intended as part of the **signal** or content – the incidental but irrelevant movements must be set aside (we call this the segregation problem); second, those that seem to be part of the message have to be paired with their counterparts (as when we say 'There!' and point), and simultaneity alone turns out to be an unreliable cue (this is our binding problem). In this Opinion article, we ask how the multiple signals carried by multiple articulators and on different modalities can be combined rapidly to build the phenomenology of a coherent message in the temporally demanding context of conversational speech.

The Scope of the Problem

There is a huge literature on spoken language processing, but most looks only at the linguistic signal. The visual articulators (the nonvocal counterparts to tongue, lips, and velum) are many: the face alone contains 43 muscles, each hand is controlled by a further 34, and an abundance of muscles is involved in moving the arms, head, and torso. This makes evident the sheer amount, complexity, and variation of the articulators and the potential signals they produce when talking.

Moreover, these multiple layers of visual signals are offset in time rather than neatly aligned; we may blink and then nod while gesturing with the hands and the nod may be followed by a head tilt outlasting the manual gesture. All this is layered onto the lexical segments and prosodic boundaries constituting the vocal signal, resulting in a plethora of disaligned signal onsets and offsets. For example, lip movements tend to precede corresponding phonation by up to 100–300 ms in natural speech [1], referential gestures tend to precede corresponding lexical information by several hundred milliseconds to several seconds even [2–4] (see also [5,6]), with this timing being slightly tighter for pitch accents and kinematic points of emphasis [7]. However, these

Highlights

Multiple layers of visual (and vocal) signals, plus their different onsets and offsets, represent a significant semantic and temporal binding problem during face-to-face conversation.

Despite this complex unification process, multimodal messages appear to be processed faster than unimodal messages.

Multimodal gestalt recognition and multilevel prediction are proposed to play a crucial role in facilitating multimodal language processing.

The basis of the processing mechanisms involved in multimodal language comprehension is hypothesized to be domain general, coopted for communication, and refined with domain-specific characteristics.

A new, situated framework for understanding human language processing is called for that takes into consideration the multilayered, multimodal nature of language and its production and comprehension in conversational interaction requiring fast processing.

[1]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, The Netherlands
[3]Centre for Language Studies, Radboud University Nijmegen, Nijmegen, The Netherlands

*Correspondence: j.holler@donders.ru.nl (J. Holler).

multimodal signals are perceived as unified, synchronous percepts and are integrated effortlessly by the brain even at quite large temporal asynchronies [8–13]. One could argue that the difference in transduction time for visual and auditory stimuli (the pressure waves constituting sounds are translated into a perceived auditory image more quickly than the retinal receptors translate light into a perceived visual image) may account for part of this, but this is to some extent counteracted by differences in transmission time between the two modalities in the opposite direction (i.e., light travelling faster than sound) [14], and at a standard conversational seating distance it would not account for temporal lags of several hundred milliseconds.

These asynchronies constitute a crucial binding problem. Utterances unfold over time (with turns usually lasting from a few hundred milliseconds to several seconds in length [15,16]), with different visual signals appearing at various points throughout this duration (Figure 1). Moreover, these multiple, temporally offset signals need to be integrated not just in a pair-wise fashion at small scales (e.g., lips and phonemes, gesture and word) but across the different articulators and **levels** of processing, including holistically, at the message level – for that is what interlocutors in conversation have to respond to.

And this is not all. An additional layer of processing is required to segregate the movements that are meant to contribute to the message versus incidental or accidental motions (e.g., flicking one's hair, scratching the lower arm, or grasping a drink while talking). Further, signals that perform 'housekeeping' rather than message-level functions also need to be segregated (such as the return of gaze signaling the end of a speaking turn, or thankful nod and glance at the waiter bringing a drink while, in parallel, producing a multimodal message for one's interlocutor). Thus, a highly efficient parsing mechanism has to be in place to filter out the actions not relevant to message content and to bind the remaining signals into a unified message.

It is clear that such a flood of apparently disaligned signals may represent a significant challenge to the limited processing capacities of our cognitive system when trying to process communicative messages in face-to-face interaction, especially in the light of the very tight temporal constraints under which conversation operates: the average time that elapses between two turns is in the ballpark of 0–200 ms, with responses to questions being issued typically without any perceivable gap [17–19]. Long gaps are pragmatically meaningful and may be interpreted as reluctance to confirm or agree [20]. Given the latencies involved in speech production (over 500 ms), a speaker must start preparing his or her response while the incoming turn is still in progress [19]. Message comprehension thus has to be both fast and predictive and run partially in parallel with next-turn planning [21–23]. Taking these tight temporal constraints together with the multimodal binding problem might suggest that adding visual signal processing to auditory language comprehension would put significant extra strain on our processing system.

Surprisingly, exactly the opposite appears to be the case. When responding to questions that have an accompanying manual and/or head gesture, next speakers respond faster than to questions without such visual components [24]. This is in line with experimental evidence where participants respond faster to speech–gesture combinations than to their speech-only counterparts [25–28]. Thus, at first sight, human face-to-face communication confronts us with a puzzle and an apparent paradox: processing more signals simultaneously is faster than processing speech alone. This finding might be thought to resemble multisensory facilitation observable outside the domain of communication. For example, responses to a simple auditory stimulus (a single tone) combined with a simple visual stimulus (a light flash) are faster than responses to the unimodal equivalents [29–35] and objects are recognized more quickly when perceived multimodally (e.g., identifying a sheep by picture plus the sound it makes) [36,37]. In addition to many studies demonstrating multimodal enhancement in humans, we also find it in a range of

Glossary

Articulators: while traditionally the term 'articulator' denotes the vocal organs above the larynx (i.e., the tongue, lips, teeth, and hard palate), sign languages indicate that a broader definition is needed. We here use the term based on a multimodal perspective, thus defining 'articulator' as including the tongue, lips, and mouth as well as the head, the face including the forehead and eyebrows, the upper and lower eyelids, the muscles around the nose, cheeks, and mouth, the hands, arms, and shoulders, the upper torso, and, in principle, the lower torso, legs, and feet, although they tend to be less systematically used.
Gestalt: gestalt psychologists argue that the perceptual system organizes the stream of information that it encounters on the basis of a set of principles, or 'laws', to derive a meaningful pattern. The laws fuse elements from the information stream together so they emerge as unified percepts (i.e., gestalts) that stand out against the background of surrounding information. For example, the law of 'similarity' states that visual elements that share visual features are seen as belonging together. The law of proximity states that visual elements that are positioned closer together relative to surrounding objects are seen as belonging together; a temporal version might group simultaneous events. Gestalt recognition is often deemed a primary process [122] and considered to lead to more than the sum of its parts [123] – an independent percept in its own right. Originally, the gestalt laws were defined to explain visual perception, but they also apply to other domains of cognition, such as music perception (Box 2). Here, we use the term 'gestalt' to refer to signals grouped together at the perceptual level when associated with meaning at higher levels.
Levels: percepts are processed, it is generally agreed, through a hierarchy of representational levels; in language, for example, from the acoustics through the phonology and the morphology to the syntax and semantics. Signals carried by the individual articulators are bound together through both low- and high-level processing (Figure 2) and prediction happens within and across these levels of processing.
Modalities: information in human communication comes in many modalities or perceptual senses

nonhuman species (e.g. [38–41]). However, the multimodal integration process in these examples is considerably simpler than for multimodal utterances, involving low-level integration of nonsemantic stimuli in the case of flashes and sounds and, in the case of multimodal object recognition, a process of fusing a sound and an image based on a conceptual representation in which this stimulus combination is permanently fixed. Multimodal utterances, by contrast, are relatively unpredictable and complex in composition. In integrating a 'yes' with a nod – probably one of the simplest multimodal utterances – the visual and auditory signals are conventionalized and highly congruent, but this is not representative of the bulk of multimodal utterances (e.g., many manual gestures are not redundant but rather add semantic and pragmatic information [42]). Moreover, utterances allow speakers to express an infinite number of thoughts and ideas, resulting in an unparalleled degree of potential ambiguity. Here, we try to unravel the complexities underlying multimodal utterance processing and propose possible cognitive mechanisms that may underpin it.

Potential Mechanisms Underpinning the Binding of Multimodal Language

How is the binding problem resolved in the tight time frames allowed in conversation? Some **gestalt**-like principles – that is, fast integration of stimuli that 'make sense' together and are recognized as holistic percepts – would seem to be involved. The essential prerequisite for producing prompt relevant responses is deriving a holistic message corresponding to a whole turn at talk: parsing mechanisms operating in a gestalt way must integrate information hierarchically, allowing us to bind separate units of information together at increasing levels of complexity and to distinguish overall figure from ground. These mechanisms may have domain-general roots (Box 1).

Gestalts and Stable Form-Meaning Mappings

Such a gestalt-based mechanism crucially presupposes statistical regularities in the co-occurrence of multimodal signals together with the communicative meanings that the whole ensemble is intended to convey, and these regularities must outweigh idiosyncratic message encoding to be effective. There is qualitative and quantitative evidence for this: statements expressing negation in conversation may be accompanied by the 'not face' (a complex facial expression involving a combination of muscle movements typically characterizing the facial expressions of anger, disgust, and contempt) [43,44]. The 'facial shrug' comprises an eyebrow flash with one corner of the mouth being retracted (and sometimes additional components [45]) and has been ascribed the function of signaling 'I don't know', 'Oh well', or 'OK' (i.e., functions similar to the shoulder shrug) [44,46]. Raised eyebrows often function as question markers and signals of non-understanding in spoken [47–51] as well as in many signed [52,53] languages. The so-called thinking face (withdrawing gaze, with furrowed or raised brows) has been shown to act as a pragmatic marker and signal of delay [44,46,54]. Similarly, smiles can comment on descriptions of events as being ironic, humorous, or sarcastic [46,55,56]. Also, surprisingly, even the subtle movements involved in blinking seem to fulfil important pragmatic functions, such as signaling understanding and the sufficiency of information provided [57].

Further, we find consistent patterns also when looking beyond the face, such as in forward-leaning body postures accompanying questions (in many sign languages [53] and in spoken language [58]) and in the form of certain manual gestures, especially those defined by specific kinematic and formational (shape, orientation) parameters – so-called gesture families (e.g. [42,59,60]). For example, the 'Open Hand Supine' gesture (palm facing up, extended into frontal gesture space) is used by speakers metaphorically to offer, give, present, or receive something, such as a point, opinion, or explanation or, when combined with a movement towards the interlocutor, it is used to acknowledge something as having been contributed by the interlocutor earlier or inviting

(olfactory, haptic, visual, auditory). Here, the term 'modalities' refers primarily to visual and auditory signals, but touch, as in handshakes, embraces, and so forth, is clearly important.

Multimodal gestalts: at higher levels of processing, multiplex signals (see below) form multimodal gestalts, meaning-bearing performances, that are likely to be identified through both bottom-up and top-down processes.

Multiplex signals: signals from the individual articulators are bound together, or unified, into complex multimodal signals at lower (presemantic) levels of processing. Here, we use the term 'multiplex signal' for groupings formed at the perceptual level and the term 'gestalt' only to refer to bound elements associated with meaning at the semantic/pragmatic level (Figure 2).

Signals: a behavior that, pretheoretically, makes a meaning difference without itself necessarily having a full intended meaning, like a phonemic contrast between p and b (as in pin vs bin) or an extended index finger. A single articulator can carry different kinds of signals, so that, for example, the eyes can shift gaze, blink, squint, or wink and the hands perform different kinds of manual movements.

Social action: the 'social move' or 'speech act' (but here considered multimodally) that an utterance performs in conversation. An utterance expresses an individual's intention, not only by conveying meaning but by performing a social action, such as requesting, inviting, informing, criticizing, teasing, and so forth.

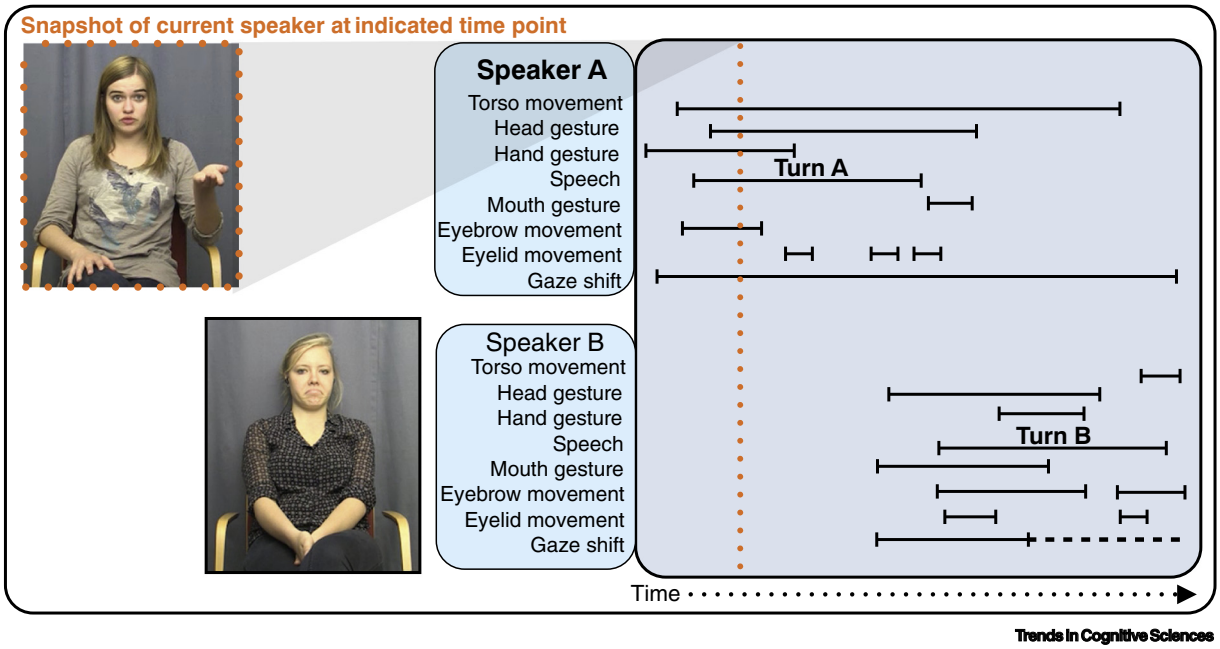

Figure 1. Visual Communicative Behaviors Distributed across Articulators and Time. The two light-blue boxes in the middle give a list of potential communicative behaviors produced by two interlocutors in interaction (persons A and B). The black lines on the righthand side aligned with these entries indicate the hypothetical occurrence of behaviors produced as part of conversational turns produced by speakers A and B along a temporal axis (time flows from left to right, indicating the unfolding of the utterances). The upper-left image (encased by the orange broken line) captures a snapshot of person A producing her conversational turn at the time point indicated by the orange broken vertical line in the image of the unfolding multimodal utterance (right). The bottom image captures a snapshot of person B producing addressee feedback at the same time point. This is followed by person B producing the next conversational turn as a response to turn A (but occurring somewhat in overlap, at least with regard to the visual signals produced by persons A and B), which can be seen to the right of the orange broken line.

a response [42,60–63]. Combined with a lateral movement of the hands away from the speaker's midline, the palm-presenting gesture is often used in contexts that relate to the 'absence of knowledge' [63,64], such as in interrogative contexts when posing proper or rhetorical questions [42,53,60,64]. Another specimen is the 'Away Gesture Family' [65,66] where the hand is moved away from the speaker, to sweep, brush, hold, or throw away an imaginary entity located in frontal gesture space to express ideas of rejection, refusal, negative assessment, or negation. A further stable form-meaning mapping can be found in the 'cyclic gesture' [67] characterized by a continuous circular movement of the hand performed away from the body to express the notion of cyclic continuity, such as in enumerations or word and concept searches.

Even the gestures of young children can fulfil speech-act-like functions, such as requesting (reaching, pointing) and offering (holding out hand with object) [68–70], with some parallels in the behavior of nonhuman primates [71,72] suggesting that these **social actions** may be deeply rooted in our evolutionary history. An observation that resonates with this idea is that quite a few of the above-mentioned visual form-meaning mappings occur across different cultures [43,47,50,53,64,66], pointing to the possibility of a common biological origin. While more systematic, large-scale, quantitative studies on conversational corpora are needed, there is already some convincing evidence for systematic associations between specific bodily signals (or combinations of several signals) and conversational social actions, representing the basis for efficient gestalt recognition.

Temporal Organization and Prediction

In the context of gestalt recognition, the distribution of signals across the different articulators and their off-set temporal organization may actually be quite beneficial to efficient communication

Box 1. Domain-General Mechanisms Foundational to Multimodal Language Processing?

Comprehending multimodal utterances in conversation may rest on at least two processes that may have a domain-general basis: low-level multisensory integration and higher-level gestalt recognition.

Multimodal Low-Level Integration as a General Cognitive Principle

The brain's tendency to combine information from different modalities is illustrated by various multimodal illusions; for example, the McGurk effect (where lip movement suggesting one sound and acoustics suggesting another are blended into a third sound [124]) and the ventriloquism illusion (whereby sounds coming from one direction are fused with lip movements from another). Considering that we are equipped with multisensory neurons deep in the superior colliculus [125] and that we find those effects not only in human adults but also in young infants, our brain might be held to be 'integration ready' by its very nature (subject only to the maturation of the relevant neural pathways). This makes sense from an evolutionary perspective: integrating sight and sound of predator or prey into a single percept would be beneficial. In addition, activating parallel pathways that can race one another and later converge can account for multisensory facilitation [29–35]. This relatively 'simple' integration mechanism might be expected, then, to be domain general and may be argued to be the necessary foundation for multimodal integration in human face-to-face communication.

Top-Down Prediction as a Domain-General Process

To respond to message content in the tight timing of conversation, both multimodal bottom-up and powerful top-down processes must be involved. On the one hand, we know that multimodal signal integration occurs on a time scale indicative of immediate low-level integration [126], applying also to speech-related stimuli [127]. On the other hand, critical top-down information seems to be predictive at higher processing levels. Studies of the attribution of overall message content, or social action, in conversation support this by pointing to the projective power of speaking turns, which creates a strong expectation for the social action of the next turn, such as an invitation projecting a response in the form of an acceptance or a declination [128] (although these studies focus on unimodal speech). Similarly, strong top-down processes from scene-context knowledge are assumed to influence the perception of visual scenes and the recognition of individual scene components outside the domain of communication ([129]; and they may combine with gestalt-like perceptions also found at lower levels [130]), thus providing another potential source of domain-general mechanisms that may underpin multimodal utterance processing (see [131] for similar parallels in multisensory perception and social domains of metacognition).

In sum, potentially domain-general bottom-up and top-down processes (lower-level integration and higher-level contextual, holistic processing) may run in parallel, occur incrementally, and interact with one another in the comprehension of multimodal utterances. However, the speed of processing required by conversation, combined with the complexity of semantic and pragmatic processing, the unpredictability of novel messages in multimodal communication, and the frequently complementary rather than redundant mappings of signals that are perceived as multimodal gestalts, may also involve specialist mechanisms (see Outstanding Questions).

rather than a challenge to fast processing. Recent research emphasizes that prediction is fundamental to cognition not only when we perceive, navigate, and interact with the world around us [73–75] but also when we process verbal language [76–79] as well as auditory scenes of other kinds, including music (Box 2). We suggest that the temporally disaligned and distributed signals across different articulators and modalities may facilitate predictive language processing in face-to-face communication. On this account, prediction happens on different time scales, covering both shorter and longer time windows, as well as on different levels (Figure 2), thus building on recent predictive coding accounts developed for verbal language (e.g. [80]). In the following, we will unpack the processes we hypothesize to be involved.

Streams of discontinuous signals produced by the various articulators are bound into **multiplex signals** based on our experience of statistical associations between the signals in everyday language use (e.g., 'X is typically followed by Y in conjunction with Z') without hierarchical structure coming into play at this stage: any signal may prime and be bound with any other. Prediction at this level is thus based on stimulus-induced bottom-up processes resting on statistical correlations and operating mostly within shorter time windows of the unfolding multimodal utterance. For example, a frequent combination of signals may involve a specific lip formation predicting the articulation of a certain sound that a speaker is about to utter, occurring together with a

Box 2. Binding and Prediction in Multimodal Language and Music

How does our perceptual system organize the stream of information it is exposed to in face-to-face communication into meaningful units? The mechanism we propose for multimodal utterance processing bears both similarities and differences to the mechanisms proposed for auditory scene analysis and music perception. Two kinds of binding principles have been identified for auditory scene analysis [132], both of which can be related to gestalt perception.

The first entails the grouping of perceived sounds on the basis of gestalt laws; for example, in music perception sounds may be grouped together when they are perceived as similar or proximal in terms of some acoustic feature such as pitch or timbre [132,133] or when they follow one another in closer succession than neighboring sounds, meaning that long distances and pauses create the impression of grouping boundaries [134]. With regard to multimodal utterances, cross-modal grouping based on perceptual similarity may not be straightforward, since the visual and verbal modalities are qualitatively different. Even binding of individual signals within the visual modality would be difficult based on perceptual similarity alone, since the signals vary hugely in form, size, and where they are articulated in space (however, grouping based on the repetition of identical movements may be more feasible, such as for successive nods or repeated cyclic finger movements depicting 'rolling'). Temporal proximity, however, may be one of the principles on which grouping is based, such as co-occurring prosodic and movement emphasis converging in the perception of linguistic stress or co-occurring facial movements combining into a more complex facial expression. However, not communicatively intended movements (e.g., a scratch) also co-occur with speech and other visual signals, meaning that grouping by temporal proximity is not sufficient. Statistical associations are likely to account for most of the process of low-level grouping in multimodal utterance processing, possibly interacting with some perceptual principles such as temporal proximity and possibly similarity in form.

The second binding principle for auditory scene analysis is sequential predictability. It has been suggested that sounds are stored linked to sounds that preceded them, thus constituting perceptual units or 'auditory object representations' [132], providing the basis for sequential prediction. When listening to music, the gestalt law of 'good continuation' facilitates the expectation of particular next sounds, chords, or harmonies [134]. If we consider that utterances unfold over time, where stable, form-meaning mappings may comprise signal sequences that span the length of an utterance, the parallel is quite obvious. The beginning of such a sequence may give rise to the expectation of the next signal in the stored sequence and so forth. However, an extra layer of processing in the form of top-down semantic analysis and pragmatic inferencing is likely to be required to allow us to cope with the complexity of binding, gestalt perception, prediction, and meaning conveyed in conversation. These processes still require much future investigation.

brow raise and gaze towards the interlocutor, followed by the hand being lifted with the palm facing upwards just a few milliseconds later. Thus, particular combinations of signals that frequently occur in close temporal proximity are perceived as unified, multiplex signals at the lowest level of processing and, again by means of statistical association, they can prime subsequent multiplex signals (note that what we have termed 'multiplex signals' may be referred to as 'gestalts' elsewhere, whereas we reserve this term for multiplex signals associated with meaning).

Layered on top of this is the prediction of gestalt-like configurations at different semantic levels of the message, including words, sentential units, meaningful gestures, syntactical structures, and social actions. Single or multiplex signals feed upwards, shaping predictions at these higher levels. For example, the lips shaping to produce a *w*-sound may restrict the search space for predictions about upcoming words to a phonetically congruent set of candidates. This candidate set may be pruned further by other components of the multiplex signal in which this lip formation is embedded, such as raised eyebrows and a lifted palm-up open hand, since at the social action level this **multimodal gestalt** may trigger the prediction of a question being produced. This prediction, in turn, will then feed downwards, thus further raising the anticipation of a *wh*-word being uttered, plus a question-typical syntactical structure, and so forth, finally influencing, top-down, the binding of signals from the individual articulators at the lowest level. The bottom-up–top-down interaction between levels is a continuous, dialectic process, leading to immediate, incremental unification while the utterance unfolds [81].

Cross-level prediction is a core component of the process of situated language comprehension and must be substantially facilitated by the formation of multiplex signals and multimodal gestalts at the different levels. One crucial aspect of this prediction process is that visual signals

often precede corresponding vocal or verbal signals. Based on this typical temporal architecture, in conjunction with the predictive binding principles outlined above, visual signals can serve a channeling function in the parsing of the linearly unfolding speech stream, at many hierarchical levels. What we suggest, then, is a hypothesis about how multimodal language may be processed in face-to-face interaction. Experimental findings demonstrating that mouth movements can facilitate the prediction of upcoming sounds [9,82,83] and that manual gestures can prime subsequent words and semantic concepts are in line with this hypothesis [11,84,85].

This channeling of predictive processes by the visual modality resonates with constraint-based models of language processing [86,87], where information processing is immediate and incremental, with alternative sentence interpretations running in parallel. One dynamic, probabilistic model of language processing is particularly compatible with our framework; namely. a recent version of the Sentence Gestalt model [88]. Here the presence of visual bodily signals would form part of the input that shapes the meaning representation of an utterance by updating probabilistic activations of possible next elements in the unfolding utterance. Visual bodily signals thus may reduce uncertainty at the message level, continuously contributing to changing activation patterns (estimates of conditional probabilities) underlying the meaning representation conveyed by an unfolding utterance. A similar notion is captured by the NOLB model accounting for the neural basis of language processing in context (including visual signals), with context reducing ambiguity and facilitating prediction of the upcoming language input [89].

Constantly updated unification of this kind would require a multimodal semantic memory buffer to keep parallel representations active just long enough in case reanalysis is required. In a turn-taking context, there comes a point when the comprehender must commit to one of those activations to begin planning a matching response that can be issued on time, and this point appears to occur surprisingly early in the incoming turn [22,90] (but see [23]). This early commitment is probably greatly aided by multiplex signals and multimodal gestalts, since they should enhance the fidelity of the predictions we make thus reducing the likelihood of prediction errors and increasing the differences in activation between alternative meaning representations. Despite this, reanalysis of an utterance's meaning must remain possible if further incoming signals require it, hence the need for multimodal buffers as suggested above.

Such a memory buffer is also required by the segregation problem due to the need to set aside articulator movements that do not seem to be communicative at first sight. Unification of signal meaning operates on the basis of 'all the news that fits', resulting in orphaned movements. However, if signals encountered downstream change the utterance interpretation, earlier segregation decisions may have to be revised leading to the integration of previously orphaned behaviors. Reanalysis of this kind is no doubt costly but may be necessary from time to time due to the speed of the binding process required by conversational turn-taking and perhaps also the use of fast and frugal heuristics in the segregation process itself. A memory buffer for parallel activations and segregated behaviors is in line with constraint-based models such as [88] but contrasts with the idea of the 'now or never bottleneck' [91] – some ability to recover and reanalyze is shown by details of the interactional repair mechanism, where, in response to an interlocutor's clarification request, speakers demonstrate the ability to revisit utterances they produced earlier to then repeat or revise them [92].

So far, then, we have argued that the compositional and temporal architecture of multimodal utterances facilitates predictive coding on multiple levels, leading to a processing advantage over unimodal utterances. Further, the rapid chunking into multiplex signals and multimodal gestalts may free cognitive resources [91]. Additional processes no doubt play a role, such as those

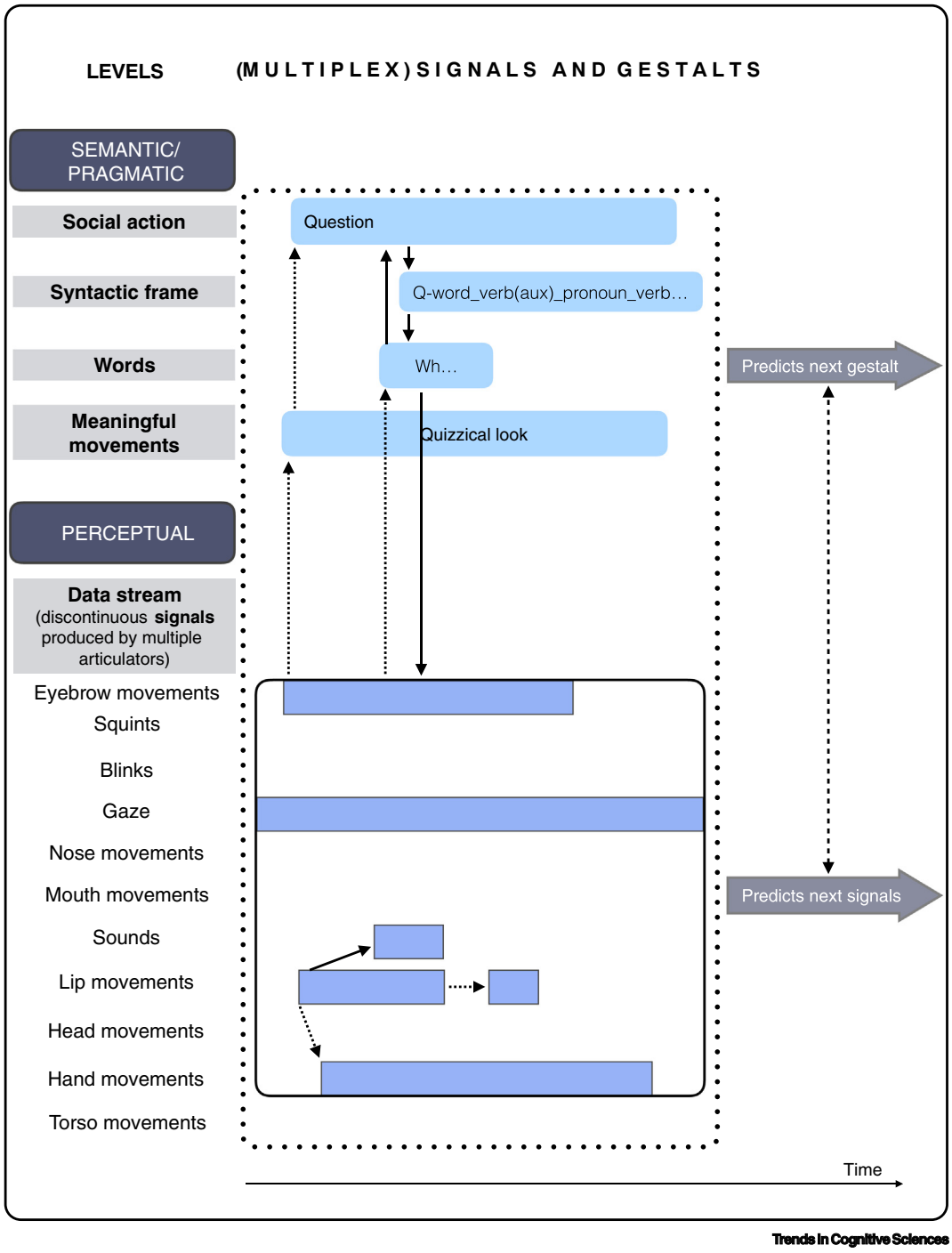

Figure 2. Multimodal Utterance Binding and Prediction across Levels. Examples of predictions are indicated through black arrows (broken black arrows indicate weaker predictions than unbroken black arrows). In the present example, the occurrence of an eyebrow movement is bound with signals occurring in close temporal proximity (a lip movement and a hand gesture) based on statistical association; this ensemble of visual signals alone is predictive of questionhood at the speech-act level (indicated by the black broken arrow pointing up towards the social-action level). Individual signals of this first ensemble predict further signals based on statistical association, such as a particular sound (e.g., w-) matching the first lip movement. This sound, especially in conjunction with the preceding visual signal ensemble, is

(Figure legend continued at the bottom of the next page.)

involved in the perception of biological motion, which may increase attention or alertness and thus the efficiency of cognitive processing.

A Language-*In-Situ* Framework for Understanding Human Language Processing

These observations suggest the need for a radical rethink of the cognitive processing involved in human communication: multimodal signaling in a dialogic setting involves an order of complexity missing from most extant models of human language processing. Efforts have been made to understand dialogic processing in recent years, but most of this still neglects multimodal aspects of language (e.g. [15,22,79,93–96]). However, progress has been made in understanding the multimodal processing involved in pairings of words and gestures (e.g., production [97–100], comprehension [26]) but rarely in a fully dialogic setting. Further, facial signaling has been largely neglected, and where examined mostly in the context of emotion research (but see [101–104] for exceptions).

Meanwhile, in psycholinguistics, the focus remains on cognitive processing in the individual in isolation listening to or producing unimodal speech, the rationale being that this constitutes a modular element of the larger picture (see [89,105] for similar points in making a case for contextualized neurobiological models allowing us to capture the neural basis of natural language processing). However, if our argument above is correct, multimodal processing influences linguistic processing (e.g., by cross-priming linguistic and gestural elements), so that unimodal language processing may involve rather different processes (ranging from subtle differences in the speed of processing to potential differences in propositional or speech act comprehension).

Elsewhere we have argued that the human capacity for structured social interaction provides the foundation for face-to-face communication [106]. This natural predisposition can be seen already in newborns and even fetuses showing strong social sensitivities [107–109] and it is evident in adulthood in neuronal activity that favors face-to-face over back-to-back interactions [110]. Our cognitive predisposition for engaging in social face-to-face interaction is matched with human bodies optimally equipped by evolution for multimodal communication: our hands have been freed for gesturing through bipedal locomotion; the white sclera and the much darker pupil of the eye render gaze direction easily detectable [111]; the fine motor control of the face and its 43 muscles as well as the fine orchestration of the multitude of muscles controlling the hands, arms, head, and neck; and, finally, the lack of facial and bodily hair renders even small muscular movements visible and communicatively effective, including squints and blinks of the eyes [57].

The possibility that multimodal and unimodal utterances may in part involve different psycholinguistic processing, and that the human communication system seems to be built for multimodal social interaction, highlights the need for a language-*in-situ* framework to allow us to answer new fundamental questions about human language processing (see Outstanding Questions). Such a processing framework (Figure 3) would need to incorporate the notion of language as a joint action, involving a speaker and an active addressee [112], and to recognize that the tight time constraints of conversation affect psycholinguistic processing in conversation, requiring some degree of parallel processing for next-turn response planning [19]. Crucially, it would assume that when taking a turn in face-to-face conversation, person A produces a stream of multimodal behavior involving many bodily articulators and a host of multiplex visual–auditory signals (including not

predictive of a wh-word, which strengthens the prediction of questionhood (hence the unbroken black arrow between 'wh-word' and 'question'). The prediction of a question is also predictive of a typical interrogative syntactic frame (first black unbroken downward arrow), which in turn strengthens the prediction of a wh-word, and this downward cascading set of predictions enhances the binding of the individual, incrementally occurring signals into multiplex signals (indicated by the unbroken-lined black square). As the utterance unfolds, the multiplex signal and its processing on higher, semantic and pragmatic levels results in the perception of a multimodal gestalt (broken-lined black square). Both the multiplex signal and the multimodal gestalt prime the next upcoming signals and gestalts.

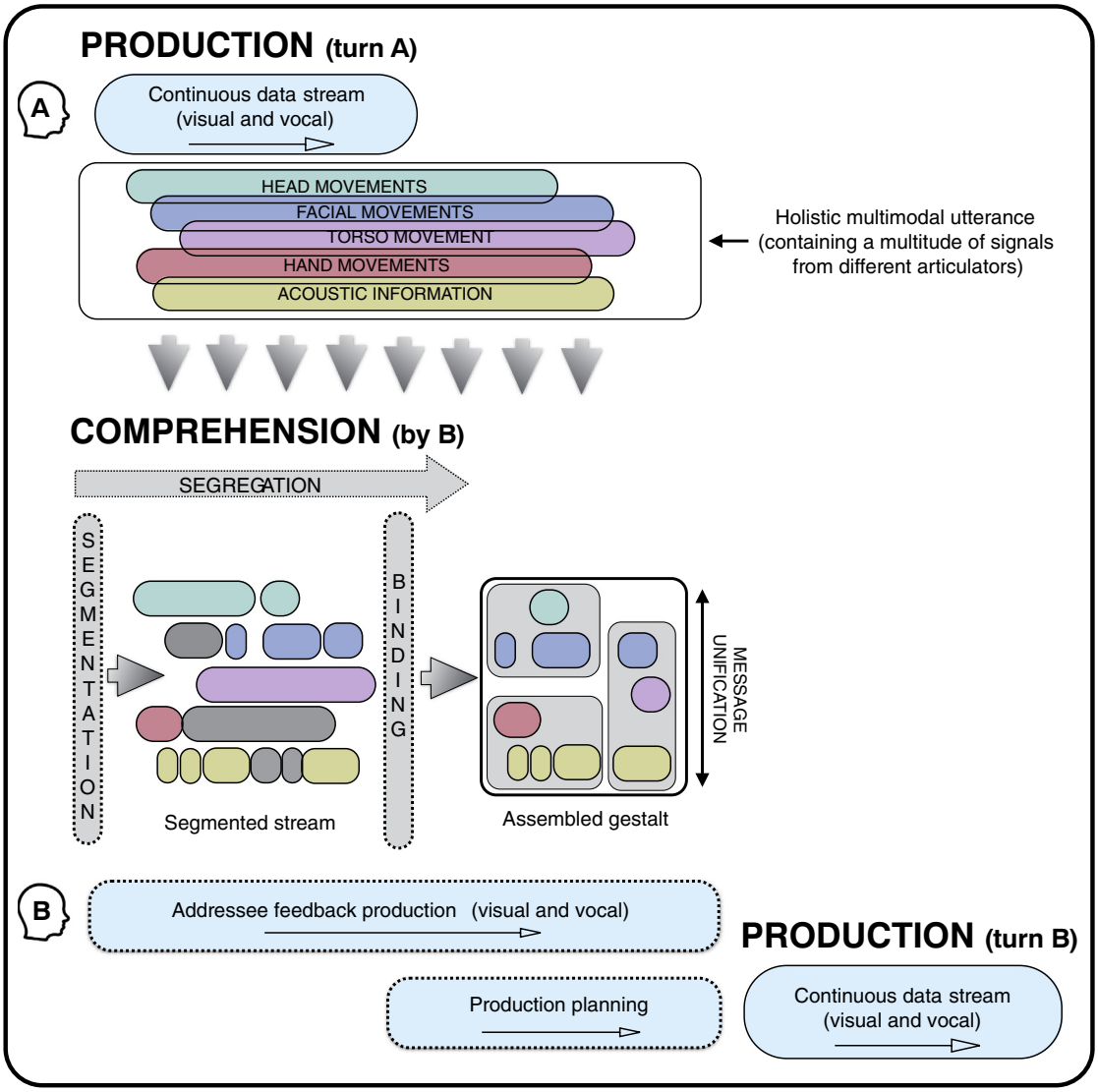

Figure 3. Multimodal Language Processing in Interactional *Situ*. Time flows from left to right and the parallel nature of the production of one conversational turn (by person A) and the comprehension of this turn by the interlocutor (person B), plus the production of addressee feedback and next-turn planning (by person B), are depicted, finally leading to the production of the next conversational turn (by person B). The turn that person A produces comprises a multimodal behavior stream containing signals emitted by the different articulators. This behavior stream is segregated into communicative movements and those that do not form part of the message. The remaining signals are bound into multiplex signals and multimodal gestalts. When the process of binding and gestalt assembling has begun, a pragmatically matching response is being planned. (Although depicted as a serial process, segregation and binding run in parallel while the utterance unfolds.)

only bodily movements and verbal content, but also intonation, breathing, pausing, and so forth). Person B parses this stream of information according to the mechanisms of binding and segregation proposed above, resting significantly on gestalt recognition and prediction of the unfolding turn and resulting in the interpretation of a unified, holistic message. However, parallel probabilistic activations of different message interpretations need to be maintained during this process so that earlier binding and segregation decisions can be revised in case A's turn should unfold in unpredicted ways.

This proposed framework capturing situated language processing in a face-to-face, dialogic environment illustrates the complexity of the cognitive and behavioral processes involved. We have here sketched the leanest form of dialogic interaction; further complexity is added by conversations often being embedded in parallel activities (e.g., cooking, hairdressing), taking place in noisy or busy environments, which can obscure information in the auditory and visual channels, and involving more than two interlocutors, which requires negotiation of who speaks when, the monitoring of and responding to multiple recipients' signals of understanding or lack thereof, and the design of utterances that take into account differences in knowledge status, common ground, etc. [112–114]. A psycholinguistic model that aims to fully capture how we process language *in situ* needs to account for these complexities.

We have focused here on comprehension, but the language-*in-situ* framework proposed above also incorporates multimodal utterance production. There has been much work suggesting that language production and comprehension are two aspects of a shared system [79,115–117]. It has even been suggested that we use analysis by synthesis, with language production forming the basis of predictions during comprehension [79,118], but this may cause difficulties for the need to perform production planning of a response well in advance during conversational turn-taking [92]. However, processes may be shared regarding some components of the processes of binding and orchestration. Production entails the programming of hierarchically structured behaviors comprising multiplex signals, requiring the selection of the signal components, the generation of different timestamps for their various on- and offsets, and withholding their execution until called on [119,120]. In 1951, Lashley [121] argued for parallel activation of all of the components contained in a behavior chain, overlaid by an 'action schema' functioning as the ordering system. One possibility is that such action schemata function as cognitive templates for both the perception and the production of the sort of multimodal gestalts we have here hypothesized.

Concluding Remarks

Fast multimodal integration, gestalt recognition, and prediction are core to the survival of most species. The cognitive mechanisms that enable these processes are thus likely to be phylogenetically ancient and may have been coopted for human communication over the course of evolutionary history, providing us with a natural predisposition for multimodal social interaction. Here, we have argued that this predisposition is core to the phenomenon of multimodal language processing and have proposed a framework that embraces this notion and cognitive mechanisms that may scaffold how we process multimodal language *in situ*.

Outstanding Questions

How similar or different is the *in situ* processing of unimodal versus multimodal utterances in terms of the psycholinguistic, cognitive, and neurobiological processes that underpin it?

How, precisely, do multiplex signals influence the language comprehension process? Does their binding lead to the recognition of multimodal gestalts? Do these multiplex signals and multimodal gestalts facilitate prediction during language processing?

Is the temporal organization of multiplex signals an inherent feature of the multimodal gestalts that they may form? Are the individual signals that constitute them temporally organized such that they facilitate fast language processing?

How does the integration of signals from the visual and auditory modalities compare with audiovisual integration outside the domain of communication? How domain general or domain specific is multimodal integration during language processing?

Is it justified to assume that multimodal processing is the default mode for human language comprehension? Is unimodal language processed more slowly because part of the message is missing, thus incurring extra costs?

What roles do cultural specificity and cultural universality in multimodal signal composition play in language processing? Can cultural universals in multimodal gestalt recognition be explained in terms of some form of 'innateness'? Are such gestalts recognized more easily and thus processed faster than others?

Acknowledgments

We thank two anonymous reviewers for their invaluable comments on an earlier version of this manuscript. We also thank the Max Planck Gesellschaft for their financial support of J.H. and S.C.L. as well as the European Research Council for their support through grants #773079 and #269484, respectively.

References

1. Chandrasekaran, C. *et al.* (2009) The natural statistics of audiovisual speech. *PLoS Comput. Biol.* 5, e1000436
2. Morrel-Samuels, P. and Krauss, R.M. (1992) Word familiarity predicts temporal asynchrony of hand gestures and speech. *J. Exp. Psychol. Learn. Mem. Cogn.* 18, 615–622
3. Bergmann, K. *et al.* (2011) The relation of speech and gestures: temporal synchrony follows semantic synchrony. In *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction (GeSpIn 2011)*, pp. 3–8, University of Bielefeld
4. Kok, K.I. (2017) Functional and temporal relations between spoken and gestured components of language. *Int. J. Corpus Linguist.* 22, 1–26
5. Schegloff, E.A. (1984) On some gestures' relation to talk. In *Structures of Sound Action: Studies in Conversation Analysis* (Atkinson, J.M. and Heritage, J., eds), pp. 266–296, Cambridge University Press
6. Mondada, L. (2018) Multiple temporalities of language and body in interaction: challenges for transcribing multimodality. *Res. Lang. Soc. Interact.* 51, 85–106
7. Wagner, P. *et al.* (2014) Gesture and speech in interaction: an overview. *Speech Commun.* 57, 209–232
8. Kelly, S.D. *et al.* (2004) Neural correlates of bimodal speech and gesture comprehension. *Brain Lang.* 89, 253–260
9. van Wassenhove, V. *et al.* (2007) Temporal window of integration in auditory–visual speech perception. *Neuropsychologia* 45, 598–607
10. Wu, Y.C. and Coulson, S. (2005) Meaningful gestures: electrophysiological indices of iconic gesture comprehension. *Psychophysiology* 42, 654–667
11. Wu, Y.C. and Coulson, S. (2007) Iconic gestures prime related concepts: an ERP study. *Psychon. Bull. Rev.* 14, 57–63

12. Bernardis, P. *et al.* (2008) Behavioural and neurophysiological evidence of semantic interaction between iconic gestures and words. *Cogn. Neuropsychol.* 25, 1114–1128
13. Habets, B. *et al.* (2010) The role of synchrony and ambiguity in speech–gesture integration during comprehension. *J. Cogn. Neurosci.* 23, 1845–1854
14. Recanzone, G.H. (2009) Interactions of auditory and visual stimuli in space and time. *Hear. Res.* 258, 89–99
15. Levinson, S.C. and Torreira, F. (2015) Timing in turn-taking and its implications for processing models of language. *Front. Psychol.* 6, 731
16. Roberts, S.G. *et al.* (2015) The effects of processing and sequence organization on the timing of turn taking: a corpus study. *Front. Psychol.* 6, 509
17. Stivers, T. *et al.* (2009) Universals and cultural variation in turn-taking in conversation. *Proc. Natl. Acad. Sci. U. S. A.* 106, 10587–10592
18. Heldner, M. (2011) Detection thresholds for gaps, overlaps, and no-gap–no-overlaps. *J. Acoust. Soc. Am.* 130, 508–513
19. Levinson, S.C. (2016) Turn-taking in human communication – origins and implications for language processing. *Trends Cogn. Sci.* 20, 6–14
20. Bögels, S. *et al.* (2015) Never say no… How the brain interprets the pregnant pause in conversation. *PLoS One* 10, e0145474
21. Boiteau, T.W. *et al.* (2013) Interference between conversation and a concurrent visuomotor task. *J. Exp. Psychol. Gen.* 143, 295
22. Bögels, S. *et al.* (2015) Neural signatures of response planning occur midway through an incoming question in conversation. *Sci. Rep.* 5, 12881
23. Sjerps, M.J. and Meyer, A.S. (2015) Variation in dual-task performance reveals late initiation of speech planning in turn-taking. *Cognition* 136, 304–324
24. Holler, J. *et al.* (2018) Processing language in face-to-face conversation: questions with gestures get faster responses. *Psychon. Bull. Rev.* 25, 1900–1908
25. Holle, H. *et al.* (2008) Neural correlates of the processing of co-speech gestures. *Neuroimage* 39, 2010–2024
26. Kelly, S.D. *et al.* (2010) Two sides of the same coin: speech and gesture mutually interact to enhance comprehension. *Psychol. Sci.* 21, 260–267
27. Nagels, A. *et al.* (2015) Feeling addressed! The role of body orientation and co-speech gesture in social communication. *Hum. Brain Mapp.* 36, 1925–1936
28. Wu, Y.C. and Coulson, S. (2015) Iconic gestures facilitate discourse comprehension in individuals with superior immediate memory for body configurations. *Psychol. Sci.* 26, 1717–1727
29. Murray, M.M. *et al.* (2001) Visuo-spatial neural response interactions in early cortical processing during a simple reaction time task: a high-density electrical mapping study. *Neuropsychologia* 39, 828–844
30. Teder-Sälejärvi, W.A. *et al.* (2002) An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Cogn. Brain Res.* 14, 106–114
31. Molholm, S. *et al.* (2002) Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cogn. Brain Res.* 14, 115–128
32. Talsma, D. and Woldorff, M.G. (2005) Selective attention and multisensory integration: multiple phases of effects on the evoked brain activity. *J. Cogn. Neurosci.* 17, 1098–1114
33. Senkowski, D. *et al.* (2006) Oscillatory beta activity predicts response speed during a multisensory audiovisual reaction time task: a high-density electrical mapping study. *Cereb. Cortex* 16, 1556–1565
34. Romei, V. *et al.* (2007) Occipital transcranial magnetic stimulation has opposing effects on visual and auditory stimulus detection: implications for multisensory interactions. *J. Neurosci.* 27, 11465–11472
35. Moran, R.J. *et al.* (2008) Changes in effective connectivity of human superior parietal lobule under multisensory and unisensory stimulation. *Eur. J. Neurosci.* 27, 2303–2312
36. Molholm, S. *et al.* (2004) Multisensory visual–auditory object recognition in humans: a high-density electrical mapping study. *Cereb. Cortex* 14, 452–465
37. Suied, C. and Viaud-Delmon, I. (2009) Auditory–visual object recognition time suggests specific processing for animal sounds. *PLoS One* 4, e5256
38. Roberts, J.A. *et al.* (2007) Consequences of complex signaling: predator detection of multimodal cues. *Behav. Ecol.* 18, 236–240
39. Gingras, G. *et al.* (2009) The differing impact of multisensory and unisensory integration on behavior. *J. Neurosci.* 29, 4897–4902
40. Uetz, G.W. *et al.* (2009) Multimodal communication and mate choice in wolf spiders: female response to multimodal versus unimodal signals. *Anim. Behav.* 78, 299–305
41. Ręk, P. (2018) Multimodal coordination enhances the responses to an avian duet. *Behav. Ecol.* 29, 411–417
42. Kendon, A. (2004) *Gesture: Visible Action as Utterance,* Cambridge University Press
43. Benitez-Quiroz, C.F. *et al.* (2016) The not face: a grammaticalization of facial expressions of emotion. *Cognition* 150, 77–84
44. Chovil, N. (1991) Discourse-oriented facial displays in conversation. *Res. Lang. Soc. Interact.* 25, 163–194
45. Ekman, P. (2009) *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage,* Norton
46. Bavelas, J. and Chovil, N. (2018) Some pragmatic functions of conversational facial gestures. *Gesture* 17, 98–127
47. Ekman, P. (1979) About brows: emotional and conversational signals. In *Human Ethology* (Cranach, M., *et al.*, eds), pp. 169–249, Cambridge University Press
48. Borràs-Comes, J. and Prieto, P. (2011) 'Seeing tunes'. The role of visual gestures in tune interpretation. *Lab. Phonol.* 2, 355–380
49. Crespo Sendra, V. *et al.* (2013) Perceiving incredulity: the role of intonation and facial gestures. *J. Pragmat.* 47, 1–13
50. Enfield, N.J. *et al.* (2013) Huh? What? A first survey in twenty-one languages. In *Conversational Repair and Human Understanding* (Hayashi, M., *et al.*, eds), pp. 343–380, Cambridge University Press
51. Borràs-Comes, J. *et al.* (2014) Audiovisual correlates of interrogativity: a comparative analysis of Catalan and Dutch. *J. Nonverbal Behav.* 38, 53–66
52. Pfau, R. and Quer, J. (2010) Nonmanuals: their grammatical and prosodic roles. In *Sign Languages* (Brentari, D., ed.), pp. 381–402, Cambridge University Press
53. Zeshan, U. (2004) Interrogative constructions in signed languages: crosslinguistic perspectives. *Language* 80, 7–39
54. Goodwin, M. and Goodwin, C. (1986) Gesture and coparticipation in the activity of searching for a word. *Semiotica* 62, 51–76
55. González-Fuente, S. *et al.* (2015) Gestural codas pave the way to the understanding of verbal irony. *J. Pragmat.* 90, 26–47
56. Caucci, G.M. and Kreuz, R.J. (2012) Social and paralinguistic cues to sarcasm. *Humor* 25, 1–22
57. Hömke, P. *et al.* (2018) Eye blinks are perceived as communicative signals in human face-to-face interaction. *PLoS One* 13, e0208030
58. Li, X. (2014) Leaning and recipient intervening questions in Mandarin conversation. *J. Pragmat.* 67, 34–60
59. Bressem, J. and Müller, C. (2014) 119. A repertoire of German recurrent gestures with pragmatic functions. In *Body-Language Communication: An International Handbook on Multimodality in Human Interaction* (Müller, C., *et al.*, eds), pp. 1575–1591, De Gruyter Mouton
60. Müller, C. (2017) How recurrent gestures mean: conventionalized contexts-of-use and embodied motivation. *Gesture* 16, 277–304
61. Bavelas, J.B. *et al.* (1992) Interactive gestures. *Discourse Process.* 15, 469–489
62. Bavelas, J.B. *et al.* (1995) Gestures specialized for dialogue. *Personal. Soc. Psychol. Bull.* 21, 394–405
63. Chu, M. *et al.* (2014) Individual differences in frequency and saliency of speech-accompanying gestures: the role of cognitive abilities and empathy. *J. Exp. Psychol. Gen.* 143, 694–709
64. Cooperrider, K. *et al.* (2018) The palm-up puzzle: meanings and origins of a widespread form in gesture and sign. *Front. Commun.* 3, 23

65. Bressem, J. and Müller, C. (2017) The "negative-assessment-construction" – a multimodal pattern based on a recurrent gesture? *Linguist. Vanguard* 3, 1–9
66. Bressem, J. *et al.* (2017) Multimodal language use in Savosavo: refusing, excluding and negating with speech and gesture. *Pragmatics* 27, 173–206
67. Ladewig, S.H. and Bressem, J. (2013) New insights into the medium hand: discovering recurrent structures in gestures. *Semiotica* 2013, 203–231
68. Bates, E. *et al.* (1979) *The Emergence of Symbols: Cognition and Communication in Infancy,* Academic Press
69. Cameron-Faulkner, T. *et al.* (2015) The relationship between infant holdout and gives, and pointing. *Infancy* 20, 576–586
70. Boundy, L. *et al.* (2016) Exploring early communicative behaviours: a fine-grained analysis of infant shows and gives. *Infant Behav. Dev.* 44, 86–97
71. Rossano, F. and Liebal, K. (2014) "Requests" and "offers" in orangutans and human infants. In *Studies in Language and Social Interaction* (Vol. 26) (Drew, P. and Couper-Kuhlen, E., eds), pp. 335–364, John Benjamins
72. Pollick, A.S. and de Waal, F.B.M. (2007) Ape gestures and language evolution. *Proc. Natl. Acad. Sci. U. S. A.* 104, 8184–8189
73. Kilner, J.M. *et al.* (2007) Predictive coding: an account of the mirror neuron system. *Cogn. Process.* 8, 159–166
74. Friston, K. (2010) The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138
75. de Lange, F.P. *et al.* (2018) How do expectations shape perception? *Trends Cogn. Sci.* 22, 764–779
76. Federmeier, K.D. and Kutas, M. (1999) A rose by any other name: long-term memory structure and sentence processing. *J. Mem. Lang.* 41, 469–495
77. DeLong, K.A. *et al.* (2005) Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nat. Neurosci.* 8, 1117–1121
78. Van Berkum, J.J.A. *et al.* (2005) Anticipating upcoming words in discourse: evidence from ERPs and reading times. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 443–467
79. Pickering, M.J. and Garrod, S. (2013) An integrated theory of language production and comprehension. *Behav. Brain Sci.* 36, 329–347
80. Kuperberg, G.R. and Jaeger, T.F. (2016) What do we mean by prediction in language comprehension? *Lang. Cogn. Neurosci.* 31, 32–59
81. Hagoort, P. and Van Berkum, J.J.A. (2007) Beyond the sentence given. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 362, 801–811
82. Venezia, J.H. *et al.* (2016) Timing in audiovisual speech perception: a mini review and new psychophysical data. *Atten. Percept. Psychophys.* 78, 583–601
83. Schroeder, C.E. *et al.* (2008) Neuronal oscillations and visual amplification of speech. *Trends Cogn. Sci.* 12, 106–113
84. Yap, D-F. *et al.* (2011) Iconic gestures prime words. *Cogn. Sci.* 35, 171–183
85. Obermeier, C. *et al.* (2010) What iconic gesture fragments reveal about gesture–speech integration: when synchrony is lost, memory can help. *J. Cogn. Neurosci.* 23, 1648–1663
86. MacDonald, M.C. *et al.* (1994) The lexical nature of syntactic ambiguity resolution. *Psychol. Rev.* 101, 676–703
87. Tanenhaus, M.K. and Trueswell, J.C. (1995) Sentence comprehension. In *Speech, Language, and Communication* (2nd edn) (Miller, J.L. and Eimas, P.D., eds), pp. 217–262, Academic Press
88. Rabovsky, M. *et al.* (2018) Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nat. Hum. Behav.* 2, 693–705
89. Skipper, J.I. (2015) The NOLB model: a model of the natural organization of language and the brain. In *Cognitive Neuroscience of Natural Language Use*, pp. 101–134, Cambridge University Press
90. Gisladottir, R.S. *et al.* (2015) Conversation electrified: ERP correlates of speech act recognition in underspecified utterances. *PLoS One* 10, e0120068
91. Christiansen, M.H. and Chater, N. (2016) The now-or-never bottleneck: a fundamental constraint on language. *Behav. Brain Sci.* 39, e62
92. Levinson, S.C. (2016) "Process and perish" or multiple buffers with push-down stacks? *Behav. Brain Sci.* 39, e62
93. Pickering, M.J. and Garrod, S. (2004) Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27, 169–190
94. Magyari, L. *et al.* (2014) Early anticipation lies behind the speed of response in conversation. *J. Cogn. Neurosci.* 26, 2530–2539
95. Magyari, L. *et al.* (2017) Temporal preparation for speaking in question–answer sequences. *Front. Psychol.* 8, 211
96. Bögels, S. *et al.* (2018) Planning versus comprehension in turn-taking: fast responders show reduced anticipatory processing of the question. *Neuropsychologia* 109, 295–310
97. De Ruiter, J.P. and McNeill, D. (2000) The production of gesture and speech. In *Language and Gesture*, pp. 284–311, Cambridge University Press
98. Kita, S. and Özyürek, A. (2003) What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. *J. Mem. Lang.* 48, 16–32
99. Hostetter, A.B. and Alibali, M.W. (2008) Visible embodiment: gestures as simulated action. *Psychon. Bull. Rev.* 15, 495–514
100. Kita, S. *et al.* (2017) How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychol. Rev.* 124, 245–266
101. Bavelas, J. *et al.* (2014) Effect of dialogue on demonstrations: direct quotations, facial portrayals, hand gestures, and figurative references. *Discourse Process.* 51, 619–655
102. Staudte, M. *et al.* (2014) The influence of speaker gaze on listener comprehension: contrasting visual versus intentional accounts. *Cognition* 133, 317–328
103. Holler, J. *et al.* (2015) Eye'm talking to you: speakers' gaze direction modulates co-speech gesture processing in the right MTG. *Soc. Cogn. Affect. Neurosci.* 10, 255–261
104. Clark, H.H. (2016) Depicting as a method of communication. *Psychol. Rev.* 123, 324–347
105. Hasson, U. *et al.* (2018) Grounding the neurobiology of language in first principles: the necessity of non-language-centric explanations for language comprehension. *Cognition* 180, 135–157
106. Levinson, S.C. and Holler, J. (2014) The origin of human multimodal communication. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 369, 20130302
107. Castiello, U. *et al.* (2010) Wired to be social: the ontogeny of human interaction. *PLoS One* 5, e13199
108. Farroni, T. *et al.* (2002) Eye contact detection in humans from birth. *Proc. Natl. Acad. Sci. U. S. A.* 99, 9602–9605
109. Reid, V.M. *et al.* (2017) The human fetus preferentially engages with face-like visual stimuli. *Curr. Biol.* 27, 1825–1828.e3
110. Jiang, J. *et al.* (2015) Leader emergence through interpersonal neural synchronization. *Proc. Natl. Acad. Sci. U. S. A.* 112, 4274–4279
111. Kobayashi, H. and Kohshima, S. (1997) Unique morphology of the human eye. *Nature* 387, 767–768
112. Clark, H.H. (1996) *Using Language,* Cambridge University Press
113. Sacks, H. *et al.* (1974) A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696–735
114. Goodwin, C. (1981) *Conversational Organization: Interaction Between Speakers and Hearers,* Academic Press
115. Menenti, L. *et al.* (2011) Shared language: overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. *Psychol. Sci.* 22, 1173–1182
116. Silbert, L.J. *et al.* (2014) Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proc. Natl. Acad. Sci. U. S. A.* 111, E4687–E4696
117. Chater, N. *et al.* (2016) Language as skill: intertwining comprehension and production. *J. Mem. Lang.* 89, 244–254
118. Garrod, S. and Pickering, M.J. (2015) The use of content and timing to predict turn transitions. *Front. Psychol.* 6, 751
119. Chu, M. and Hagoort, P. (2014) Synchronization of speech and gesture: evidence for interaction in action. *J. Exp. Psychol. Gen.* 143, 1726–1741
120. Levelt, W.J.M. *et al.* (1985) Pointing and voicing in deictic expressions. *J. Mem. Lang.* 24, 133–164
121. Lashley, K.S. (1951) The problem of serial order in behavior. In *Cerebral Mechanisms in Behavior; The Hixon Symposium*, pp. 112–146, Wiley

122. Wertheimer, M. (1923) Untersuchungen zur Lehre von der Gestalt. II. *Psychol. Forsch.* 4, 301–350 (in German)
123. Koffka, K. (1935) *Principles of Gestalt Psychology,* Harcourt, Brace
124. McGurk, H. and Macdonald, J. (1976) Hearing lips and seeing voices. *Nature* 264, 746–748
125. Wallace, M.T. *et al.* (2006) The development of cortical multisensory integration. *J. Neurosci.* 26, 11844–11849
126. Schroeder, C.E. and Foxe, J. (2005) Multisensory contributions to low-level, 'unisensory' processing. *Curr. Opin. Neurobiol.* 15, 454–458
127. Noppeney, U. *et al.* (2008) The effect of prior visual information on recognition of speech and sounds. *Cereb. Cortex* 18, 598–609
128. Levinson, S.C. (2013) Action formation and ascription. In *The Handbook of Conversation Analysis* (Stivers, T. and Sidnell, J., eds), pp. 101–130, Wiley-Blackwell
129. Bar, M. (2004) Visual objects in context. *Nat. Rev. Neurosci.* 5, 617–629
130. Rasche, C. and Koch, C. (2002) Recognizing the gist of a visual scene: possible perceptual and neural mechanisms. *Neurocomputing* 44–46, 979–984
131. Deroy, O. *et al.* (2016) Metacognition in multisensory perception. *Trends Cogn. Sci.* 20, 736–747
132. Winkler, I. *et al.* (2012) Multistability in auditory stream segregation: a predictive coding view. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 367, 1001–1012
133. Winkler, I. *et al.* (2009) Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends Cogn. Sci.* 13, 532–540
134. Jackendoff, R. and Lerdahl, F. (2006) The capacity for music: what is it, and what's special about it? *Cognition* 100, 33–72