Multimodal communication and language origins: integrating gestures and vocalizations

Marlen Fröhlich[1,*], Christine Sievers[2], Simon W. Townsend[3], Thibaud Gruber[4,5,†] and Carel P. van Schaik[1,†]

[1]*Department of Anthropology, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland*

[2]*Department of Philosophy and Media Studies, Philosophy Seminar, University of Basel, Holbeinstrasse 12, 4051 Basel, Switzerland*

[3]*Department of Comparative Linguistics, University of Zurich, Plattenstrasse 54, 8032 Zurich, Switzerland*

[4]*Swiss Center for Affective Sciences, CISA, University of Geneva, Chemin des Mines 9, 1202 Geneva, Switzerland*

[5]*Department of Zoology, University of Oxford, 11a Mansfield Road, OX1 3SZ Oxford, UK*

[*]Author for correspondence (E-mail: marlen.froehlich@uzh.ch; Tel.: +41 (0) 44 635 5413).

[†]Equal senior authors.

ABSTRACT

The presence of divergent and independent research traditions in the gestural and vocal domains of primate communication has resulted in major discrepancies in the definition and operationalization of cognitive concepts. However, in recent years, accumulating evidence from behavioural and

neurobiological research has shown that both human and non-human primate communication is inherently multimodal. It is therefore timely to integrate the study of gestural and vocal communication. Herein, we review evidence demonstrating that there is no clear difference between primate gestures and vocalizations in the extent to which they show evidence for the presence of key language properties: intentionality, reference, iconicity and turn-taking. We also find high overlap in the neurobiological mechanisms producing primate gestures and vocalizations, as well as in ontogenetic flexibility in gestures and vocalizations. These findings confirm that human language had multimodal origins. Nonetheless, we note that in great apes, gestures seem to fulfil a carrying (i.e. predominantly informative) role in close-range communication, whereas the opposite holds for face-to-face interactions of humans. This suggests an evolutionary shift in the carrying role from the gestural to the vocal stream, and we explore this transition in the carrying modality. Finally, we suggest that future studies should focus on the links between complex communication, sociality and cooperative tendency to strengthen the study of language origins.

Key words: evolution of language, cognition, ontogeny, gestural origins, vocal origins, primates, comparative approach, learning, multimodality.

CONTENTS

I. INTRODUCTION: LANGUAGE EVOLUTION AND THE COMPARATIVE APPROACH

Language is predominantly manifested in face-to-face interactions (Vigliocco, Perniss & Vinson, 2014), and probably evolved primarily in this context: the "core ecological niche for language use" (Roberts, Torreira & Levinson, 2015, p. 119). Along with speech, humans transmit a great deal of information to others through body language (Ekman & Friesen, 1969; Ekman, 1973), consistently integrating visual and acoustic (and sometimes also tactile) components. Across all cultures and ages, human speech is accompanied by visual signals and cues such as gestures, postures, facial expressions and eye gaze (Ekman & Friesen, 1969; McNeill, 2000; Levinson & Holler, 2014). Likewise, speech has been considered as a form of 'action' where mouth actions are dynamically integrated with manual gestures as well as other bodily movements, forming a unified communication system (Kendon, 1980, 2000, 2004). This additional information afforded by multimodality matters greatly. McGurk and MacDonald (1976) demonstrated that the acoustic

perception of speech is modified by the accompanying articulatory gestures serving as a visual cue (termed the 'McGurk effect'). Moreover, Massaro's perceptual experiments (Massaro & Egan, 1996; Massaro, 1998), which manipulated the degree of conflicting audio and visual speech information, suggested that humans rely on both channels to understand the signal, giving more weight to the channel with the most reliable information. If we remove these additional communicative acts of the body, the comprehension of language is often impaired: 'emojis', for instance, were invented to remove the ambiguity in text messages (Lo, 2008; Kaye, Malone & Wall, 2017). In light of the universal use of bodily signals and cues to complement our words and refine our messages (e.g. Goldin-Meadow, 1999; McNeill, 2000; Kendon, 2004; Holle & Gunter, 2007), it is therefore surprising that human communication has traditionally been equated to speech (Hockett, 1960; Lieberman, 1993; Hauser, Chomsky & Fitch, 2002). In addition, in light of the tight integration between gesture, speech and other bodily actions in human face-to-face communication, a 'dual-modality view' focusing principally on vocalizations and gesture might not be very useful (Kendon, 2000).

The evolutionary roots of language have puzzled researchers for more than 150 years (e.g. Fiske, 1863; Tylor, 1866). Human communication clearly has a strong biological basis in the brain and the vocal apparatus (the critical elements for speech production), yet understanding its evolution has been hampered by the obvious fact that these anatomical features do not fossilize (Ghazanfar & Rendall, 2008). Nonetheless, there is ample evidence that many components of language are shared with other animals (Hauser *et al.*, 2002; Fitch, 2010; Levinson & Holler, 2014; van Schaik, 2016; Townsend, Koski, Byrne *et al.*, 2017), and the comparative approach has increasingly been used to gain insight into the cognitive building blocks and selective pressures shaping the human communication system.

As a consequence of the widely held dual-modality view of language (Kendon, 2000), researchers have long debated whether the foundation for language evolution lay in the gestural ('gesture-first

theory of language origins') or the vocal domain ('vocal-first theory of language origins') (for reviews see Fitch, 2010; Liebal *et al.*, 2013; Kendon, 2017). In essence, the argument is about whether one or the other of these modalities shows more key cognitive characteristics of language, such as intentionality, reference, iconicity, combinatoriality, turn-taking, neural control and ontogenetic plasticity (Arbib, Liebal & Pika, 2008; Tomasello, 2008; Liebal *et al.*, 2013; Levinson & Holler, 2014). Comparative research on human and non-human communicative interactions has therefore focused especially on these cognitive mechanisms (Liebal *et al.*, 2013; Sievers & Gruber, 2016; Townsend, Engesser, Stoll *et al.*, 2018). First, intentional communication has been broadly defined as signalling voluntarily and in a goal-directed way, with different orders of intentionality distinguished depending on the underlying degree of mental state attribution (Dennett, 1983; Townsend *et al.*, 2017). Second, reference concerns signals that draw attention to relevant external objects and events (e.g. Zuberbühler, 2003; Leavens, 2004). Third, iconicity can be considered as the resemblance between form and function of a signal and as the opposite of arbitrariness; it has recently received much attention among evolutionary linguists (Perniss, Thompson & Vigliocco, 2010; Perlman & Cain, 2014; Vigliocco *et al.*, 2014; Lockwood & Dingemanse, 2015). Fourth, combinatorial capacities in non-human species have been demonstrated in the combination of both meaningless elements or meaningful signals into larger meaningful structures (e.g. Arnold & Zuberbühler, 2008; Ouattara, Lemasson & Zuberbühler, 2009; Engesser, Ridley & Townsend, 2016). Fifth, turn-taking, or a rapid exchange of turns, has recently received interest in comparative research due to the fact that most language usage is interactive and conversational (Rossano, 2013; Levinson, 2016). Sixth studies of ontogenetic plasticity ask to what extent the signal repertoire is innate or acquired (Tomasello *et al.*, 1994; Brainard & Doupe, 2002; Beecher & Brenowitz, 2005; Watson *et al.*, 2015).

Until recently, the debate on language origins largely ignored the notion that language is fundamentally multimodal. Most of the comparative work to date has focused on extant non-human

primates, given their close phylogenetic proximity to humans and the insights they can offer into the communicative abilities of our hominin ancestors. We therefore start with reviewing the evidence for multimodality in animal (in particular non-human primate) communication (Section II) to assess whether it too is fundamentally multimodal. Having confirmed the multimodal nature of primate communication and discussed its function, we then re-assess to what extent the six crucial properties of language listed above are present in both gestures and vocalizations, taking into account that the research traditions on the gestural and vocal domains (but also the cognitive and ecological perspectives of primate communication) have diverged substantially over past decades (Section III). We also take a closer look at recent work on the neural mechanisms and cognition involved in multimodal integration in primates and examine the overlap in proximate mechanisms (from neurobiological pathways to degree of voluntary control) underlying gestures and vocalizations. Finding substantial overlap between gestures and vocalizations in these features, and thus concluding that language origins must also have been multimodal, we declare an end to the debate over the likeliest language precursor. This conclusion has important implications for future research, outlined in Section IV. We return to the gestures-first *versus* vocalizations-first debate to summarize the various points of contention between the two sides but also note changes in each modality where both fail to provide parsimonious explanations. Finally we discuss the implications of our perspective for the evolution of language as a multimodal system, and suggest fruitful ways to explore the relationship further between communicative complexity and other traits that are thought to have been critical in human evolution.

Before we start, we acknowledge that this review might paradoxically be seen as adopting a 'dual-modality view' of gestures and vocalizations, since the evidence we present is biased towards gesture and vocalization. We acknowledge that human multimodal communication goes far beyond speech and gesture, and also incorporates facial expressions, bodily movements/actions and non-speech affective vocalizations (e.g. laughter, crying) (Ekman & Friesen, 1969; Levinson & Holler, 2014).

However, because one of our major aims is to discuss and reconcile the evidence for cognitive and neural mechanisms underlying gestures and vocalization (the most heavily debated communication domains in the field of language evolution), a detailed discussion of other signal types is beyond our scope. However, while we will principally focus on gestures and vocalizations, we touch on other channels too, such as facial expressions. As our review will show, we aim to provide an integrated model of communication that is multimodal by nature rather than one focused on modalities.

II. THE RISE OF 'MULTIMODALISM' IN COMPARATIVE RESEARCH

In recent years, the communication of non-human primates has come to be divided into three behavioural modalities: gestures, facial expressions and vocalizations (Liebal *et al.*, 2013). Most comparative researchers would nowadays agree that studying vocalizations and bodily and facial movements in isolation fails to disentangle the function of communicative acts and to describe fully the communication systems of primates (e.g. Slocombe, Waller & Liebal, 2011; Higham & Hebets, 2013; Hobaiter, Byrne & Zuberbühler, 2017; Fröhlich & van Schaik, 2018). It thus appears as if most behavioural ecologists, comparative psychologists and evolutionary linguists have become 'multimodalists' by now. But what do we mean when we talk about a 'multimodal signal' or 'multimodal signal combinations'?

Multimodality is defined in various ways (Partan & Marler, 1999; Higham & Hebets, 2013; Liebal *et al.*, 2013; Levinson & Holler, 2014) and is used in somewhat different meanings in comparative psychology and behavioural ecology. Comparative psychologists have typically focused on signal production in human and non-human primates – particularly great apes – and refer to signal categories such as vocalization, gesture, or facial expression as a 'modality' of communication (Fröhlich & Hobaiter, 2018; Fröhlich & van Schaik, 2018). Multimodal signals are then described as arising from the simultaneous or sequential integration of signals from at least two of the 'modalities', e.g. gesture and facial expression (Liebal *et al.*, 2013). However, outside of great ape

communication, the term ‘modality’ is typically used to refer to perception through the sensory channels of vision, touch, hearing, olfaction, etc. (Rowe, 1999; Partan & Marler, 2005). For example, a single gesture (e.g. a visual–audible ‘slap object’) can be multimodal from the perspective of a behavioural ecologist (focusing on the sensory channels of perception), but not from the perspective of a comparative psychologist (focusing on the communicative channels of production) (Higham & Hebets, 2013). By contrast, a visual–silent gesture such as an ‘arm wave’ combined with a (visual) facial expression would be classified as multimodal by a comparative psychologist, but unimodal (visual) by a behavioural ecologist (Wilke *et al.*, 2017). Here we focus on what we think is the core of multimodality, namely that in production and/or perception different input channels must be integrated to form a communicative unit and/or to identify a distinct message. *Multimodal signals* are then defined as signals consisting of two or more components of different sensory modalities which are obligatorily coupled (e.g. lip-smacking with a salient visual and auditory component), whereas in *multimodal signal combinations* two or more distinct signals, which incorporate different sensory modalities (e.g. silent non-contact gesture plus vocalization), are flexibly coupled (Fröhlich & Hobaiter, 2018; Fröhlich & van Schaik, 2018).

Some communication is naturally unimodal. Examples include signals displaced in space (bird song, e.g. Brumm & Slabbekoorn, 2005; primate long-distance vocalizations, e.g. Bornean orangutans’ *Pongo pygmaeus* long call; Spillmann, Dunkel, Van Noordwijk *et al.*, 2010)] or time (olfactory signals, e.g. marmoset’s pheromones; Barrett, Abbott & George, 1990)]. However, behavioural and neuro-ethological research has firmly established that primates commonly integrate the modalities of gesture, facial expression and vocalization in their short-range communication (for review see Liebal *et al.*, 2013). Most studies on primate multimodal communication to date have focused either on the *production* of flexible combinations of gesture and vocalization (e.g. Pollick & De Waal, 2007; Genty *et al.*, 2014; Hobaiter *et al.*, 2017; Wilke *et al.*, 2017) or the *perception* of vocal (auditory) and facial (visual) components within a signal (Partan, 1999, 2002; Ghazanfar, 2013).

While the function of complex (i.e. multi-component and multimodal) communication in non-primate animals has received much interest during the past decade (Hebets & Papaj, 2005; Higham & Hebets, 2013; Partan, 2013), work in non-human primates, the major model system for studies on the evolution of language, has only recently started on this issue (Micheletta *et al.*, 2013; Hobaiter *et al.*, 2017; Wilke *et al.*, 2017). The function as well as the socio-ecological drivers of flexible gestural–vocal combinations have been of particular interest to primatologists interested in the underlying cognitive mechanisms (reviewed in Fröhlich & van Schaik, 2018). For instance, recent work showed that bonobos (*Pan paniscus*) use the same vocalization ('contest-hoot') in playful and aggressive contexts but add gestures to distinguish between the two (Genty *et al.*, 2014), thereby clarifying an ambiguous message sent in one channel by adding a more specific component in another channel. For wild chimpanzees (*Pan troglodytes*), Wilke *et al.* (2017) recently provided evidence that responses to multimodal signals (gesture + vocal) were more likely to match the response of the gestural than the vocal components. In line with this study, Hobaiter *et al.* (2017) subsequently showed that wild chimpanzees, after perceived goals were not achieved, switched to bi-modal signal use (gesture + vocalization) only if the initial signals were exclusively vocal. These examples suggest that the gestural mode seems to carry great ape close-distance communication, i.e. play the dominant role as information carrier, with other signals mainly helping to disambiguate (Table 1). Even in conditions of good visibility and plenty of social exposure (i.e. in captivity), signals of different sensory modalities are frequently combined into multimodal sequences or signals (Taglialatela, Russell, Pope *et al.*, 2015). This also supports the hypothesis that many signal combinations in great apes are non-redundant and, because meanings (or conveyed information) of gestures and vocalizations often overlap to some extent, signal combinations might be predominantly used for disambiguation. Thus, gesture–vocal combinations seem to function primarily to refine the message in primates (Hobaiter *et al.*, 2017; Fröhlich & van Schaik, 2018), whereas co-speech gestures frequently have more of an additive, complementary function (e.g. Cassell, McNeill &

McCullough, 1999; Kendon, 2004; see also Tab. 2). Nonetheless, as we will discuss, these human/non-human comparisons should be viewed with caution, since primate gestures and co-speech gestures, as well as primate and human vocalizations might have distinct evolutionary origins (Section IV.2).

As a result, it has been hypothesized that multimodal communication functions largely to increase comprehension by disambiguating and/or complementing a message (Hebets & Papaj, 2005; Partan & Marler, 2005; Fröhlich & van Schaik, 2018), and recent empirical work indeed strongly supports the notion of 'multimodalism' (Wacewicz & Zywiczynski, 2017). This inherent multimodality of primate communication has led researchers to propose that there is continuity in multimodal communication from primates to humans (Taglialatela, Russell, Schaeffer *et al.*, 2011; Liebal *et al.*, 2013; Levinson & Holler, 2014; Kendon, 2017; Wacewicz & Zywiczynski, 2017). We should therefore assume that the communication of our pre-linguistic ancestors was, just like language, already multimodal.

III. COGNITIVE MECHANISMS IDENTIFIED IN NON-HUMAN GESTURES AND VOCALIZATIONS

For a long time, a gulf was thought to separate animal from human communication (Chomsky, 1959; Bickerton, 1992). For instance, animal vocalizations were thought merely to reflect the emotional arousal of the producer and thus not to represent intentional communication (Tomasello, 2008). Recent research, however, not only shows that language might be much older than previously recognized (Krause *et al.*, 2007; Dediu & Levinson, 2013; Atkinson *et al.*, 2018), but also is best regarded as an 'evolutionarily stratified system' (i.e. consisting of different abilities of different evolutionary origins; Levinson & Holler, 2014). While language as a complex expressive system is undoubtedly a derived trait of humans, some critical cognitive building blocks of the human communication system are shared with animal communication (e.g. Hauser *et al.*, 2002; Arbib *et al.*,

2008; Levinson & Holler, 2014; Engesser *et al.*, 2016; Arbib, Aboitiz, Burkart *et al.*, 2018). In the study of language origins, comparative researchers have typically focused on these distinct cognitive building blocks and investigated their presence in the gestural, vocal or facial communicative acts of non-human species. Here, we discuss cognitive mechanisms in gesture and vocalization which have been the focus of comparative studies in non-human species and are thought to have played a major role in the evolution of language: intentionality, reference, iconicity, combinatoriality, turn-taking, neural control and ontogenetic plasticity (Marler, Evans & Hauser, 1992; Liebal *et al.*, 2013; Perniss & Vigliocco, 2014; Townsend *et al.*, 2017; Zuberbühler, 2018). In particular, we ask whether each of these cognitive building blocks are found only in gestural communication, as claimed by proponents of the gestures-first model of language origins (e.g. Hewes, 1973; Corballis, 2002; Armstrong & Wilcox, 2007; Tomasello, 2008), or are also found in vocal communication.

(1) Intentionality

We humans constantly engage in intentional communication. The study of intentionality therefore started out as an endeavour to describe characteristics of human intentional communication, which Grice (1957) characterized as *ostensive* on the production side, and *inferential* on the comprehension side. Thus, human communication fundamentally combines cognitive and cooperative components. From a cognitive perspective, *ostensive* signalling implies that human signallers openly communicate their intentions to inform the receiver by not just producing sentences but also by using visual signals, such as gestures and facial expressions, to make it salient to the receiver that the signaller indeed has the intention to inform about x (Sperber & Wilson, 1986). This makes human ostensive intentional communication by definition multimodal (Sievers, Wild & Gruber, 2017). Communication is *inferential* and thus successful (i.e. understanding is achieved) when the receiver recognizes those intentions. In addition, it must be noted that ostensive signal production and comprehension, at least in the traditional interpretation (but see Moore, 2016; 2017), requires

metacognition, in that signallers aim to influence the mental state of the receiver by displaying an intention to inform a conspecific, and the meta-intention of intending conspecifics to grasp this informative intention, the so-called communicative intention. Furthermore, receivers must be capable of inferring these intentions of the signaller. To Tomasello (2008) this implies that successful human communication in adult humans requires that sender and recipient share their attention ('joint attention') and goals ('shared intentionality') and enough common ground in that both accept the conventional aspects of the communicative repertoire.

Human language would not work without it being cooperative: it requires communicators to attend to the partner's signalling and to react appropriately in accordance with the context of the interaction and the partner's produced signals (Grice, 1975). Human communication is, therefore, in its essence a cooperative endeavour, with individuals attending to the communicative partner's signalling behaviour by taking communicative turns (Levinson, 2006; Rossano, 2013). In fact, a characteristic feature of human communication is its declarative use of language, with individuals sharing information (e.g. Hurford, 2007), even if not all language use is driven exclusively by cooperative intentions and locutionary acts such as threats and deception do not require any additional cooperative motivations in the communicators.

In comparison, much, although not all, of primate communication is primarily imperative (Grice, 1957; Hurford, 2007; Tomasello, 2008). However, it is possible that we have just not conducted the appropriate experiments yet to distinguish between imperative *versus* declarative signal use in non-human animals (Lyn, Russell & Hopkins, 2010; Crockford, Wittig & Zuberbühler, 2017; Leavens, Bard & Hopkins, 2017). The scarcity of declarative events in wild apes and their infrequent occurrence in enculturated individuals suggests that apes have the biological capability to declare, but specific environmental triggers must be present for their expression (Lyn *et al.*, 2010). Some playback studies in the wild seem to suggest that chimpanzees have the cognitive capacity to inform ignorant group members (i.e. those individuals not yet aware) of an imminent danger (Crockford,

Wittig, Mundry *et al.*, 2012; Crockford *et al.*, 2017). Cooperative contexts or situations in which the signaller has in fact no or only a limited direct benefit from signalling might therefore be a fruitful avenue to address declarative communication, such as social object play, tool use, scavenging (Bickerton & Szathmáry, 2011) or alarm calling.

Note that it is tempting to interpret terms like declarative and imperative, which arose in linguistics, as reflecting the sender's intention, although they can also be used in a purely functional sense. Scholars of animal communication working on vocalizations have focused on this functional sense (Hurford, 2007), while those mainly working in the gestural modality have focused on a notion of declarative and imperative reflecting the sender's intentions (e.g. Leavens, Russell & Hopkins, 2005*b*).

The study of intentional communication in non-human animals started out as an endeavour in the gestural modality (Liebal & Oña, 2018). Tomasello (Tomasello, George, Kruger *et al.*, 1985; Liebal, Call & Tomasello, 2004a) used a Grice-inspired approach to develop criteria for potentially intentional communication in the gestural modality in great apes. They defined intentional communication in its simplest form as signallers producing signals in order to achieve a goal (Tomasello *et al.*, 1985). This description of intentional communication is cognitively simpler than that presented by Grice, as it does not require meta-representations and ostensive signalling. Based on this description, and therefore focusing on signaller behaviour, Liebal (Liebal, Pika, Call *et al.*, 2004b, pp. 379-380) provided three criteria to decide that signal production was intentional: (1) the signaller is producing the signal for an audience, that is, it is used socially, which implies audience checking; the signaller checks the state of attention of the recipient. After signal production, (2) response waiting is expected, i.e. stopping, maybe gaze checking to monitor the behaviour of the targeted conspecific. Furthermore, if the recipient's response is not satisfactory, (3) the signaller should display persistence, that is, repeat the produced signal. Applying these criteria to empirical

data, Tomasello *et al.* (1985) concluded that great ape gestures but not vocalizations are produced intentionally.

Leavens *et al*. (2005*b*) provided comparable criteria for identifying intentional communication in great apes, likewise focusing on criteria for the gestures and gazes of signallers. Where Tomasello *et al.* (1985) emphasized persistence, Leavens *et al.* (2005*b*) add an emphasis on elaborative behaviour. Unlike persistence, elaboration is not about repeating the same signal until one's goal is achieved but rather using a different signal if the original one did not lead to the intended result. From a cognitive point of view elaboration is the more flexible behaviour.

Investigations in the vocal modality started out representing communication as information transmission, not focusing on a potentially intentional component, but rather on explaining how information transmission in concrete cases of call production ultimately served an adaptive function (e.g. Cheney & Seyfarth, 1981). Only later did research in this modality also focus on intentional communication, possibly triggered by the claims of e.g. Tomasello (2008) that only gestures in non-human primates are intentionally produced signals, while vocalizations are involuntarily produced. Initial attempts in monkeys mostly supported this conclusion (Cheney & Seyfarth, 1985, 1996), paralleling results in ground squirrels (Sherman, 1977), roosters (Gyger, Karakashian & Marler, 1986) or downy woodpeckers (*Dryobates pubescens*) (Sullivan, 1985). All these cases concerned high-urgency alarm calls. More recent work focused on cases where immediate flight responses are not required. Thus, Wich and de Vries (2006) found evidence that Thomas langurs (*Presbytes thomasi*) take the audience's awareness of a predator into account by persisting in emitting alarm calls at a tiger until the last female in the group has vocally acknowledged they heard them.

One of the first experimental investigations specifically designed to capture intentional communication in the vocal modality was conducted on wild chimpanzees in a field experiment by Schel *et al.* (2013). Chimpanzees were confronted with moving snake models (to which they respond with alertness rather than naked fear) to determine whether individuals would intentionally inform

others. The authors predicted that if this was the case, their calling would be dependent on the audience's gazing towards the snake. They looked for signallers displaying audience checking and gaze alternation between recipient and snake as well as evidence of persistence until everyone was informed of the presence of the snake. Furthermore, they looked for stopping rules in signal production: if the conspecific is informed, the signaller stops producing the signal, because the goal to inform is achieved. Schel *et al*. (2013) concluded that the criteria for gestural communication could be applied to vocal signals. Other studies, both observational (Gruber & Zuberbühler, 2013) and experimental (Crockford *et al.*, 2012; Crockford, Wittig & Zuberbühler, 2015; Crockford *et al.*, 2017), have provided further evidence of intentional vocal behaviour in chimpanzees, possibly even signalling an intent of changing conspecifics' mental states.

(2) Reference

For the notion of reference, the suggestion of Liebal *et al*. (2013) to adopt a multimodal approach is even more pressing because studies of gestures and vocalizations have so far used very different frameworks. On the theoretical side, the philosophy of language and linguistics employs two major notions of referential signals in humans which potential multimodal approaches could be based on. Both notions – Semantic Reference and Speaker's Reference – assume concrete cognitive processes underlying reference (Sievers & Gruber, 2016): speakers are required to intend to refer and hearers to form representations about the referent in question, as well as to infer what is referred to or to understand the referential meaning based on previously formed associations.

Semantic Reference centres on the idea that words themselves refer. It is concerned with words that are conventionally used to refer to one particular referent. For instance, the proper name "Mount Everest", if used according to convention, refers to the particular mountain in the Himalayas. The second notion of reference is labelled Speaker's Reference (Bach, 2006), focusing on the signaller's intentions to refer, and thus asking to what extent a speaker has the intention to point out something

to the receiver. Without context, the sender's intended referent, and therefore the meaning of words and sentences in a dialogue is often hard to understand. Therefore, spoken language in face-to-face communication – despite the symbolic nature of words – is relatively context dependent (Table 1). As opposed to Speaker's Reference, Semantic Reference perceives semantically referential words as referring in a context-independent manner. That is, to understand the sentence "Mount Everest is beautiful", the hearer only needs to perceive the utterance, and does not need to take into account further contextual cues, as "Mount Everest" according to *convention* is always used to refer to the particular mountain. As Sievers and Gruber (2016) point out, only the notion of Speaker's Reference is of interest for comparative research in non-human animals, as it does not rely on existing conventions of signal uses.

In the biological sciences, the interest in reference in vocal signals is traditionally determined by classifying a certain group of animal signals that function to refer to the presence of an external entity when produced. As originally proposed by Macedonia and Evans (1993), for a signal to be functionally referential it should be "elicited by a special class of stimuli and capable of causing behaviours adaptive to such stimuli in absence of contextual cues" (p. 117–178). Such a functional approach of reference remains agnostic about the cognitive processes involved, such as whether animals have a representation of the referent. Some experiments with wild Diana (*Cercopithecus diana*) and Campbell's monkeys (*Cercopithecus campbelli*) suggest that receivers attend to the call's meaning rather than to acoustical features, and thus might form some kind of mental representation upon perceiving a call (Zuberbühler, 2000*b*, *a*). However, a number of researchers argue that effects on receivers can be explained by simpler mechanisms, and mental representations may not be necessary to produce and respond appropriately to functionally referential signals (Marler, Evans & Hauser, 1992a; Owren & Rendall, 2001; Seyfarth, Cheney, Bergman *et al.*, 2010).

Other than issues associated with identifying underlying cognitive processes, the functionally referential framework has been identified as problematic for additional reasons. Wheeler and Fischer

(2012), for example, point out that context independence of a call's embedded information at the receiver's side is actually rare to non-existent. Most calls in the animal kingdom are necessarily linked to external contextual cues in order to cause the appropriate reaction in the receiver (e.g. Zuberbühler, Cheney & Seyfarth, 1999; Clay, Smith & Blumstein, 2012; Arnold & Zuberbühler, 2014). Ducheminsky, Henzi and Barrett (2014) point out that in eagle-alarm-call production of vervet monkeys (*Chlorocebus pygerythrus*), the receiver looks up into the sky before it responds appropriately to the situation (i.e. to run deep into the foliage), which they take as evidence for a context-dependent reaction. However, it could also be argued that such a response is equally supportive of a functional reference-based interpretation. If the signaller's call was indeed ambiguous in meaning, we might expect that the recipient should look first to the signaller to obtain contextual information. However, the fact that the monkey looks into the direction from where the specific predator might come (i.e. the sky) indicates that the signal has conveyed a distinct message, namely a threat from above (see also Schel, Tranquilli & Zuberbühler, 2009). There might also be additional reasons for the recipient to look up first before reacting: for example, if the monkey expects an aerial predator, it still may need to locate the predator in order to find an effective hiding place relative to the position of the predator.

Nevertheless, examples in animal vocal communication where receivers do not consider the context before reacting seem to be uncommon (Clay *et al.*, 2012; Arnold & Zuberbühler, 2014; Tab. 2). These vocalizations might nonetheless still fulfil the criteria for Speaker's Reference if they are accompanied by markers indicating the intention to 'denote' something to the receiver (Schel *et al.*, 2013; Crockford *et al.*, 2017). Studies on chimpanzees' alarm and food calls, for example, have provided evidence that these calls are intentionally directed, given that they were socially used (e.g. directed at specific recipients) (Crockford *et al.*, 2012, 2017) and associated with audience checking, gaze alternation and goal persistence (Schel *et al.*, 2013). Our overview thus does not deny that functionally referential calls can be cognitively complex, but it does emphasize that the concept of

functional reference cannot inform as much regarding cognitive complexity as perhaps previously thought. In conclusion, the cognitive mechanisms underlying vocal reference in non-human primates remain unclear.

In contrast to vocal research, gestural research determining referential uses of signals started out very differently, by relying on resemblance to human pointing behaviour as a criterion for referential uses of gestures, often referring to the lack of pointing gestures, or indeed any deictic gesture (one whose meaning depends on context) in other primate species (Liebal *et al.*, 2013). Primate equivalents of deictic gesture use have been mostly studied in captive apes, who produce pointing gestures solely in interactions with human caretakers (e.g. Leavens, Hopkins & Bard, 2005*a*; Tomasello, 2006; Lyn *et al.*, 2010). Their ability to comprehend them is also very limited (Tomasello, 2006), probably due to the lack of common ground (Moore, 2013). For intra-specific interactions, pointing behaviour in wild great apes was only reported once in bonobos (Veà & Sabater-Pi, 1998), but at least some possible cases of deictic gesturing have been described in the wild (chimpanzees: Pika & Mitani, 2006; Hobaiter, Leavens & Byrne, 2014) and captivity (bonobos: Genty & Zuberbühler, 2014). In principle, gaze alternation between an object and a conspecific (Tomasello *et al.*, 1994; Leavens *et al.*, 2005b; Schel *et al.*, 2013) might also qualify as deictic gesturing. In the example of the chimpanzee alarm call to a moving snake discussed above, it seems that gaze alternation functions to point out the intended referent of a concurrent vocalization (Schel *et al.*, 2013). However, it is controversial whether gaze alternation can reliably indicate intentional communication (as discussed in Liebal *et al*., 2013, p. 178): it may simply indicate two competing entities of interest (i.e. the object and the conspecific) in the signaller's mind, so that gaze alternation may simply represent vacillation between these two.

When turning to other animals, including corvids (Pika & Bugnyar, 2011) and coral reef fish (Vail, Manica & Bshary, 2013), the approach slightly changes to applying a list of criteria to decide whether a gesture is referential: the behaviour is supposed to be (1) directed towards a referent; (2)

mechanically ineffective; and (3) meant to be perceived by a specific recipient. In these examples identifying referential gesture use in non-primate species, the very idea of reference revolves around non-vocal behaviours that physically point something out to someone.

These criteria appear difficult to apply to vocally referential signals, raising doubts about the feasibility of a multimodal outlook on referential uses of signals. Therefore, Sievers and Gruber (2016) propose to follow the notion of Speaker's Reference and thus look at proximate mechanisms involved (just as gesture research on reference does). Thus, for both modalities, reference is understood as the intention of the signaller to refer the recipient to something. They label this account Signaller's Reference and build their framework around the flexibility with which signallers produce their signals, such as changing the modality of the signal depending on the outer circumstances to enable the receiver to grasp the reference.

Comparing gestural and vocal frameworks for reference, a major problem within the gestural domain is that experimental paradigms to test for gesture semantics, that is for an actual referent, are virtually absent. While the *apparently satisfactory outcome* might be a fruitful start to tackle this issue (Hobaiter & Byrne, 2014), it is still unclear whether and how the referential meaning of the gesture is really processed by receivers. At present, there seems to be clearer evidence for rudimentary referential signal usage in animal vocalizations (although limited to a small set of objects, such as major predators or foods) than gestures, which led numerous researchers to favour a vocal scenario of language origins (Cheney & Seyfarth, 2005; Zuberbühler, 2005; Arnold & Zuberbühler, 2006). Nonetheless, major discrepancies in methodology and concepts prevent any final conclusions (Liebal *et al.*, 2013).

(3) Iconicity

Iconicity is the phenomenon whereby a sign's meaning can be predicted from its structure. It has been argued that iconicity might have played a central role in establishing displacement (referring to

something not immediately present) in language evolution, and in supporting referentiality (learning to assign linguistic labels to objects and events in the world), in language development (Perniss & Vigliocco, 2014). The opposite of iconicity is arbitrariness (one of Hockett's design features of language), which means that there is no direct connection between the signal (word) and what is being referenced (meaning); such communication is symbolic (Hockett, 1960). A common argument by proponents of the gesture-first theory of language origins is that vocalizations have been bootstrapped on gestures due to the latter's greater opportunities for iconic productivity in visual space, as evident in forms of pantomime (Armstrong & Wilcox, 2007; Tomasello, 2008; Fay, Arbib & Garrod, 2013).

In contrast to intentionality and reference, the majority of studies focusing on this cognitive domain have involved adult humans. For instance, two experiments compared the suitability of non-linguistic vocalization and gestures for iconic representation in the creation of novel communication systems (Fay *et al.*, 2013; Fay *et al.*, 2014). Gestural communication proved to be both more effective (i.e. successful) and more efficient (i.e. faster) than non-linguistic vocal communication in creating signalling systems from scratch, and both studies showed that combining gestures and vocalizations did not improve performance beyond gestures alone. In an experimental study on children, iconic gestures were understood better than iconic vocalizations by 24- and 36-month-olds (but not 18-months-olds), suggesting that iconic gestures support language ontogeny (Bohn, Call & Tomasello, 2019). In fact, iconic signals appear to emerge with the instantiation of most human communication systems: many signals of American Sign Language, in which gestures and facial expressions carry the full communicative burden (Goldin-Meadow, 1999), arose as iconic representations of objects or actions before gradually losing the original resemblance between form and referent (Frishberg, 1975). In human participants who were experimentally prohibited from using their existing language, a shift from iconic to increasingly symbolic and language-like signals occurred to establish mutual understanding across repeated interactions (Garrod *et al.*, 2007, 2010). Communicating novel

meanings *via* iconic visual production is also demonstrated by signers who, faced with an unfamiliar object or event, tend to create an iconic sign for it (Klima & Bellugi, 1979). Human research has thus firmly established that iconic gesture, by communicating through resemblance, represents a more precise modality than non-linguistic vocalization for establishing reference in the absence of language (Fay *et al.*, 2014). It might therefore be better-suited for early communication systems before these get shaped through many instances of repeated interaction.

Nonetheless, iconicity also plays a role in word (and hence, vocalization) learning. Imai & Kita (2014) developed the 'sound–symbolism bootstrapping hypothesis', in which iconic sounds (i.e. signals with sound–meaning mappings) would enable human infants to understand that perceived sounds refer to things in the environment and focus on specific form–meaning mappings. There is now increasing evidence for the potential for vocal iconicity beyond onomatopoeia and sound symbolism (Perniss *et al.*, 2010; Imai & Kita, 2014; Lockwood & Dingemanse, 2015). We know that human minds are capable of perceiving similarities not only within but also between different sensory modalities and cognitive domains. In fact, 'cross-modal iconicity' – resemblance of signal and meaning crossing the borders between sensory structures and cognitive configurations – is thought to be extremely widespread (Ahlner & Zlatev, 2010; Elleström, 2017). In their experimental study with Swedish students, Ahlner & Zlatev (2010) varied vowels and consonants in fictive word-forms, and concluded that both types of sounds play a role in perceiving an iconic connection between the word-forms and visual figures. As such, cross-modal iconicity appears to connect the two extremes of pure (unimodal) sensory resemblance (e.g. a whistle signifying a bird's song) and purely cognitive resemblance (e.g. the notion of a wheel standing for the passing of time), enabled by the tight connection between perception and cognition (Elleström, 2017). In addition, growing evidence demonstrates that particular phonemes – the human speech sounds that constitute words – are associated with particular semantic contents, with differences found between voiced and voiceless phonemes, rounded and sharp phonemes, or phonemes produced in the front or back of the

mouth (reviewed in Myers-Schulz *et al.*, 2013; Schmidtke, Conrad & Jacobs, 2014). Although it is often unclear whether these psychological associations are due to physical features of articulations (e.g. visual perception of associated facial expressions), or due to the listener's previous auditory experience of acoustic features, their existence supports the idea that vocalizations may also be iconic to some extent.

While evidence for vocal iconicity in humans seems to be increasing (Perlman & Cain, 2014; Perlman, 2017), there is currently no evidence that iconicity exists in animal vocalizations. However, this could at least partly be due to the difficulty of operationalizing iconicity for this communicative domain, and of disentangling the informational from the emotional component ('the higher the fundamental frequency, the higher the danger'); the right experiments may have not yet been conducted. Adopting the human definition, comparative researchers categorized a gesture as being iconic "when its motion path in space or on another animal's body follows a path of movement or form of an action which is inferred to be desired of another animal by a gesturing animal" (Tanner & Byrne, 1996; p. 164). In principle, this definition would include numerous cases of described intention movements and gesture types where form resembles function, such as a beckoning gesture in sexual initiations (Genty & Zuberbühler, 2014), a "hip shimmy" (involving rapid lateral hip movements) resulting in genito-genital rubbing (Douglas & Moscovice, 2015), or an outstretched leg of a mother persuading her infant to climb on her for joint travel (Fröhlich, Wittig & Pika, 2016*b*). Although experimental studies reveal that great apes have difficulties with spontaneously comprehending iconic signals, it was also shown that they learn them faster than arbitrary gestures (Bohn, Call & Tomasello, 2016; Bohn *et al.*, 2019). In conclusion, the presence of iconicity is more established and striking in the motor-visual domain, and has not been shown in the vocal-auditory domain of non-human species. Because gesture thus seems to trump vocalizations at creating communication systems from scratch (where vocalizations are 'piggy-backed' on gesture) (Fay *et al.*,

2013, 2014), studies on human communication seem to support 'gesture-first' theories of language origins for this particular feature (Hewes, 1973).

(4) Combinatoriality

The ability to generate an infinite number of expressions with a novel meaning from a finite number of signal elements is another key feature of language (Hockett, 1960; Hauser *et al.*, 2002; Fitch, 2010), and has received increasing attention by comparative researchers in recent years (for reviews see Townsend *et al.*, 2018; Zuberbühler, 2018). Human speech is composed of discrete, mainly meaningless phonemes forming discrete meaningful words (phonology), which in turn are combined into more complex phrases or expressions (compositional syntax) (Hockett, 1960). Both observations and experiments have demonstrated that certain primate species are able to combine context-specific, meaningful vocal signals into sequences similar to the combinatorial structures found in language. Male Campbell's monkeys, for example, produce predator-specific alarm calls ('krak' and 'hok' in response to leopards and eagles, respectively) that can be affixed with an acoustic modifier ('-oo') to broaden the respective call's meaning (Ouattara *et al.*, 2009). Given that the affix changes the meaning of the stem alarm calls in a predictable way (from specific to more general) this has been argued to represent a rudimentary form of compositionality, akin to abstract meaning operators in language, such as 'like' (see Collier, Bickel, van Schaik *et al.*, 2014). Furthermore, male putty-nosed monkeys (*Cercopithecus nictitans*), another guenon species, combine two predator-specific alarm calls into a higher-order sequence, with the resulting combination eliciting the initiation of group movement in non-predatory contexts (Arnold & Zuberbühler, 2008). However, since the individual calls here do not appear to contribute to sequence meaning, these call combinations have instead been analysed as idiomatic rather than compositional structures (Arnold & Zuberbühler, 2012). Outside the primates, evidence for syntax comes from two bird species. Pied babblers (*Turdoides bicolor*) combine two functionally distinct vocalizations into a larger sequence when encountering a

terrestrial threat that requires recruiting group members, and playback experiments have indicated that receivers process the call combination compositionally by linking the meaning of the independent parts (Engesser *et al.*, 2016). Experiments on Japanese great tits (*Parus minor*) have also revealed that receivers extract different meanings from 'ABC' (scan for danger) and 'D' notes (approach the caller), and a compound meaning from 'ABC–D' combinations (Suzuki, Wheatcroft & Griesser, 2016).

Outside of sign language systems, non-vocal combinatorial structuring in humans seem to be prevalent in pantomime. Pantomime has been defined as non-verbal and non-conventionalized means of communication, which is executed primarily in the visual channel by coordinated, successive movements of the whole body, but might also incorporate non-linguistic vocalizations (Żywiczyński, Wacewicz & Sibierska, 2018). These movements symbolically encode and communicate meaning independently of language (Xu *et al.*, 2009), and can refer to a potentially unlimited repertoire of events, or sequences of events. For these and other reasons, pantomimic scenarios of language origins have recently gained popularity among experts in the field (e.g. Tomasello, 2008; Zlatev, 2008; Arbib, 2012). However, pantomime as self-contained communicative acts cannot be easily isolated into component parts, as segments would lack obvious discrete boundaries and may not be freely combinable. Individual gestures (e.g. emblems, home-sign systems), by contrast, do have a discrete onset–termination structure and can combine and recombine to form systematic, compositional messages (McNeill, 1992; Goldin-Meadow, 1999; Kendon, 2004; Goldin-Meadow & Alibali, 2013; Clay *et al.*, 2014).

Interestingly, in the non-human gestural domain, evidence for sequences resembling combinatorial structuring (with meaning) is entirely absent. Focusing on gesture sequences in captive chimpanzees, Liebal *et al*. (2004*a*) showed that the majority consisted of repetitions of the same gestures, which were mainly tactile and related to the play context. The emergence of gesture sequences was seen as a by-product of the recipient's lack of responsiveness rather than as a systematic combination of

gestures to increase the efficiency of particular gestures (Liebal *et al.*, 2004*a*). Similarly, Genty, Breuer, Hobaiter *et al.* (2009); and Hobaiter and Byrne (2011b), who examined serial gesturing in gorillas (*Gorilla gorilla*) and chimpanzees respectively, found no evidence for syntactic effects of sequential combinations. Hobaiter & Byrne (2012), like Liebal *et al*. (2004*a*), concluded that gesture bouts are a consequence of persistence in the face of failure. Taken together, at present there is much clearer evidence for combinatorial capacities in animal vocalizations than there is for gestures. Whether human syntax evolved gradually from animal combinatoriality, or emerged more recently as a functional change from non-linguistic operations is still subject to debate (Fitch, 2011; Zuberbühler, 2018).

(5) Turn-taking

The cooperative and rapid exchange of turns is a universal feature of human linguistic interactions (Sacks, Schegloff & Jefferson, 1974; Stivers *et al.*, 2009) and has primarily been studied from a developmental perspective (Levinson & Holler, 2014). Using conversation analysis, the sequential organization of social action *via* turns is investigated to grasp how mutual understanding and the successful engagement of cooperative interactions is achieved (Sacks *et al.*, 1974).

In recent years, turn-taking has gained much research attention in the field of animal communication, building on the premise that it was a key prerequisite for the language system (Takahashi, Narayanan & Ghazanfar, 2013; Henry, Craig, Lemasson *et al.*, 2015; Fröhlich, 2017; Demartsev, Strandburg-Peshkin, Ruffner *et al.*, 2018; Pika, Wilkinson, Kendrick *et al.*, 2018). Vocal turn-taking, defined as the precise, stereotyped coordination of vocal contributions by two individuals (Farabaugh, 1982), evolved independently in a wide variety of taxa, including insects (Bailey, 2003), frogs (Tobias, Viswanathan & Kelley, 1998; Wong *et al.*, 2004), bats (Vernes, 2016) and primates (Geissmann & Orgeldinger, 2000). Nonetheless, turn-taking is best-studied in birds, where duetting and antiphony are widely found (Slater & Mann, 2004; Hall, 2009; Dahlin & Benedict, 2014; Henry *et al.*, 2015). In

non-human primates, evidence for vocal turn-taking has been gathered from all the major clades, among them lemurs, marmosets, titi monkeys, squirrel monkeys and siamangs (for review see Levinson, 2016).

Gestural turn-taking has so far been studied only in great apes, building on a conversation-analytic framework to understand the role of gestural communication, but also other communicative means such as gaze and body orientation. For example, Rossano (Rossano, 2013; Rossano & Liebal, 2014) and Fröhlich (Fröhlich, Kuchenbuch, Müller *et al.*, 2016a) examined the structural, temporal, and even spatial patterns underlying the coordination of joint travel in bonobo and chimpanzee mother–offspring pairs. These and other studies suggest that vocal and gestural exchanges in particular contexts might resemble simple forms of turn-taking, providing evidence that cooperative joint actions might have evolved earlier than previously thought and perhaps even preceded the evolution of language (Rossano, 2013; Fröhlich *et al.*, 2016*a*; Levinson, 2016). However, as long as we ignore that communicative turns can be exchanged *via* multiple means and sensory channels, our knowledge of the complexity inherent to turn-taking behaviour remains incomplete (Fröhlich, 2017). Therefore, the combination of the two paradigms – multimodality and turn-taking – *via* conversation-analytic approaches might allow a more dynamic, holistic study of animal communication by better taking the roles of both signaller and receiver into account. This would allow us to draw more accurate comparisons to the human 'interaction engine' – a package of underlying propensities in human communication, including the face-to-face character that affords the use of gestures and gazes (Levinson, 2006).

(6) Neural control

Neuroscience provides an important line of evidence in favour of a multimodal approach to communication. The purpose of this section is not to cover in detail the neural basis of vocal and gestural production and processing in non-human primates (for this, see the extensive review in

Liebal *et al.*, 2013), but to focus on neural areas that have been highlighted in multimodal processing. In particular, because the different modalities rely on different body parts or organs, we will here cover the processing and integration of signals rather than their initial perception.

The neural bases of mammalian vocalizations, for both production and perception, have been extensively documented (for a recent review, see Hage & Nieder, 2016). Processes in the larynx and supra-laryngeal vocal tract are controlled by several nuclei located in the pons and medulla (Jürgens, 2002; Liebal *et al.*, 2013), with the activity of the motor nuclei controlled in the medulla, itself mediated by the periaqueductal grey of the midbrain (Jürgens, 2002; Liebal *et al.*, 2013). The apparent sole involvement of these subcortical structures has long been relied upon to justify the notion that much of primate vocalizations may be innate. This view was additionally apparently supported by the fact that there was no apparent direct connection between the motor cortex and the basal motor nuclei in primates, which were seen as controlled exclusively by the reticular formation of the medulla, in contrast to humans, where a direct connection between the nucleus ambiguus and the motor cortex allows a direct control of the larynx by the cortex without relying on the medulla network (Jürgens, 1976). However, in an earlier study, Kuypers (1958) had in fact observed these direct connecting neurons in one of three chimpanzee subjects, contradicting what has paradoxically become known as the Kuyper–Jürgen's model (Lameira, 2017). At least for chimpanzees, it thus seems the extent of vocal control is not yet fully understood.

Overall, voluntary control over communicative signals appears to require the involvement of cortical structures, particularly the dorso- and mediofrontal cortices, including the anterior cingulate gyrus and the supplementary and pre-supplementary motor areas (Jürgens, 2002). Thus, cortical involvement can serve as a criterion for voluntary control. In more recent years, some cortical areas (e.g. anterior cingulate gyrus or prefrontal cortex) have been connected to the learning and voluntary production of primate vocalizations, in particular in food-based paradigms (Gemba, Miki & Kazuo, 1995; Coudé *et al.*, 2011; Taglialatela *et al.*, 2012; Hage & Nieder, 2013). A network of ventrolateral

frontal (VLF) and dorsomedial frontal regions (DMF) appears to be homologous across monkeys and humans and could allow cognitive control of vocalization production across primate species (Loh, Petrides, Hopkins *et al.*, 2017). Furthermore, Kumar, Croxson & Simonyan (2016) studied the structural organization of the laryngeal motor cortical (LMC) network in humans and rhesus monkeys, revealing a large overlap in the structural network, although there were differences with regard to connection strength between the LMC and inferior parietal cortices. With respect to processing, two main pathways emerging from the auditory (parietal) cortex, the 'anteroventral' and 'posterodorsal' streams, are well described and both project in the prefrontal cortex (PFC), where the information contained in the calls is integrated (Hage & Nieder, 2016). The anteroventral stream is thought to encode auditory identity ('what?') while the posterodorsal stream is thought to primarily encode auditory space ('where?') (Rauschecker & Scott, 2009). Thus, there is increasing evidence for cortical involvement in both production and processing of auditory stimuli in primates.
Comparatively, the neural network allowing the production and perception of gestures is less well documented (however, see e.g. Meguerditchian *et al.*, 2012), and has been mostly described for the processing of facial expressions (Liebal *et al.*, 2013). Nevertheless, as in the processing of auditory stimuli, a dual stream has also been proposed for visual information (Ungerleider & Mishkin, 1982), with a ventral stream enabling object identification ('what?') and a dorsal stream enabling spatial information ('where?'). The processing of facial stimuli occurs in various areas such as the inferior temporal cortex or the superior temporal sulcus, responding to facial identity or expression (Hasselmo, Rolls & Baylis, 1989). The superior temporal sulcus (STS), in particular, has been shown to react differently to aggressive or affiliative facial expressions compared to neutral ones; its activation is also often identified during the emotional processing of vocalizations and prosody (Gruber & Grandjean, 2017), showing its involvement in both the visual and auditory modality.
In the quest for the neurobiological underpinnings of action/gesture recognition, researchers proposed that perceived visual information is mapped onto its motor representation in the brain

(Rizzolatti, Fogassi & Gallese, 2001). Mirror neurons, originally discovered in the ventral premotor cortex of rhesus macaques (*Macaca mulatta*), were shown to discharge both when an individual performs an action or observes the same action by another individual (Di Pellegrino *et al.*, 1992; Gallese *et al.*, 1996). The discovery of this class of visuo-motor neurons has been widely used to explain how conspecifics can mutually understand each other's actions, bridging the gap from action to communication, as the link between actor and observer resembles the link between the sender and the receiver of each message (Rizzolatti & Arbib, 1998). The 'mirror system hypothesis' states that language eventually became possible because of the mutual understanding of grasping actions enabled by the mirror system, a strong argument in favour of the gestural-first theory (Arbib, 2012). However, this claim now seems controversial given that the precise role of mirror neurons in the processes through which individuals understand the actions of others (rather than those involved in an individual's own motor control) remains unclear (Hickok, 2009; Kendon, 2017).

With respect to multimodality, a region of major interest appears to be the PFC, which includes several well-established areas for studies of language such as Broca's area, located in the inferior frontal gyrus (IFG) of the granular ventrolateral PFC (vlPFC). While, the involvement of Broca's area in language processing is well described, recent studies have shown that the IFG, at least in humans, is able to process and integrate speech and gestures simultaneously (Homae, Hashimoto, Nakajima *et al.*, 2002; Xu *et al.*, 2009). The vlPFC is the homologue of the IFG in the primate brain; it is to be noted that the lateral PFC also contains face-selective neurons (Scalaidhe, Wilson & Goldman-Rakic, 1997; Tsao, Schweers, Moeller *et al.*, 2008) and that generally, the vlPFC allows integrating information from several communicative means. As described by Hage and Nieder (2016), the "largely segregated visual and auditory pathways converge in the vlPFC to give rise to neurons that represent higher-order multisensory and categorical representations of communicative signals". While Hage and Nieder (2016) suggest that this may allow the integration of the fixed correspondence between vocalizations and facial expressions associated with them, we may also

hypothesize that such an integration centre could also more generally facilitate multimodal communication, both in its production and perception aspects. There appears to be strong overlap in the brain structures and circuits involved in production and processing of visual or gestural and vocal stimuli. Moreover, neurocognitive studies of motor representations of speech sounds, action-related language, sign language and co-speech gestures present strong evidence that the processing of gestures and vocalizations is tightly interlinked in the brain (Kimura, 1993; Willems & Hagoort, 2007). The fact that hand and mouth actions are controlled by very closely related systems underpins the co-involvement of hand and mouth in both language and practical activities. Taken together, research on the proximate mechanisms underlying primate communication, both with regard to production and perception, points towards a tight integration in the brain of co-occurring signals from multiple domains, in non-human primates and humans alike, as expected when communication is inherently multimodal.

(7) Ontogenetic plasticity: the impact of learning

Apart from neurobiological mechanisms underlying communicative acts, an understanding of how communication develops is essential for the proximate perspective on behaviour. Importantly, the ability to produce, actually use, and comprehend signals may all show different developmental pathways**,** suggesting that different cognitive prerequisites might be involved (Liebal *et al.*, 2013). The extent to which structure and usage of vocalizations and gestures are impacted by the social and physical environment during development has been under much discussion (Cheney & Seyfarth, 2018; Fröhlich & Hobaiter, 2018). Research on gestural communication in great apes in captivity made the case for an individual learning mechanism, 'ontogenetic ritualization', in which gestures originate *via* shortening of a functional action sequence (Tomasello *et al.*, 1994; Halina, Rossano & Tomasello, 2013). By contrast, studies on wild communities provided evidence that the available repertoires are largely innate and species-typical (Genty *et al.*, 2009; Hobaiter & Byrne, 2011a;

Graham, Furuichi & Byrne, 2016). After a lively debate on the mechanisms of gesture acquisition, several researchers concluded that the forms of gesture types (the 'tool-set' or available repertoires) are largely genetically anchored (Byrne *et al.*, 2017), whereas their usage in relation to their context and social environment (the 'tool application') is affected by interactional experiences throughout life (Bard *et al.*, 2014; Fröhlich & Hobaiter, 2018; Liebal, Schneider & Errson-Lembeck, 2018; Pika & Fröhlich, 2018). Early gestural communication in chimpanzees, for instance, is influenced by the number of interaction partners and interaction rates with non-maternal conspecifics (Fröhlich, Müller, Zeiträg *et al.*, 2017). It is also possible that the captive setting fosters more gestural "inventions" and idiosyncratic gestures than the wild, due to richer social opportunities and more repeated interactions with the same individuals. While different mechanisms of acquiring gesture types might be involved, it is always critical to ensure sufficient sampling effort across research settings.

Vocal production learning is often highlighted as a critical stepping-stone in language evolution, but the evidence from primates remains scarce and is mostly limited to unvoiced calls not involving the vocal tract (Hopkins, Taglialatela & Leavens, 2007; Wich *et al.*, 2009; Lameira *et al.*, 2016; however see Watson *et al*., 2015; Crockford *et al*., 2004). Cases of vocal inventions have been described but seem to be exceptional, like the customary 'throat scrape' produced by orang-utan mothers at Tuanan, Borneo (van Schaik, van Noordwijk & Wich, 2006). Although the structure of primate vocalization types appears to be largely fixed, social input from the environment can substantially influence the usage of specific vocalizations (e.g. Snowdon & Hausberger, 1997; Laporte & Zuberbühler, 2011; Katsu, Yamada & Nakamichi, 2017; Lameira, 2017; Cheney & Seyfarth, 2018). Studies on vocal development in birds and mammals have demonstrated that individual experiences accumulated through social interactions (e.g. responses of conspecifics) can play a substantial role by introducing new sounds and encouraging improvization (Snowdon & Hausberger, 1997). Therefore, although the morphology and structure of signals seems to be genetically channelled in both the

gestural and the vocal domains, we see profound developmental plasticity in the usage of both these communicative modalities.

Building on pleas for more explicit cross-modal study designs (Liebal *et al.*, 2013; Townsend *et al.*, 2017), Fröhlich, Wittig and Pika (2018) recently provided the first study on the developmental trajectory of intentional communication in chimpanzees by using an multimodal approach. Both gestures and gesture–vocal combinations, but also vocal signals, were frequently accompanied by specific intentionality markers (i.e. audience checking, goal persistence and sensitivity to recipient's attention). Their findings showed that intentional signal use is not only affected by age, but also by variables related to social circumstances, such as communicative context, interaction partner and group membership. In light of accumulating studies demonstrating a substantial impact of social experiences on communicative development, it is now vital to understand the role of learning and social experience in both unimodal and multimodal signal production (see also Higham & Hebets, 2013). This will further elucidate whether the same underlying mechanisms are at play as in human language acquisition and origins. For instance, unimodal signals might develop at an earlier age in primates, because multimodal signal combinations may require key socio-cognitive skills that develop only later in ontogeny. A similar pattern is found in human children, who employ gestures first ('pre-linguistic stage') and acquire language passing a so-called 'one-word stage' with a frequent use of bimodal combinations (see for review Bretherton & Bates, 1979). An alternative developmental scenario, however, suggests that multimodal communication emerges first and is later tuned to the most effective, unimodal signals (Liebal *et al.*, 2013). Preliminary support for this explanation comes from studies on chimpanzees (Bard *et al.*, 2014; Fröhlich *et al.*, 2016*b*), but much more empirical and theoretical work is needed to confirm this. In sum, evidence for flexibility, effects of social exposure and interactional experiences on communicative usage is increasing for the gestural as well as the vocal domain.

IV. IMPLICATIONS OF 'MULTIMODALISM' FOR LANGUAGE-EVOLUTION SCENARIOS

(1) Taking multimodality seriously – burying the hatchet

With the exception of combinatorial signal sequences and perhaps iconicity, evidence for the other widely acknowledged language components – intentionality, reference, turn-taking, ontogenetic plasticity – is found in both the gestural and vocalization domains in animals (Table 2). Most of the cognitive properties that have been emphasized in human communication have been identified in both gestures and vocalizations of non-human species, albeit to various extents. Unfortunately, we cannot draw firmer conclusions due to major discrepancies in the definition and operationalization of cognitive concepts and the extent of the parallels in terms of cognitive processing and frequencies of use in the repertoire within and across species. Even so, recent work on proximate control involved in primate communication is also in support of 'multimodalism'. The heated debate around the likeliest language precursor therefore can be put to rest: a gestures-first *versus* vocalizations-first opposition of language origins can no longer be supported. We therefore conclude that it is time to bury the hatchet about whether the origins of human communication lie in the vocal or gestural modality and acknowledge that the puzzle of language evolution can only be tackled by investigating animal communication as multimodal signalling. In this spirit, investigations recently started endorsing a focus on both signallers and receivers in a communicative interaction, emphasizing the importance of flexible, multimodal interactions and turn-taking behaviour (Slocombe *et al.*, 2011; Waller *et al.*, 2013; Levinson, 2016; Fröhlich, 2017; Sievers *et al.*, 2017; Rossano, 2018).

Given that both the ancestral state and the eventual outcome were multimodal, we endorse recent proposals that a multimodal origin is the most likely scenario for the evolution of modern language (e.g. Taglialatela *et al.*, 2011; Lameira, Hardus & Wich, 2012; Gillespie-Lynch *et al.*, 2014; Wacewicz & Zywiczynski, 2017). By itself, however, this conclusion does not solve the problem of

the relationship between the vocal and gestural domains in language evolution. Below we address some major remaining issues as well as promising research avenues.

(2) The transition problem

A key question remains. While both human and non-human primate communication are both multimodal and share numerous features thought to be essential for language, such as multimodality, intentionality, flexibility and ontogenetic plasticity, this does not mean that the same signals used by primates continued to be the main information carriers in human communication. Here, we ask how gestures and vocalizations changed from the ancestral state to human language, under the assumption that human language evolved from a communication system very similar to that of the extant great apes.

(*a*) *Human 'co-speech gestures'* versus *non-human 'gestures'*

In comparative research on non-human primates, gestures are usually defined as socially directed, goal-oriented, mechanically ineffective movements of the extremities or body, or body postures (Pika, 2008; e.g. Bard *et al.*, 2014; Hobaiter & Byrne, 2014), which, unlike human co-speech gestures, are thought to affect the receiver *via* three sensory channels: vision, audition and touch. In light of the pervasive use of manual gestures across human cultures and ages, researchers have turned particularly to non-human primates to seek homologues (e.g. Hewes, 1973; Corballis, 2002; Armstrong & Wilcox, 2007). Since a remarkable number of gesture types in the naturally used repertoire is shared among genera and species, they are thought to be phylogenetically quite old (Byrne *et al.*, 2017). What historically began with a focus on manual gestures limited to movements of the upper extremities in enculturated apes trained to use sign language (Gardner & Gardner, 1969; Premack & Premack, 1972; Patterson, 1978) has gradually expanded to include most of the bodily communication of primates. Gestures as defined by comparative research encompass movements of

the entire body and body postures that are by definition inseparable from animal displays. This has led to a number of misconceptions, ultimately leading to a conceptual divorce from research on other communication modes (Fröhlich & Hobaiter, 2018). Nonetheless, the use of the term gesture became increasingly popular, and other researchers then demonstrated that the intentionality criterion was also fulfilled in 'less manual' species such as corvids (Pika & Bugnyar, 2011) and even fish (Vail *et al.*, 2013).

Despite the fundamentally different definitions of 'gesture' in human and non-human research, comparative research often evaluates findings as if they constitute the very same thing. On the one hand, increasing evidence shows that a substantial proportion of gesture types used by human children is present in the ape repertoire (Blake, 2000; Gillespie-Lynch *et al.*, 2014; Juvrud, Bakker, Kaduk *et al.*, 2018; Kersken, Gómez, Liszkowski *et al.*, 2018). However, this is not found for speech-accompanying ('co-speech') gestures, which are the best-studied facet of human non-conventional gestural communication (Goldin-Meadow, 1999; McNeill, 2000; Kendon, 2004). They are mainly described as manual action in visual space and tightly connected with talk in timing, meaning, and function (McNeill, 1992; Kendon, 2004). These 'illustrators' (beats, iconics, deictics and metaphorics) are thought to represent derived forms of bodily communication that emerged after the onset of speech in the human lineage. Beat gestures consist of short, repetitive, rhythmic movements that mark the tempo of speech, deictic gestures point out referents of speech. Both iconic and metaphoric gestures exploit imagery to elaborate the contents of speech. While iconic gestures capture aspects (spatial images, actions, objects, people) of the semantic content of speech (e.g. an arced cupped hand in the air when describing how water was poured from a glass into a dish), metaphoric gestures spatially represent abstract ideas or concepts (e.g. a circling movement of the hand when describing the passing of time). While some have argued that we find homologous examples of iconic gesture use among non-human primates (Tanner & Byrne, 1996; Perlman, Tanner & King, 2012; Kendon, 2017), we surmise that the non-human gestures studied to date are in fact

very different from the majority of speech-accompanying gestures (in particular beats and metaphorics) used in human communication. While some iconic and deictic gestures are also used in the primate order (e.g. pointing and pantomime, outstretched hands or spread arms when we approach someone to greet, or the gestures used by hunters to direct hunting partners; Hindley, 2014), emblems, beat and metaphorical gestures are derived forms of communication without any homologous equivalent in the primate order. We therefore conclude that some major transitions must have taken place in the gestural domain that still need to be understood. Indeed, speech-accompanying gestures probably evolved without replacing existing gestures.

(*b*) *Human words* versus *primate vocalizations*

Similar to the study of gesture, we see the same discrepancy between the vocalizations of non-human primates and human words. It has been argued that the human equivalents of animal vocalizations are rather non-verbal affective expressions, such as laughing and crying (Ekman & Friesen, 1969; Corballis, 2002; Scherer, Johnstone & Klasmeyer, 2003). Laughter and crying may represent the leftover fragments of an originally much larger innate call system, which may have been critical for the transition between innate and learned vocalizations (Deacon, 1992). By contrast, human speech is composed of discrete, mostly meaningless segments (phonemes and syllables) forming discrete meaningful units (words and phrases), which has been referred to as "duality of patterning" (Hockett, 1960). Using a limited repertoire of phonemes and syllables, and a larger but still finite repertoire of words, we are able to generate an unlimited number of ideas and concepts. Within the cultural diversity of sounds and words of human language, phonemes are the innate components that are rapidly channelled through early experience (Ruben, 1997; Kuhl, 2004).

Despite increasing evidence for call combinations in non-human species (Arnold & Zuberbühler, 2008; Ouattara *et al.*, 2009; Engesser *et al.*, 2016), signal combinations of non-human animals do not necessarily involve homologous elements to human words. In addition, findings demonstrating

voluntary control and learning capacities in great apes often focused on unvoiced calls, that is, sounds not involving the vocal tract (Hopkins *et al.*, 2007; Wich *et al.*, 2009; Lameira *et al.*, 2016; however see Watson *et al*., 2015). Hence, there might be a similarly large evolutionary gap between a beckoning gesture and human pointing as between a context-specific 'hoo' vocalization (e.g. Crockford, Gruber & Zuberbühler, 2018) and a spoken word. In other words, there are large gaps between the human and non-human forms of both gestures and vocalizations, and they cannot be easily bridged by current comparative evidence.

(3) The switch of carrying roles in human language

Only recently have studies begun to focus on gestural–vocal combinations in great ape face-to-face communication. Since gestures seem to carry most communicative meaning, they (together with the rich body of work emphasizing high gestural frequencies and repertoires across social contexts) appear to suggest that gestures in great apes possess a dominant, carrying role in close-range communication (Genty *et al.*, 2014; Hobaiter *et al.*, 2017; Wilke *et al.*, 2017; see Tab. 2). Importantly, we see the opposite pattern in humans: speech-accompanying gestures complement and refine the message conveyed in speech, but they (in particular beat and metaphorical gestures) clearly would not work on their own (Goldin-Meadow, 1999). This contrast therefore implies that there has been an evolutionary switch in carrying roles for close-distance communication (see also Mühlenbernd *et al.*, 2014): while the gestural domain seems to play the major role in information transfer in short-distance communication in great apes, both on its own and for disambiguating the vocal message (Genty *et al.*, 2014; Hobaiter *et al.*, 2017; Wilke *et al.*, 2017; Fröhlich *et al.*, 2018), in (adult) human communication they mainly serve to support the vocal medium (McNeill, Cassell & McCullough, 1994; Goldin-Meadow, 1999; Kendon, 2004). Fig. 1 sketches this transition in the carrying roles of the gestural and vocal domains. With this view we are not returning to the gestural

theory on language origins, because it does not see vocalizations as being merely 'bootstrapped' on gestures (Tomasello, 2008; Fay *et al.*, 2013).

(4) Scenarios for language evolution

Let us examine scenarios for the changes involved in the evolution of human language. In the scenario recently proposed by Levinson & Holler (2014), human communication gradually added new layers of communicative abilities over phylogenetic time (Fig. 2A). They argue that ritualized gestures, dyadic turn-taking and later on iconic gestural representations (i.e. the 'interaction engine') preceded the development of the voluntary breathing control that enables complex vocalization. Accordingly, vocalization complemented the pre-existing repertoire of iconic and deictic gestures, and subsequently coevolved with it for nearly a million years, resulting in the tight integration of vocal and gestural modalities in human communication (Kendon, 2000; McNeill, 2000; Kendon, 2004). Because gesturing and speaking are elements of a single process of utterance generation, it seems unreasonable to hold up a dual-modality view of language rather than that of a unified system: speech-related mouth actions have always been dynamically integrated with manual gestures as well as other bodily actions (Kendon, 2004; Gentilucci & Corballis, 2006). As Kendon (2017, p.168) notes, "the co-involvement of gesturing with speech—where gestures are schematic forms abstracted from practical action—indicates that languaging is derived from practical action", suggesting "that speaking, like co-occurring manual gesturing, is manipulatory activity in the abstract". While this scenario certainly has great explanatory power, it ignores the fact that many of the speech-accompanying gestures we see every day in modern human communication might have very little to do with those evolutionarily old ritualized gestures. Rather, language might have originated from iconic communication coordinated across both the vocal *and* gestural modalities (see also Perlman, 2017).

We therefore propose a modification to this scenario, in which pressures on signal efficacy decreased during human evolution, due to increasing visibility of social partners, reduced inter-individual distance, and thus decreasing environmental noise (Fig. 2B). This enabled more opportunities for dyadic, face-to-face social interactions between social partners with high levels of familiarity (i.e. common interactional experience). Because efficacy was no longer an issue, it became possible to consistently transmit different information *via* multiple channels simultaneously, resulting in increased information content (complementarity) of multimodal signals. Joint attention (i.e. two individuals coordinating their attention to an entity of mutual interest: Tomasello, 1995) increasingly involved multiple sensory channels and paved the way for complex declarative interactions, which are considered a critical precondition for the emergence and development of language (Butterworth, 2001; Bard & Leavens, 2009; Leavens & Racine, 2009). With the rise of shared intentionality based on established common psychological ground, human communication has therefore become much more cooperative than communication systems in the rest of the primate order (Tomasello, Carpenter, Call *et al.*, 2005; Tomasello & Carpenter, 2007; Aureli, Perucchini & Genco, 2009; Burkart, Hrdy & van Schaik, 2009). Although human communication is not only used in cooperative (but also in deceptive or imperative) ways (Zuckerman, DePaulo & Rosenthal, 1981; Porter & Yuille, 1996), it is widely held that it was the motivation to donate information which facilitated the origin of human language (Tomasello, 2008; van Schaik, 2016).

The key point is that richer messages could be created more easily in the vocal stream, which thus came to play the leading role in the evolution of human language following the emergence of vocal-production learning. Initially, the vocal stream was inadequate for this: great ape's vocalizations, compared to gestures, are characterized by small repertoires and modest plasticity (Table 1). But the use of unvoiced sounds such as whistles, raspberries, clicks, smacks, etc. (Hopkins *et al.*, 2007; Wich *et al.*, 2009; Lameira *et al.*, 2016) might have enabled our ancestors to bridge this gulf. Given how widespread redundancy (i.e. different signals or signal components conveying the same meaning and

thus eliciting the same response; e.g. Partan & Marler, 1999) is in multisensory animal communication, it is highly intriguing that redundancy and flexibility of use seem to play a minor role in primate vocal communication compared to gestures. This has previously been related to the higher urgency of vocalizations (Tomasello & Zuberbühler, 2002), but recent studies have shown that gestures can also be used in 'urgent' contexts (Hobaiter & Byrne, 2012; Fröhlich *et al.*, 2016b), similar to vocalizations being used in non-urgent contexts (e.g. Crockford *et al.*, 2018). Fitch *et al.* (2016) recently demonstrated that monkey's vocal tracts are 'speech-ready', hence selection could rapidly enrich the vocal plasticity and thus repertoire. These shifts from great ape to human close-range communication are illustrated in Fig. 1. However, what remains unknown is how richer call repertoires could give rise to phonemes and syllables.

While speech-accompanying gestures serve to underscore and support emphases in verbal communication, the vocal modality can carry far more detailed and diverse meaning than the gestural stream (sign language can be learned but is not a naturally developing, ubiquitous activity, and thus likely piggybacks on the abilities of production and perception that evolved for the vocal stream). Our communication system depends fundamentally on cooperative information sharing, common ground and semantic reference (Tomasello, 2008), and it would be useful to focus more on those derived features in future work. We need to stress that there are clear motivational differences in sharing information between humans and great apes. Specifically, after communication became truly cooperative and our ancestors' '*Mitteilungsbedürfnis'* (i.e. the need or eagerness to inform others and share meanings; Fitch, 2010) arose with it, it was probably the vocal stream that became enhanced. Prosocial behaviour and common ground might have been the fundamental starting ground for this (Fig. 2B): it is striking that great ape pointing and referential signalling is commonly observed in interactions with caretakers 'altruistically' providing food but rare in intra-specific interactions (Leavens *et al.*, 2005*a*). The gestural stream responded by becoming the supportive activity to

improve the efficacy of the vocal stream, to the point that vocal streams alone, stripped of all stresses added by our gestures in the broad sense, is much harder to understand.

V. CONCLUSIONS

(1) We aimed to summarize current evidence from comparative research in favour of a multimodal origin of language. We showed that there are no clear differences in the extent to which key language components are present in gesture and vocalization (despite discrepancies in form and function), and that the neurobiological regulation of both shows great overlap and integration, as well as developmental plasticity. This conclusion indicates that animal communication is fundamentally multimodal, and thus shows the futility of a gestures-first or a vocalizations-first origin of human communication.

(2) An exciting future avenue would be to examine in more detail how and why such a switch or 'role reversal' between the vocal and the gestural stream in combinations, from supporter to carrier and *vice versa*, took place during human evolution. A first step might be to look at the rich opportunities for compositionality in primate vocalizations (e.g. Arnold & Zuberbühler, 2008), as opposed to the redundancy present in the ape gestural repertoire (e.g. Byrne *et al.*, 2017).

(3) Specifically, signal structure in the vocal domain seems to be considerably more rigid and discrete, and repertoire sizes smaller than in the gestural domain – which animals apparently overcome through combinatoriality. By contrast, compositionality in signal sequences is much less evident in the gestural domain (Liebal *et al.*, 2004a; Hobaiter & Byrne, 2011b), where many different signal types are used to achieve the same goal and are thus redundant in meaning (Tomasello *et al.*, 1994).

(4) At the same time, it is critical to consider more derived features like the fundamental cooperative nature of language, by examining the links between sociality, cooperative tendency and

communicative complexity across species. Consistency in definitions will be immensely important for future research efforts.

VI. ACKNOWLEDGEMENTS

We are grateful to Judith Burkart and two anonymous reviewers for useful comments on the manuscript, and to Sabrina Engesser, Cat Hobaiter, Colin Wagner, Maria van Noordwijk and Uli Knief for insightful discussions on multimodality. M.F. was supported by the Forschungskredit Postdoc of the University of Zurich (grant FK-17-106), and a Research Fellowship of the German Research Foundation (grant FR 3986/1-1). T.G. was supported by the Swiss National Science Foundation (grants CR13I1_162720 and P300PA_164678).

VII. REFERENCES

AHLNER, F. & ZLATEV, J. (2010). Cross–modal iconicity: a cognitive semiotic approach to sound symbolism. *Sign Systems Studies* **38**, 298–348.

ARBIB, M. A. (2012). *How the Brain Got Language: The Mirror System Hypothesis*. Oxford University Press.

ARBIB, M. A., ABOITIZ, F., BURKART, J. M., CORBALLIS, M., COUDÉ, G., HECHT, E., LIEBAL, K., MYOWA–YAMAKOSHI, M., PUSTEJOVSKY, J. & PUTT, S. (2018). The comparative neuroprimatology 2018 (CNP–2018) road map for research on How the Brain Got Language. *Interaction Studies* **19**, 370–387.

ARBIB, M. A., LIEBAL, K. & PIKA, S. (2008). Primate vocalization, gesture, and the evolution of human language. *Current Anthropology* **49**, 1053–1063.

ARMSTRONG, D. F. & WILCOX, S. E. (2007). *The Gestural Origin of Language*. Oxford University Press, Oxford.

ARNOLD, K. & ZUBERBÜHLER, K. (2006). Language evolution: semantic combinations in primate calls. *Nature* **441**, 303.

ARNOLD, K. & ZUBERBÜHLER, K. (2008). Meaningful call combinations in a non–human primate. *Current Biology* **18**, R202–R203.

ARNOLD, K. & ZUBERBÜHLER, K. (2012). Call combinations in monkeys: compositional or idiomatic expressions? *Brain and Language* **120**, 303–309.

ARNOLD, K. & ZUBERBÜHLER, K. (2014). Primate pragmatics: putty–nosed monkeys use contextual information to disambiguate the cause of alarm calls. *Folia Primatologica* **84**, 375–376.

ATKINSON, E. G., AUDESSE, A. J., PALACIOS, J. A., BOBO, D. M., WEBB, A. E., RAMACHANDRAN, S. & HENN, B. M. (2018). No Evidence for Recent Selection at FOXP2 among Diverse Human Populations. *Cell* **174**, 1424–1435.

AURELI, T., PERUCCHINI, P. & GENCO, M. (2009). Children's understanding of communicative intentions in the middle of the second year of life. *Cognitive Development* **24**, 1–12.

BACH, K. (2006). What Does it Take to Refer? In *The Oxford Handbook of Philosophy of Language* (ed. E. Lepore and B. C. Smith), pp. 516–555. Clarendon Press, Oxford.

BAILEY, W. J. (2003). Insect duets: underlying mechanisms and their evolution. *Physiological Entomology* **28**, 157–174.

BARD, K. A., DUNBAR, S., MAGUIRE–HERRING, V., VEIRA, Y., HAYES, K. G. & MCDONALD, K. (2014). Gestures and social–emotional communicative development in chimpanzee infants. *American Journal of Primatology* **76**, 14–29.

BARD, K. A. & LEAVENS, D. A. (2009). Socio–emotional factors in the development of joint attention in human and ape infants. In *Learning from animals? Examining the nature of human uniqueness* (ed. L. S. Röska–Hardy and E. M. Neumann–Held), pp. 89–104. Psychology Press, Hove, East Sussex.

BARRETT, J., ABBOTT, D. & GEORGE, L. (1990). Extension of reproductive suppression by pheromonal cues in subordinate female marmoset monkeys, Callithrix jacchus. *Journal of Reproduction and Fertility* **90**, 411–418.

BEECHER, M. D. & BRENOWITZ, E. A. (2005). Functional aspects of song learning in songbirds. *Trends in Ecology & Evolution* **20**, 143–149.

BICKERTON, D. (1992). *Language and species*. University of Chicago Press.

BICKERTON, D. & SZATHMÁRY, E. (2011). Confrontational scavenging as a possible source for language and cooperation. *BMC Evolutionary Biology* **11**, 261.

BLAKE, J. (2000). *Routes To Child Language: Evolutionary And Developmental Precursors*. Cambridge University Press.

BOHN, M., CALL, J. & TOMASELLO, M. (2016). Comprehension of iconic gestures by chimpanzees and human children. *Journal of Experimental Child Psychology* **142**, 1–17.

BOHN, M., CALL, J. & TOMASELLO, M. (2019). Natural reference: a phylo– and ontogenetic perspective on the comprehension of iconic gestures and vocalizations. *Developmental Science* **22,** e12757.

BRAINARD, M. S. & DOUPE, A. J. (2002). What songbirds teach us about learning. *Nature* **417**, 351–358.

BRETHERTON, I. & BATES, E. (1979). The emergence of intentional communication. *New Directions for Child and Adolescent Development* **1979**, 81–100.

BRUMM, H. & SLABBEKOORN, H. (2005). Acoustic Communication in Noise. In *Advances in the Study of Behavior*, vol. Volume 35, pp. 151–209. Academic Press.

BURKART, J. M., HRDY, S. B. & VAN SCHAIK, C. P. (2009). Cooperative breeding and human cognitive evolution. *Evolutionary Anthropology: Issues, News, and Reviews* **18**, 175–186.

BUTTERWORTH, G. (2001). Joint visual attention in infancy. In *Blackwell Handbook of Infant Development* (ed. J. G. Brenner and A. Fogel), pp. 213–240. Blackwell, Oxford, England.

BYRNE, R. W., CARTMILL, E., GENTY, E., GRAHAM, K. E., HOBAITER, C. & TANNER, J. (2017). Great ape gestures: intentional communication with a rich set of innate signals. *Animal Cognition* **20,** 755–769.

CASSELL, J., MCNEILL, D. & MCCULLOUGH, K.–E. (1999). Speech–gesture mismatches: evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition* **7**, 1–34.

CHENEY, D. L. & SEYFARTH, R. M. (1981). Selective forces affecting the predator alarm calls of vervet monkeys. *Behaviour* **76**, 25–60.

CHENEY, D. L. & SEYFARTH, R. M. (1985). Social and non–social knowledge in vervet monkeys. *Philosophical Transactions of the Royal Society B: Biological Sciences* **308**, 187–201.

CHENEY, D. L. & SEYFARTH, R. M. (1996). Function and intention in the calls of non–human primates. In *Proceedings of the British Academy*, vol. 88, pp. 59–16.

CHENEY, D. L. & SEYFARTH, R. M. (2005). Constraints and preadaptations in the earliest stages of language evolution. *The Linguistic Review* **22,** 135–159.

CHENEY, D. L. & SEYFARTH, R. M. (2018). Flexible usage and social function in primate vocalizations. *Proceedings of the National Academy of Sciences*.

CHOMSKY, N. (1959). A review of BF Skinner's verbal behavior. *Language* **35**, 26–58.

CLAY, Z., POPLE, S., HOOD, B. & KITA, S. (2014). Young children make their gestural communication systems more language–like: segmentation and linearization of semantic elements in motion events. *Psychological Science* **25**, 1518–1525.

CLAY, Z., SMITH, C. L. & BLUMSTEIN, D. T. (2012). Food–associated vocalizations in mammals and birds: what do these calls really mean? *Animal Behaviour* **83**, 323–330.

COLLIER, K., BICKEL, B., VAN SCHAIK, C. P., MANSER, M. B., & TOWNSEND, S. W. (2014). Language evolution: syntax before phonology? *Proceedings of the Royal Society B: Biological Sciences* **281**, 20140263.

CORBALLIS, M. C. (2002). *From Hand to Mouth: The Origins of Language*. Princeton University Press, Princeton, NJ.

COUDÉ, G., FERRARI, P. F., RODÀ, F., MARANESI, M., BORELLI, E., VERONI, V., MONTI, F., ROZZI, S. & FOGASSI, L. (2011). Neurons controlling voluntary vocalization in the macaque ventral premotor cortex. *PLoS ONE* **6**, e26822.

CROCKFORD, C., GRUBER, T. & ZUBERBÜHLER, K. (2018). Chimpanzee quiet hoo variants differ according to context. *Royal Society Open Science* **5**, 172066.

CROCKFORD, C., HERBINGER, I., VIGILANT, L. & BOESCH, C. (2004). Wild chimpanzees produce group‐specific calls: a case for vocal learning? *Ethology* **110**, 221‐243.

CROCKFORD, C., WITTIG, R. M., MUNDRY, R. & ZUBERBÜHLER, K. (2012). Wild chimpanzees inform ignorant group members of danger. *Current Biology* **22**, 142–146.

CROCKFORD, C., WITTIG, R. M. & ZUBERBÜHLER, K. (2015). An intentional vocalization draws others' attention: A playback experiment with wild chimpanzees. *Animal Cognition* **18**, 581–591.

CROCKFORD, C., WITTIG, R. M. & ZUBERBÜHLER, K. (2017). Vocalizing in chimpanzees is influenced by social–cognitive processes. *Science Advances* **3**, e1701742.

DAHLIN, C. R. & BENEDICT, L. (2014). Angry birds need not apply: a perspective on the flexible form and multifunctionality of avian vocal duets. *Ethology* **120**, 1–10.

DEACON, T. W. (1992). The neural circuitry underlying primate calls and human language. In *Language Origins: A Multidisciplinary Approach*, pp. 121–162. Springer.

DEDIU, D. & LEVINSON, S. (2013). On the antiquity of language: the reinterpretation of Neandertal linguistic capacities and its consequences. *Frontiers in Psychology* **4**, 397.

DEMARTSEV, V., STRANDBURG–PESHKIN, A., RUFFNER, M. & MANSER, M. (2018). Vocal turn–taking in meerkat group calling sessions. *Current Biology* **28**, 3661–3666.e3.

DENNETT, D. (1983). Intentional systems in cognitive ethology: the 'Panglossian paradigm' defended. *Behavioral and Brain Sciences* **6**, 343–390.

DI PELLEGRINO, G., FADIGA, L., FOGASSI, L., GALLESE, V. & RIZZOLATTI, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research* **91**, 176–180.

DOUGLAS, P. H. & MOSCOVICE, L. R. (2015). Pointing and pantomime in wild apes? Female bonobos use referential and iconic gestures to request genito–genital rubbing. *Scientific Reports* **5**, 13999.

DUCHEMINSKY, N., HENZI, S. P. & BARRETT, L. (2014). Responses of vervet monkeys in large troops to terrestrial and aerial predator alarm calls. *Behavioral Ecology* **25**, 1474–1484.

EKMAN, P. (1973). Cross–cultural studies of facial expressions. In *Darwin and facial expressions* (ed. P. Ekman), pp. 169–222. Academic Press, New York, London.

EKMAN, P. & FRIESEN, W. (1969). The repertoire of non–verbal behavior: categories, origins, usage and coding. *Semiotica* **1**, 49–98.

ELLESTRÖM, L. (2017). Bridging the gap between image and metaphor through cross–modal iconicity. *Dimensions of Iconicity* **15**, 167.

ENGESSER, S., RIDLEY, A. R. & TOWNSEND, S. W. (2016). Meaningful call combinations and compositional processing in the southern pied babbler. *Proceedings of the National Academy of Sciences* **113**, 5976–5981.

FARABAUGH, S. M. (1982). The ecological and social significance of duetting. *Acoustic communication in birds* **2**, 85–124.

FAY, N., ARBIB, M. & GARROD, S. (2013). How to bootstrap a human communication system. *Cognitive Science* **37**, 1356–1367.

FAY, N., LISTER, C. J., ELLISON, T. M. & GOLDIN–MEADOW, S. (2014). Creating a communication cystem from scratch: gesture beats vocalization hands down. *Frontiers in Psychology* **5**, 354.

FISKE, J. (1863). The evolution of language. *The North American Review* **97**, 411.

FITCH, W. (2011). The evolution of syntax: an exaptationist perspective. *Frontiers In Evolutionary Neuroscience* **3**, 9.

FITCH, W. T. (2010). *The Evolution of Language*. Cambridge University Press, Cambridge.

FITCH, W. T., DE BOER, B., MATHUR, N. & GHAZANFAR, A. A. (2016). Monkey vocal tracts are speech–ready. *Science Advances* **2**, e1600723.

FRISHBERG, N. (1975). Arbitrariness and iconicity: historical change in American Sign Language. *Language* **51,** 696–719.

FRÖHLICH, M. (2017). Taking turns across channels: conversation–analytic tools in animal communication. *Neuroscience & Biobehavioral Reviews* **80**, 201–209.

FRÖHLICH, M. & HOBAITER, C. (2018). The development of gestural communication in great apes. *Behavioural Ecology and Sociobiology* **72**, 194.

FRÖHLICH, M., KUCHENBUCH, P., MÜLLER, G., FRUTH, B., FURUICHI, T., WITTIG, R. M. & PIKA, S. (2016a). Unpeeling the layers of language: bonobos and chimpanzees engage in cooperative turn–taking sequences *Scientific Reports* **6**, 25887.

FRÖHLICH, M., MÜLLER, G., ZEITRÄG, C., WITTIG, R. M. & PIKA, S. (2017). Gestural development of chimpanzees in the wild: the impact of interactional experience. *Animal Behaviour* **134**, 271–282.

FRÖHLICH, M. & VAN SCHAIK, C. P. (2018). The function of primate multimodal communication. *Animal Cognition* **21,** 619–629.

FRÖHLICH, M., WITTIG, R. M. & PIKA, S. (2016b). Should I stay or should I go? Initiation of joint travel in mother–infant dyads of two chimpanzee communities in the wild. *Animal Cognition* **19**, 483–500.

FRÖHLICH, M., WITTIG, R. M. & PIKA, S. (2019). The ontogeny of intentional communication in chimpanzees in the wild. *Developmental Science* **22**, e12716.

GALLESE, V., FADIGA, L., FOGASSI, L. & RIZZOLATTI, G. (1996). Action recognition in the premotor cortex. *Brain* **119**, 593–609.

GARDNER, R. A. & GARDNER, B. T. (1969). Teaching sign language to a chimpanzee. *Science* **165**, 664–672.

GARROD, S., FAY, N., LEE, J., OBERLANDER, J. & MACLEOD, T. (2007). Foundations of representation: where might graphical symbol systems come from? *Cognitive Science* **31**, 961–987.

GARROD S., FAY, N., ROGERS, S., WALKER, B. & SWOBODA, N. (2010). Can iterated learning explain the emergence of graphical symbols? *Interaction Studies* **11**, 33–50.

GEISSMANN, T. & ORGELDINGER, M. (2000). The relationship between duet songs and pair bonds in siamangs, *Hylobates syndactylus*. *Animal Behaviour* **60**, 805–809.

GEMBA, H., MIKI, N. & KAZUO, S. (1995). Cortical field potentials preceding vocalization and influences of cerebellar hemispherectomy upon them in monkeys. *Brain Research* **697**, 143–151.

GENTILUCCI, M. & CORBALLIS, M. C. (2006). From manual gesture to speech: a gradual transition. *Neuroscience & Biobehavioral Reviews* **30,** 949–960.
GENTY, E., BREUER, T., HOBAITER, C. & BYRNE, R. W. (2009). Gestural communication of the gorilla (*Gorilla gorilla*): repertoire, intentionality and possible origins. *Animal Cognition* **12**, 527–546.
GENTY, E., CLAY, Z., HOBAITER, C. & ZUBERBÜHLER, K. (2014). Multi–modal use of a socially directed call in bonobos. *PLoS ONE* **9**, e84738.
GENTY, E. & ZUBERBÜHLER, K. (2014). Spatial reference in a bonobo gesture. *Current Biology* **24**, 1601–1605.
GHAZANFAR, A. A. (2013). Multisensory vocal communication in primates and the evolution of rhythmic speech. *Behavioral Ecology and Sociobiology* **67**, 1441–1448.
GHAZANFAR, A. A. & RENDALL, D. (2008). *Evolution of Human Vocal Production*. Cell Press, Cambridge, MA, ETATS–UNIS.
GILLESPIE–LYNCH, K., GREENFIELD, P. M., LYN, H. & SAVAGE–RUMBAUGH, S. (2014). Gestural and symbolic development among apes and humans: support for a multimodal theory of language evolution. *Frontiers in Psychology* **5**, 1228.
GOLDIN–MEADOW, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences* **3**, 419–429.
GOLDIN–MEADOW, S. & ALIBALI, M. W. (2013). Gesture's role in speaking, learning, and creating language. *Annual review of psychology* **64**, 257–283.
GRAHAM, K. E., FURUICHI, T. & BYRNE, R. W. (2016). The gestural repertoire of the wild bonobo (*Pan paniscus*): a mutually understood communication system. *Animal Cognition* **20**, 171–177.
GRICE, H. (1975). Logic and conversation. In *Syntax and Semantics*, vol. Speech Acts (ed. P. Cole and J. Morgan), pp. 43–58. Academic Press, New York.
GRICE, H. P. (1957). Meaning. *The Philosophical Review* **66**, 377–388.
GRUBER, T. & GRANDJEAN, D. (2017). A comparative neurological approach to emotional expressions in primate vocalizations. *Neuroscience & Biobehavioral Reviews* **73**, 182–190.
GRUBER, T. & ZUBERBÜHLER, K. (2013). Vocal recruitment for joint travel in wild chimpanzees. *PLoS ONE* **8**, e76073.
GYGER, M., KARAKASHIAN, S. J. & MARLER, P. (1986). Avian alarm calling: is there an audience effect? *Animal Behaviour* **34**, 1570–1572.
HAGE, S. R. & NIEDER, A. (2013). Single neurons in monkey prefrontal cortex encode volitional initiation of vocalizations. *Nature communications* **4**, 2409.
HAGE, S. R. & NIEDER, A. (2016). Dual neural network model for the evolution of speech and language. *Trends in neurosciences* **39**, 813–829.
HALINA, M., ROSSANO, F. & TOMASELLO, M. (2013). The ontogenetic ritualization of bonobo gestures. *Animal Cognition 16*, 653–666.
HALL, M. L. (2009). Chapter 3 A Review of Vocal Duetting in Birds. In *Advances in the Study of Behavior*, vol. Volume 40, pp. 67–121. Academic Press.
HASSELMO, M. E., ROLLS, E. T. & BAYLIS, G. C. (1989). The role of expression and identity in the face–selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural Brain Research* **32**, 203–218.
HAUSER, M. D., CHOMSKY, N. & FITCH, T. W. (2002). The language faculty: what is it, who has it, and how did it evolve? *Science* **298**, 1568–1579.
HEBETS, E. A. & PAPAJ, D. R. (2005). Complex signal function: developing a framework of testable hypotheses. *Behavioural Ecology and Sociobiology* **57**, 197–214.
HENRY, L., CRAIG, A. J., LEMASSON, A. & HAUSBERGER, M. (2015). Social coordination in animal vocal interactions. Is there any evidence of turn–taking? The starling as an animal model. *Frontiers in Psychology* **6**, 1416.

HEWES, G. W. (1973). Primate communication and the gestural origin of language. *Current Anthropology* **12**, 5–24.
HICKOK, G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience* **21**, 1229–1243.
HIGHAM, J. P. & HEBETS, E. A. (2013). An introduction to multimodal communication. *Behavioral Ecology and Sociobiology* **67**, 1381–1388.
HINDLEY, P. C. (2014). Nominal and imperative iconic gestures used by the Khoisan of North West Botswana to coordinate hunting. *African Study Monographs* **35**, 149–181.
HOBAITER, C. & BYRNE, R. W. (2011a). The gestural repertoire of the wild chimpanzee. *Animal Cognition* **14**, 747–767.
HOBAITER, C. & BYRNE, R. W. (2011b). Serial gesturing by wild chimpanzees: its nature and function for communication. *Animal Cognition* **14**, 827–838.
HOBAITER, C. & BYRNE, R. W. (2012). Gesture use in consortship: wild chimpanzees' use of gesture for an 'evolutionarily urgent'purpose. In *Developments in Primate Gesture Research* (ed. S. Pika and K. Liebal), pp. 129–146. John Benjamins Publishing Company.
HOBAITER, C. & BYRNE, R. W. (2014). The meanings of chimpanzee gestures. *Current Biology* **24**, 1596–1600.
HOBAITER, C., BYRNE, R. W. & ZUBERBÜHLER, K. (2017). Wild chimpanzees' use of single and combined vocal and gestural signals. *Behavioral Ecology and Sociobiology* **71**, 96.
HOBAITER, C., LEAVENS, D. A. & BYRNE, R. W. (2014). Deictic gesturing in wild chimpanzees (*Pan troglodytes*)? Some possible cases. *Journal of Comparative Psychology* **128**, 82–87.
HOCKETT, C. F. (1960). The origin of speech. *Scientific American* **203**, 88–97.
HOLLE, H. & GUNTER, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience* **19**, 1175–1192.
HOMAE, F., HASHIMOTO, R., NAKAJIMA, K., MIYASHITA, Y. & SAKAI, K. L. (2002). From perception to sentence comprehension: the convergence of auditory and visual information of language in the left inferior frontal cortex. *Neuroimage* **16**, 883–900.
HOPKINS, W. D., TAGLIALATELA, J. D. & LEAVENS, D. (2007). Chimpanzees differentially produce novel vocalizations to capture the attention of a human. *Animal Behaviour* **73**, 281–286.
HURFORD, J. R. (2007). *The Origins of Meaning: Language in the Light of Evolution*. Oxford University Press, Oxford.
IMAI, M. & KITA, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**, 20130298.
JÜRGENS, U. (1976). Projections from the cortical larynx area in the squirrel monkey. *Experimental Brain Research* **25**, 401–411.
JÜRGENS, U. (2002). Neural pathways underlying vocal control. *Neuroscience & Biobehavioral Reviews* **26**, 235–258.
JUVRUD, J., BAKKER, M., KADUK, K., DEVALK, J. M., GREDEBÄCK, G. & KENWARD, B. (2018). Longitudinal continuity in understanding and production of giving-related behavior from infancy to childhood. *Child Development* **90**, e182-e191.
KATSU, N., YAMADA, K. & NAKAMICHI, M. (2017). Influence of social interactions with nonmother females on the development of call usage in Japanese macaques. *Animal Behaviour* **123**, 267–276.
KAYE, L. K., MALONE, S. A. & WALL, H. J. (2017). Emojis: insights, affordances, and possibilities for psychological science. *Trends in Cognitive Sciences* **21**, 66–68.
KENDON, A. (1980). Gesticulation and speech: two aspects of the process of utterance. In *The Relationship of Verbal and Nonverbal Communication* (ed. M. Key), pp. 207–227. The Hague, Mouton.

KENDON, A. (2000). Language and gesture: unity or duality. In *Language and Gesture* (ed. D. McNeill), pp. 47–63. Cambridge University Press, Cambridge.

KENDON, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press, Cambridge.

KENDON, A. (2017). Reflections on the "gesture–first" hypothesis of language origins. *Psychonomic Bulletin & Review* **24**, 163–170.

KERSKEN, V., GÓMEZ, J.–C., LISZKOWSKI, U., SOLDATI, A. & HOBAITER, C. (2018). A gestural repertoire of 1– to 2–year–old human children: in search of the ape gestures. *Animal Cognition*. https://doi.org/10.1007/s10071-018-1213-z.

KIMURA, D. (1993). *Neuromotor mechanisms in human communication*. Oxford University Press, Oxford.

KLIMA, E. & BELLUGI, U. (1979). *The Signs of Language*. Harvard University Press., Cambridge, MA.

KRAUSE, J., LALUEZA–FOX, C., ORLANDO, L., ENARD, W., GREEN, R. E., BURBANO, H. A., HUBLIN, J. J., HANNI, C., FORTEA, J., DE LA RASILLA, M., BERTRANPETIT, J., ROSAS, A. & PÄÄBO, S. (2007). The derived FOXP2 variant of modern humans was shared with neanderthals. *Current Biology* **17**, 1908–1912.

KUHL, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience* **5**, 831.

KUMAR, V., CROXSON, P. L. & SIMONYAN, K. (2016). Structural organization of the laryngeal motor cortical network and its implication for evolution of speech production. *Journal of Neuroscience* **36**, 4170–4181.

KUYPERS, H. (1958). Some projections from the pericentral cortex to the pons and lower brain stem in monkey and chimpanzee. *Journal of Comparative Neurology* **110**, 221–255.

LAMEIRA, A. R., HARDUS, M. & WICH, S. (2012). Orangutan instrumental gesture–calls: reconciling acoustic and gestural speech evolution models. *Evolutionary Biology* **39**, 415–418.

LAMEIRA, A. R. (2017). Bidding evidence for primate vocal learning and the cultural substrates for speech evolution. *Neuroscience & Biobehavioral Reviews* **83**, 429–439.

LAMEIRA, A. R., HARDUS, M. E., MIELKE, A., WICH, S. A. & SHUMAKER, R. W. (2016). Vocal fold control beyond the species–specific repertoire in an orang–utan. *Scientific Reports* **6**, 30315.

LAPORTE, M. N. C. & ZUBERBÜHLER, K. (2011). The development of a greeting signal in wild chimpanzees. *Developmental Science* **14**, 1220–1234.

LEAVENS, D. A. (2004). Manual deixis in apes and humans. *Interaction Studies* **5,** 387–408.

LEAVENS, D. A., BARD, K. A. & HOPKINS, W. D. (2017). The mismeasure of ape social cognition. *Animal Cognition***.** https://doi.org/10.1007/s10071-017-1119-1

LEAVENS, D. A., HOPKINS, W. D. & BARD, K. A. (2005a). Understanding the point of chimpanzee pointing: epigenesis and ecological validity. *Current Directions in Psychological Science* **14**, 185–189.

LEAVENS, D. A. & RACINE, T. P. (2009). Joint attention in apes and humans: are humans unique? . *Journal of Consciousness Studies* **16**, 240–267.

LEAVENS, D. A., RUSSELL, J. L. & HOPKINS, W. D. (2005b). Intentionality as measured in the persistence and elaboration of communication by chimpanzees (*Pan troglodytes*). *Child Development* **76**, 291–306.

LEVINSON, S. C. (2006). On the human "interaction engine". In *Roots of Human Sociality: Culture, Cognition and Interaction* (ed. N. J. Enfield and S. C. Levinson), pp. 39–69. Berg, Oxford.

LEVINSON, S. C. (2016). Turn–taking in human communication–origins and implications for language processing. *Trends in Cognitive Sciences* **20**, 6–14.

LEVINSON, S. C. & HOLLER, J. (2014). The origin of human multi–modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**, 20130302.

LIEBAL, K., CALL, J. & TOMASELLO, M. (2004a). Use of gesture sequences in chimpanzees. *American Journal of Primatology* **64**, 377–396.

LIEBAL, K. & OÑA, L. (2018). Mind the gap – moving beyond the dichotomy between intentional gestures and emotional facial and vocal signals of nonhuman primates. *Interaction Studies* **19**, 121–135.

LIEBAL, K., PIKA, S., CALL, J. & TOMASELLO, M. (2004b). To move or not to move. How apes adjust to the attentional state of others. *Interaction Studies* **5**, 199–219.

LIEBAL, K., SCHNEIDER, C. & ERRSON–LEMBECK, M. (2018). How primates acquire their gestures: evaluating current theories and evidence. *Animal Cognition*. https://doi.org/10.1007/s10071-018-1187-x

LIEBAL, K., WALLER, B. M., BURROWS, A. M. & SLOCOMBE, K. E. (2013). *Primate Communication: A Multimodal Approach*. Cambridge University Press, Cambridge.

LIEBERMAN, P. (1993). *Uniquely Human: The Evolution of Speech, Thought, and Selfless Behavior*. Harvard University Press, Cambridge, MA.

LO, S. K. (2008). The nonverbal communication functions of emoticons in computer–mediated communication. *CyberPsychology & Behavior* **11**, 595–597.

LOCKWOOD, G. & DINGEMANSE, M. (2015). Iconicity in the lab: a review of behavioral, developmental, and neuroimaging research into sound–symbolism. *Frontiers in Psychology* **6**, 1246.

LOH, K. K., PETRIDES, M., HOPKINS, W. D., PROCYK, E. & AMIEZ, C. (2017). Cognitive control of vocalizations in the primate ventrolateral–dorsomedial frontal (VLF–DMF) brain network. *Neuroscience & Biobehavioral Reviews* **82**, 32–44.

LYN, H. RUSSELL, J. L. & HOPKINS, W. D. (2010). The impact of environment on the comprehension of declarative communication in apes. *Psychological Science* **21**, 360–365.

MACEDONIA, J. M. & EVANS, C. S. (1993). Essay on contemporary issues in ethology: variation among mammalian alarm call systems and the problem of meaning in animal signals. *Ethology* **93**, 177–197.

MARLER, P., EVANS, C. S. & HAUSER, M. D. (1992). Animal signals: motivational, referential or both? In *Nonverbal Vocal Communication: Comparative And Developmental Approaches* (ed. H. Papousek, S. Jürgens and M. Papousek), pp. 66–86. Cambridge University Press, Cambridge.

MASSARO, D. W. (1998). *Perceiving Talking Faces: From Speech Perception To A Behavioral Principle*. MIT Press, Cambridge, MA.

MASSARO, D. W. & EGAN, P. B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review* **3**, 215–221.

MCGURK, H. & MACDONALD, J. (1976). Hearing lips and seeing voices. *Nature* **264**, 746–748.

MCNEILL, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press, Chicago.

MCNEILL, D. (2000). *Language and Gesture*. Cambridge University Press, Cambridge.

MCNEILL D., CASSELL, J. & MCCULLOUGH, K. (1994). Communicative effects of speech–mismatched gestures. *Research on Language and Social Interaction* **27**, 223–237.

MEGUERDITCHIAN, A., GARDNER, M. J., SCHAPIRO, S. J. & HOPKINS, W. D. (2012). The sound of one–hand clapping: handedness and perisylvian neural correlates of a communicative gesture in chimpanzees. *Proceedings of the Royal Society B: Biological Sciences* **279**, 1959–1966.

MICHELETTA, J., ENGELHARDT, A., MATTHEWS, L. E. E., AGIL, M. & WALLER, B. M. (2013). Multicomponent and multimodal lipsmacking in crested macaques (*Macaca nigra*). *American Journal of Primatology* **75**, 763–773.

MOORE, R. (2013). Evidence and interpretation in great ape gestural communication. *Humana.Mente – Journal of Philosophical Studies* **24**, 27–51.

MOORE, R. (2016). Meaning and ostension in great ape gestural communication. *Animal Cognition* **19**, 223–231.

MOORE, R. (2017). Social cognition, Stag Hunts, and the evolution of language. *Biology & Philosophy* **32**, 797–818.

MÜHLENBERND, R., ENKE, D., VILLING, M., GAVRILOV, N. & NICK, J. D. (2014). Modality switch in human language evolution. In *Evolution of Language: Proceedings of the 10th International Conference* (ed. E. A. Cartmill, Sean Roberts, H. Lyn and H. Cornish*)*, pp. 161–168. World Scientific.

MYERS–SCHULZ, B., PUJARA, M., WOLF, R. C. & KOENIGS, M. (2013). Inherent emotional quality of human speech sounds. *Cognition & Emotion* **27**, 1105–1113.

OUATTARA, K., LEMASSON, A. & ZUBERBÜHLER, K. (2009). Campbell's monkeys use affixation to alter call meaning. *PLoS ONE* **4**, e7808.

OWREN, M. J. & RENDALL, D. (2001). Sound on the rebound: bringing form and function back to the forefront in understanding nonhuman primate vocal signalling. *Evolutionary Anthropology* **10**, 58–71.

PARTAN, S. R. (1999). Multimodal communication: the integration of visual and acoustic signals by macaques. *Behaviour* **139**, 993–1028.

PARTAN, S. R. (2002). Single and multichannel signal composition: facial expressions and vocalizations of rhesus macaques (*Macaca mulatta*). *Behaviour* **139**, 993–1027.

PARTAN, S. R. (2013). Ten unanswered questions in multimodal communication. *Behavioral Ecology and Sociobiology* **67**, 1523–1539.

PARTAN, S. R. & MARLER, P. (1999). Communication goes multimodal. *Science* **283**, 1272–1273.

PARTAN, S. R. & MARLER, P. (2005). Issues in the classification of multimodal communication signals. *American Naturalist* **166**, 231–245.

PATTERSON, F. G. (1978). The gestures of a gorilla: language acquisition in another pongid. *Brain and Language* **5**, 72–97.

PERLMAN, M. (2017). Debunking two myths against vocal origins of language: language is iconic and multimodal to the core. *Interaction Studies* **18**, 376–401.

PERLMAN, M. & CAIN, A. A. (2014). Iconicity in vocalization, comparisons with gesture, and implications for theories on the evolution of language. *Gesture* **14**, 320–350.

PERLMAN, M., TANNER, J. E. & KING, B. J. (2012). A mother gorilla's variable use of touch to guide her infant. In *Developments in Primate Gesture Research* (ed. S. Pika and K. Liebal), pp. 55–71. John Benjamins Publishing Company.

PERNISS, P., THOMPSON, R. L. & VIGLIOCCO, G. (2010). Iconicity as a general property of language: evidence from spoken and signed languages. *Frontiers in Psychology* **1**, 227.

PERNISS, P. & VIGLIOCCO, G. (2014). The bridge of iconicity: from a world of experience to the experience of language. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**, 20130300.

PIKA, S. (2008). Gestures of apes and pre–linguistic human children: similar or different? *First Language* **28**, 116–140.

PIKA, S. & BUGNYAR, T. (2011). The use of referential gestures in ravens (*Corvus corax*) in the wild. *Nature Communications* **2**, 560.

PIKA, S. & FRÖHLICH, M. (2018). Gestural acquisition in great apes: the Social Negotiation Hypothesis. *Animal Cognition*. https://doi.org/10.1007/s10071-017-1159-6

PIKA, S. & MITANI, J. (2006). Referential gestural communication in wild chimpanzees (*Pan troglodytes*). *Current Biology* **16**, R191–R192.

PIKA, S., WILKINSON, R., KENDRICK, K. H. & VERNES, S. C. (2018). Taking turns: bridging the gap between human and animal communication. *Proceedings of the Royal Society B: Biological Sciences* **285**, 20180598.

POLLICK, A. S. & DE WAAL, F. B. M. (2007). Ape gestures and language evolution. *Proceedings of the National Academy of Sciences* **104**, 8184–8189.

PORTER, S. & YUILLE, J. C. (1996). The language of deceit: an investigation of the verbal clues to deception in the interrogation context. *Law and Human Behavior* **20**, 443–458.
PREMACK, A. J. & PREMACK, D. (1972). Teaching language to an ape. *Scientific American* **227**, 92–99.
RAUSCHECKER, J. P. & SCOTT, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature Neuroscience* **12**, 718.
RIZZOLATTI, G. & ARBIB, M. A. (1998). Language within our grasp. *Trends in Neurosciences* **21**, 188–194.
RIZZOLATTI, G., FOGASSI, L. & GALLESE, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience* **2**, 661.
ROBERTS, S. G., TORREIRA, F. & LEVINSON, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: a corpus study. *Frontiers in Psychology* **6**, 509.
ROSSANO, F. (2013). Sequence organization and timing of bonobo mother–infant interactions. *Interaction Studies* **14**, 160–189.
ROSSANO, F. (2018). Social manipulation, turn–taking and cooperation in apes. *Interaction Studies* **19**, 151–166.
ROSSANO, F. & LIEBAL, K. (2014). "Requests" and "offers" in orangutans and human infants. In *Requesting in Social Interaction* (ed. P. Drew and E. Couper-Kuhlen)**,** pp. 333–362. John Benjamins Publishing.
ROWE, C. (1999). Receiver psychology and the evolution of multicomponent signals. *Animal Behaviour* **58**, 921–931.
RUBEN, R. J. (1997). A time frame of critical/sensitive periods of language development. *Acta Oto–Laryngologica* **117**, 202–205.
SACKS, H., SCHEGLOFF, E. A. & JEFFERSON, G. (1974). A simplest systematics for the organization of turn–taking in conversation. *Language* **50**, 696–735.
SCALAIDHE, S. P. Ó., WILSON, F. A. & GOLDMAN–RAKIC, P. S. (1997). Areal segregation of face–processing neurons in prefrontal cortex. *Science* **278**, 1135–1138.
SCHEL, A. M., TOWNSEND, S. W., MACHANDA, Z., ZUBERBÜHLER, K. & SLOCOMBE, K. E. (2013). Chimpanzee alarm call production meets key criteria for intentionality. *PLoS ONE* **8**, e76674.
SCHEL, A. M., TRANQUILLI, S. & ZUBERBÜHLER, K. (2009). The alarm call system of two species of black–and–white colobus monkeys (*Colobus polykomos* and *Colobus guereza*). *Journal of Comparative Psychology* **123**, 136.
SCHERER, K. R., JOHNSTONE, T. & KLASMEYER, G. (2003). Vocal expression of emotion. In *Handbook of Affective Sciences* (ed. R. J. Davidson, K. R. Scherer and H. H. Goldsmith*),* pp. 433–456. Oxford University Press, New York.
SCHMIDTKE, D., CONRAD, M. & JACOBS, A. (2014). Phonological iconicity. *Frontiers in Psychology* **5**, 80.
SEYFARTH, R. M., CHENEY, D. L., BERGMAN, T., FISCHER, J., ZUBERBÜHLER, K. & HAMMERSCHMIDT, K. (2010). The central importance of information in studies of animal communication. *Animal Behaviour* **80**, 3–8.
SHERMAN, P. W. (1977). Nepotism and the evolution of alarm calls. *Science* **197**, 1246–1253.
SIEVERS, C. & GRUBER, T. (2016). Reference in human and non–human primate communication: what does it take to refer? *Animal Cognition* **19**, 759–768.
SIEVERS, C., WILD, M. & GRUBER, T. (2017). Intentionality and flexibility in animal communication. In *The Handbook of the Philosophy of Animal Minds* (ed. K. Andrews and J. Beck), pp. 333–342. Routledge.
SLATER, P. J. & MANN, N. I. (2004). Why do the females of many bird species sing in the tropics? *Journal of Avian Biology* **35**, 289–294.
SLOCOMBE, K. E., WALLER, B. M. & LIEBAL, K. (2011). The language void: the need for multimodality in primate communication research. *Animal Behaviour* **81**, 919–924.

SNOWDON, C. T. & HAUSBERGER, M. (1997). *Social Influences on Vocal Development*. Cambridge University Press, Cambridge.

SPERBER, D. & WILSON, D. (1986). *Relevance: Communication and Cognition*. Harvard University Press, Cambridge, MA.

SPILLMANN, B., DUNKEL, L. P., VAN NOORDWIJK, M. A., AMDA, R. N., LAMEIRA, A. R., WICH, S. A. & VAN SCHAIK, C. P. (2010). Acoustic properties of long calls given by flanged male orang-utans (*Pongo pygmaeus wurmbii*) reflect both individual identity and context. *Ethology* **116**, 385–395.

STIVERS, T., ENFIELD, N. J., BROWN, P., ENGLERT, C., HAYASHI, M. & HEINEMANN, T. (2009). Universals and cultural variation in turn–taking in conversation. *Proceedings of the National Academy of Sciences* **106**, 10587–10592.

SULLIVAN, K. (1985). Selective alarm calling by downy woodpeckers in mixed–species flocks. *The Auk* **102**, 184–187.

SUZUKI, T. N., WHEATCROFT, D. & GRIESSER, M. (2016). Experimental evidence for compositional syntax in bird calls. *Nature Communications* **7**, 10986.

TAGLIALATELA, J. P., REAMER, L., SCHAPIRO, S. J. & HOPKINS, W. D. (2012). Social learning of a communicative signal in captive chimpanzees. *Biology Letters* **8**, 498–501.

TAGLIALATELA, J. P., RUSSELL, J. L., POPE, S. M., MORTON, T., BOGART, S., REAMER, L. A., SCHAPIRO, S. J. & HOPKINS, W. D. (2015). Multimodal communication in chimpanzees. *American Journal of Primatology* **77**, 1143–1148.

TAGLIALATELA, J. P., RUSSELL, J. L., SCHAEFFER, J. A. & HOPKINS, W. D. (2011). Chimpanzee vocal signaling points to a multimodal origin of human language. *PLoS ONE* **6**, e18852.

TAKAHASHI, DANIEL Y., NARAYANAN, DARSHANA Z. & GHAZANFAR, ASIF A. (2013). Coupled oscillator dynamics of vocal turn–taking in monkeys. *Current Biology* **23**, 2162–2168.

TANNER, J. E. & BYRNE, R. (1996). Representation of action through iconic gesture in a captive lowland gorilla. *Current Anthropology* **37**, 162–173.

TOBIAS, M. L., VISWANATHAN, S. S. & KELLEY, D. B. (1998). Rapping, a female receptive call, initiates male–female duets in the South African clawed frog. *Proceedings of the National Academy of Sciences* **95**, 1870–1875.

TOMASELLO, M. (1995). Joint attention as social cognition. In *Joint Attention: Its Origin and Role in Development* (ed. C. Moore and P. J. Dunham), pp. 103–130. Erlbaum.

TOMASELLO, M. (2006). Why don't apes point? In *Roots of Human Sociality: Culture, Cognition and Interaction* (ed. N. J. Enfield and S. C. Levinson), pp. 506–524. Berg, Oxford.

TOMASELLO, M. (2008). *Origins of human communication*. MIT press, Cambridge, Massachusetts.

TOMASELLO, M., CALL, J., NAGELL, K., OLGUIN, R. & CARPENTER, M. (1994). The learning and use of gestural signals by young chimpanzees: a trans–generational study. *Primates* **35**, 137–154.

TOMASELLO, M. & CARPENTER, M. (2007). Shared intentionality. *Developmental Science* **10**, 121–125.

TOMASELLO, M., CARPENTER, M., CALL, J., BEHNE, T. & MOLL, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *Behavioral and Brain Sciences* **28**, 675–691.

TOMASELLO, M., GEORGE, B. L., KRUGER, A. C., JEFFREY, M., FARRAR & EVANS, A. (1985). The development of gestural communication in young chimpanzees. *Journal of Human Evolution* **14**, 175–186.

TOMASELLO, M. & ZUBERBÜHLER, K. (2002). Primate vocal and gestural communication. In *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition* (ed. M. Bekoff, C. Allen and G. M. Burghardt), pp. 293–299. MIT Press, Cambridge.

TOWNSEND, S. W., ENGESSER, S., STOLL, S., ZUBERBÜHLER, K. & BICKEL, B. (2018). Compositionality in animals and humans. *PLoS Biology* **16**, e2006425.

TOWNSEND, S. W., KOSKI, S. E., BYRNE, R. W., SLOCOMBE, K. E., BICKEL, B., BOECKLE, M., BRAGA GONCALVES, I., BURKART, J. M., FLOWER, T., GAUNET, F., GLOCK, H. J., GRUBER, T., JANSEN, D. A. W. A. M., LIEBAL, K., LINKE, A., MIKLÓSI, Á., MOORE, R., VAN SCHAIK, C. P., STOLL, S., VAIL, A., WALLER, B. M., WILD, M., ZUBERBÜHLER, K. & MANSER, M. B. (2017). Exorcising Grice's ghost: an empirical approach to studying intentional communication in animals. *Biological Reviews* **92**, 1427–1433.

TSAO, D. Y., SCHWEERS, N., MOELLER, S. & FREIWALD, W. A. (2008). Patches of face–selective cortex in the macaque frontal lobe. *Nature Neuroscience* **11**, 877.

TYLOR, E. B. (1866). On the origin of language. *Fortnightly review* **4**, 544–559.

UNGERLEIDER, L. & MISHKIN, M. (1982). Two cortical visual systems. In *Analysis of Visual Behavior* (ed. D. J. Ingle, M. A. Goodale and R. J. W. Mansfield), pp. 549–586. MIT Press, Cambridge, MA.

VAIL, A. L., MANICA, A. & BSHARY, R. (2013). Referential gestures in fish collaborative hunting. *Nature Communications* **4**, 1765.

VAN SCHAIK, C. P. (2016). *The Primate Origins of Human Nature*. John Wiley & Sons.

VAN SCHAIK, C. P., VAN NOORDWIJK, M. A. & WICH, S. A. (2006). Innovation in wild Bornean orangutans (*Pongo pygmaeus wurmbii*). *Behaviour* **143**, 839–876.

VEÀ, J. & SABATER–PI, J. (1998). Spontaneous pointing behaviour in the wild pygmy chimpanzee (*Pan paniscus*). *Folia Primatologica* **69**, 289–290.

VERNES, S. C. (2017). What bats have to say about speech and language. *Psychonomic Bulletin & Review* **24**, 111–117.

VIGLIOCCO, G., PERNISS, P. & VINSON, D. (2014). Language as a multimodal phenomenon: implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**, 20130292.

WACEWICZ, S. & ZYWICZYNSKI, P. (2017). The multimodal origins of linguistic communication. *Language & Communication* **54**, 1–8.

WALLER, B. M., LIEBAL, K., BURROWS, A. M. & SLOCOMBE, K. E. (2013). How can a multimodal approach to primate communication help us understand the evolution of communication? *Evolutionary Psychology* **11**, 538–549.

WATSON, S. K., TOWNSEND, S. W., SCHEL, A. M., WILKE, C., WALLACE, E. K., CHENG, L., WEST, V. & SLOCOMBE, K. E. (2015). Vocal learning in the functionally referential food grunts of chimpanzees. *Current Biology* **25**, 495–499.

WHEELER, B. C. & FISCHER, J. (2012). Functionally referential signals: a promising paradigm whose time has passed. *Evolutionary Anthropology: Issues, News, and Reviews* **21**, 195–205.

WICH, S. A. & DE VRIES, H. (2006). Male monkeys remember which group members have given alarm calls. *Proceedings of the Royal Society B: Biological Sciences* **273**, 735–740.

WICH, S. A., SWARTZ, K. B., HARDUS, M. E., LAMEIRA, A. R., STROMBERG, E. & SHUMAKER, R. W. (2009). A case of spontaneous acquisition of a human sound by an orangutan. *Primates* **50**, 56–64.

WILKE, C., KAVANAGH, E., DONNELLAN, E., WALLER, B. M., MACHANDA, Z. P. & SLOCOMBE, K. E. (2017). Production of and responses to unimodal and multimodal signals in wild chimpanzees, *Pan troglodytes schweinfurthii*. *Animal Behaviour* **123**, 305–316.

WILLEMS, R. M. & HAGOORT, P. (2007). Neural evidence for the interplay between language, gesture, and action: a review. *Brain and Language* **101**, 278–289.

WONG, B., COWLING, A., CUNNINGHAM, R. B. & DONNELLY, C. F. (2004). Do temperature and social environment interact to affect call rate in frogs (*Crinia signifera*)? *Austral Ecology* **29,** 209–214.

XU, J., GANNON, P. J., EMMOREY, K., SMITH, J. F. & BRAUN, A. R. (2009). Symbolic gestures and spoken language are processed by a common neural system. *Proceedings of the National Academy of Sciences* **106**, 20664–20669.

ZLATEV, J. (2008). From proto–mimesis to language: evidence from primatology and social neuroscience. *Journal of Physiology* **102**, 137–151.

ZUBERBÜHLER, K. (2000a). Causal cognition in a non–human primate: field playback experiments with Diana monkeys. *Cognition* **76**, 195–207.

ZUBERBÜHLER, K. (2000b). Interspecies semantic communication in two forest primates. *Proceedings of the Royal Society B: Biological Sciences* **267**, 713–718.

ZUBERBÜHLER, K. (2003). Referential signalling in non–human primates: cognitive precursors and limitations for the evolution of language. *Advances in the Study of Behavior* **33**, 265–307.

ZUBERBÜHLER, K. (2005). The phylogenetic roots of language: evidence from primate communication and cognition. *Current Directions in Psychological Science* **14**, 126–130.

ZUBERBÜHLER, K. (2018). Combinatorial capacities in primates. *Current Opinion in Behavioral Sciences* **21**, 161–169.

ZUBERBÜHLER, K., CHENEY, D. L. & SEYFARTH, R. M. (1999). Conceptual semantics in a nonhuman primate. *Journal of Comparative Psychology* **113**, 33–42.

ZUCKERMAN, M., DEPAULO, B. M. & ROSENTHAL, R. (1981). Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology* **14**, 1–59.

ŻYWICZYŃSKI, P., WACEWICZ, S. & SIBIERSKA, M. (2018). Defining pantomime for language evolution research. *Topoi* **37**, 307–318.

Table 1. Face-to-face dyadic communication in great apes and humans.

Aspect of communication	**Great apes**	**Humans**
Dominant stream	Gesture	Vocalization
Support stream	Vocalization	Gesture
Largest repertoire	Gesture	Vocalization
Context dependency of signal meaning	High	Moderate
Joint attention	Unclear	Present
Major signalling need	Achieve social goal: Refinement	Cooperative information exchange: compositionality

Table 2. Evidence for language components in non-human gesture and vocalization.

Cognitive feature	**Gesture**	**Vocalization**
Intentionality	+++	++
Reference	+	++
Iconicity	+++	?
Combinatoriality	-	++
Turn-taking	++	++
Neural control	Highly overlapping	
Ontogenetic plasticity	Similar	

Figure legends

Fig. 1. Transition in face-to-face communication from the earliest hominins, proxied by great apes, to humans.

Fig. 2. The evolutionary trajectory of communicative abilities leading to modern human communication. (A) Scenario hypothesized by Levinson & Holler (2014), emphasizing the multimodal 'interaction engine'. (B) Our hypothesized scenario, emphasizing the onset of diverse origins of human gesture and prosociality. mya, million years ago.