Language & Communication 54 (2017) 73–81

Contents lists available at ScienceDirect

Language & Communication

journal homepage: www.elsevier.com/locate/langcom

Bodily constraints contributing to multimodal referentiality in humans: The contribution of a de-pigmented sclera to proto-declaratives

Juan Olvido Perea García*, Katrine Rosendal Ehlers, Kristian Tylén

Aarhus Universitet, Nordre Ringgade 1, 8000 Aarhus C, Denmark

ARTICLE INFO

Article history:
Available online 26 November 2016

Keywords:
De-pigmented sclera
Multimodality
Referentiality
Cooperation
Eye

ABSTRACT

We present the results of an empirical study that measured the contribution of a conspicuous eye-gaze (as a function of scleral de-pigmentation) of humans in conveying multimodal referentiality by combining visual and auditory cues in a naturalistic setting. We made participants interact in a cooperative task in which they had to convey referential meaning about co-presential entities. In one of the conditions, participants had no access to their interactants' eye-gaze. We interpret the results as supporting the idea that our eye morphology contributes to instantiating multimodal referentiality in cooperative tasks in peripersonal space.

1. Introduction

Purely verbal discourse without accompanying extralinguistic cues is a relatively recent cultural invention associated with the rise and expansion of writing systems, as well as a rather infrequent phenomenon in contexts of social interaction (Linell, 1982). In most of human interaction, linguistic cues typically co-occur with a rich variety of signals of a different nature, such as body orientation, facial expressions, eye-gaze, and hand gestures. This has led researchers from a range of fields to describe linguistic interaction as an inherently multimodal activity (Kendon, 2011), in which linguistic cues and other semiotic resources interact and co-evolve – on both phylogenetic and ontogenetic timescales – ratcheting on each other as they entwine in ever more complex patterns of expressive behavior.

Bearing on the assumption that gestural communication paved the way for vocalizations to convey referential meaning in our evolutionary path, most efforts have focused on addressing the interaction between hand and bodily gestures and vocalizations (Corballis, 2002). However, the link between gaze-cues and vocalizations remains largely unexplored, even though it has been considered pivotal to the ontogenetic development of linguistic abilities (Dunham et al., 1993; Mundy et al., 2007; Tomasello and Farrar, 1986; Tomasello, 2009; Baldwin, 1993). To what degree do humans actually rely on gaze-cues to jointly establish shared reference to entities in the immediate environment? As a first step towards a satisfactory answer to this question, our study sets out to test the assumption that linguistic and gaze-cues constrain each other in interaction. In this article, we present an experiment in which we measured the distinctive contribution of gaze-cues to the instantiation of referentiality. In a controlled experimental environment, we simulated the physical and interactive affordances of a naturalistic setting in which our hominin ancestors would have relied on gaze-cues as an efficient pointer – activities such as

* Corresponding author.
E-mail addresses: juan.olvido@gmail.com (J.O. Perea García), karo@cc.au.dk (K.R. Ehlers), kristian@dac.au.dk (K. Tylén).

http://dx.doi.org/10.1016/j.langcom.2016.10.007
0271-5309/

cooking, tool-making, and hunting – that is, cooperative tasks in peripersonal space. We focus especially on the affordances of the eye to indicate a gaze direction for someone to follow. This is because, among many other hypothesized functions (such as aiding in recognizing emotions: Poggi et al., 2009; or aiding in assessing the health state of potential sexual partners; Tomasello et al., 2007) the morphological properties of a de-pigmented sclera also afford using the eye-ball as a reliable spatial pointer. In co-occurrence with vocalizations, gaze allows interactants to efficiently coordinate their attention towards the relevant entities in cooperative interactions in peripersonal space. We propose that this referential function of eye-gaze, when orchestrated with vocalizations, could have been essential for our ancestors to coordinate their activity in complex tasks that either required the use of hands to manipulate objects (e.g. tool-making or cooking), or in which acoustic cues would be disfavored (e.g. hunting). Under these circumstances, gaze-cues become the only reliable visual pointer. We speculate that a general orientation towards these kinds of cooperative activities might have motivated a selective pressure for the particular characteristics of the ocular morphology of modern hominins, not dissimilar from those of other extant great apes, like Western lowland gorillas (Mayhew, 2013), or Sumatran orangutans (Perea García, 2016). This evolutionary trend in developing a conspicuous eye morphology is most undoubtedly most visible in modern humans, which led previous research to conclude that it was uniquely human (Kobayashi and Koshima, 1997, 2001).

1.1. Language as a multimodal activity

The ability to communicate linguistically is indisputably a uniquely human trait. Even though Western scholarship has entertained itself on the topic of language for centuries, it has been concerned mostly with its written form (Linell, 1982). This has led to a conception of language as an object-like *system*, rather than a social coordination process or *activity* (Love, 2004). This in spite of the fact that face-to-face conversation remains the most pervading communicative practice among humans, and critically constitute our entry into language in processes of language acquisition (Cowley, 2007; Levinson, 1983; Peräkylä, 2008; Tomasello, 2000). This bias in the Western language sciences has led to verbal signals being studied in isolation, neglecting their typical co-occurrence with other kinds of signals in the context of conversation. However, more recent developments, such as Conversation Analysis (Goodwin et al., 1987) do conceive language as only one among many communicative resources in the multimodal activity that constitutes conversation. Researchers in these and related fields thus argue in favor of considering language as essentially emerging from the interaction of different communicative semiotic resources, granting vocalizations a referential function initially on the basis of their co-occurrence with visual cues such as for instance pointing and gazing, both on phylogenetic (Altenmüller et al., 2013; Leavens et al., 2010; Levinson and Holler, 2014; Liebal et al., 2013; Partan and Marler, 1999; Taglialatela et al., 2011) and ontogenetic (see Emery, 2000 for a review) time scales.

On these views, understanding the ways in which different communicative resources interact, affecting each other's dynamics, becomes as important as understanding each resource in isolation. With their integrated message model (IMM), Bavelas and Chovil (2000) provide a useful theoretical framework to understand the interaction of cues from different modalities. Their model distinguishes between two main communicative resources in face-to-face dialogue, depending on the modality in which they are expressed: 1) Auditory Acts of Meaning, and 2) Visual Acts of Meaning (AAM, and VAM, respectively). The interaction between the basic meaning of each of these "acts" will contribute to constraining inferential processes towards a specific interpretation. A common way in which acts of meaning conveyed through different modalities interact is by pure "redundancy" – that is, the basic meaning of each act contributes to constraining inferences towards the same interpretation. For instance, a child could be smiling while saying "I'm so happy!" where both modalities indicate that he or she is happy. This contrasts with how an older individual might exclaim "I'm so happy!" with an accompanying neutral face, marking the interpretation of the IMM as not being straightforward (e.g., irony). Redundancy, however, does not mean "repetition" – rather, the authors propose that this functional overlap is only partial, and contributes to nuances in the perceived communicative intention behind the IMM.

In this study, we explore communicative strategies that exploit the functional overlap of AAM (spoken verbal utterances) and VAM (eye-gaze) in conveying referential information about entities that are immediately co-present to both interactants. This enables the establishment of "common ground" (Rączaszek-Leonardi et al., 2014) between the interactants, deemed essential for communication (see Tomasello, 2010). Note that this functional overlap is what Bavelas and Chovil (2000) described as "redundancy", since the function of both vocalizations and eye-gaze is to refer to immediately co-present entities. We compare this communicative strategy – largely overlapping in function with the linguistic category of *deictics* – and protodeclaratives as observed in human infants and extant great apes (Gómez, 1996; 2005; 2007), hypothesizing that it could a have been key for the development of critical cultural innovations and abilities (such as toolmaking, hunting, cooking) in the history of our ancestors.

1.2. The role of bodily constraints in conveying referentiality

Understanding the relationship between bodily constraints and linguistic communication (Tylén et al., 2013), as well as the development of the morphological features themselves in the history of our genus, can help us reconstruct the communicative practices of our ancestors. This can be done by inferring the effectiveness of specific communicative resources afforded by the morphological constraints of our body. For example, we assume that our hominin ancestors could point insofar as their arms were freed from the task of locomotion either by brachiating, like orangutans, or by quadrupedal locomotion, like gorillas.

By informing our theories about the communication of extinct populations with anatomical, genetic, and archaeological data, we can constrain our inferences about their capabilities. This has already been done with Neanderthals (see Johansson, 2015 for a review). For example, Mendizábal and Ferreras have increased our understanding of the evolution of human linguistic capabilities by looking at bodily constraints observable in the fossil records. Their study (Mendizábal and Ferreras, 2009) argues for the ability of other hominin species (*Homo neanderthalensis* and *Homo heidelbergensis*) to communicate in complex ways through vocalizations, much like humans today. Their conclusions were based on the identification of physiological features of the outer and middle ear that are crucial for both the production and comprehension of complex vocal streams in contemporary humans.

In a way not dissimilar to the approach of Mendizábal and Ferreras, we can base inferences about our ancestors' communicative strategies (such as protodeclaratives) by considering the perceptual affordances of their anatomical features. In particular, we are interested in the affordances of the ocular morphology for gaze-following from conspecifics (Kobayashi and Koshima, 1997; 2001; Mayhew, 2013; Mayhew and Gómez, 2015; Perea García, 2016). A first step to do this would be exploring the interaction between linguistic and eye-gaze signals in conveying referential meaning. As far as we know, there is no way to know what portion of the sclera was exposed in our hominin ancestors from the fossil record because, unlike bones, the structural integrity of the eyeball decays rapidly after cellular death. Following the proposal by Cerqueira et al. (2012), their method for the prediction of *Homo* pigmentation phenotype from genomic data could be used to investigate the pigmentation patterns of extinct hominins.

If we assume that a de-pigmented sclera enables a certain degree of visual referentiality (Kobayashi and Koshima, 1997; 2001; Tomasello et al., 2007) in the behavioral pattern of protodeclaratives it follows that referentiality as instantiated by eye-gaze depends on scleral de-pigmentation. Based on the degree of their scleral de-pigmentation involved in instantiating cross-modal reference, we can reconstruct the communicative strategies potentially available to our ancestors. Further, this motivates hypotheses about the pervasiveness of these communicative strategies in different social strata, based on the intraspecific distribution of the trait, most likely depending on factors such as age, sex, and status, etc., since scleral de-pigmentation seems to be modulated by these factors in extant great apes (Mayhew and Gómez, 2015; Perea García, 2016). Using methods such as those proposed by Cerqueira et al. (2012) to identify the stage in the evolution of our genus in which our ancestors started displaying an ocular morphology resembling our own, we could triangulate the properties of communication through vocalizations, such as the frequency and range of effectiveness of deictics enabled by eye-gaze as the only VAM to provide spatial cues about the location of the intended referent. In other words, we could determine when and under which pressures the pale sclera that today characterizes some great ape species would have become an adaptive trait, rather than a spandrel arising as the product of other adaptive trends (such as, presumably, neoteny). This could be a little piece in the broader field of evolutionary linguistics as tackled, for example, by Mendizábal and Ferreras (2009) or Dediu and Levinson (2013).

1.3. Previous studies on referentiality as instantiated through eye-gaze

It is generally assumed that eye-gaze cues provide important referential information in face-to-face communication. However, most theoretical and empirical studies have looked at the interaction between gestures (or facial expressions) and linguistic communication. Even though a handful of other studies have concerned themselves with the referential affordances of eye-gaze in isolation, the present study is, to our knowledge, the first one that experimentally manipulates eye-gaze cues to explore the interaction between eye-gaze and vocalizations in instantiating referentiality in interaction. While humans generally are able to invoke entities that are not perceptually available (*displacement*, Hockett, 1969), a good deal of human communication still involves interacting with entities that are co-present to both parties in the communicative dyad (Peräkylä, 2008). We suggest that the morphological constraints of the human eye afford a refined deictic function that partly overlaps with the referential functions of human vocalizations articulated in fully developed linguistic structures. This functional overlap is likely to have its basis in the emergence of spoken linguistic reference (AAM) from VAM (of which we are concerned mainly with the referentiality afforded by eye-gaze) both in the ontogenetic and phylogenetic development of language. If this is the case, referentiality conveyed through visual means should interact with vocal verbal referentiality. As a first step to investigate this co-development of articulated speech and conspicuous eye-gaze, our study tests the hypothesis that there is an interaction between vocal-verbal and eye-gaze cues in instantiating referentiality.

Some previous studies have empirically assessed the referential function of eye-gaze cues in isolation. For example, Gamer and Hecht (2007) measured the efficiency of eye-gaze as a referential communicative resource as a function of the range of the cone. Others have observed a correlation of eye-gaze and vocalizations in conveying referentiality in naturalistic settings, helping to outline aspects such as the temporal dynamics of eye-gaze and verbal cues in instantiating referentiality (Hanna and Brennan, 2007; Brennan et al., 2008), the involvement of eye-gaze in regulating turns of speech (Vertegaal et al., 2001; Ho et al., 2015), as well as more refined measurements such as the typical span of time between gazing and verbal cues (Allopenna et al., 1998; Griffin and Bock, 2000; Richardson et al., 2007). "Naturalistic" observations have also been extensively done in computer-mediated communication (Ishii and Kobayashi, 1992; Cherubini et al., 2010; Gale and Monk, 2000; Monk and Gale, 2002), but their relevance for the present study is, for the most part, minimal, due to the constraints of the different communicative contexts (face-to-face vs. computer-mediated). Lobmaier et al.'s (2013) approach has the potential to uncover the interaction between different communicative modalities, but they focus on the relationship between facial expressions, and eye-gazing cues (more specifically how facial expressions might affect the accuracy of gaze perception). Other studies

have addressed the interaction between linguistic and eye-gazing cues in interaction by manipulating either the referents (Hanna et al., 2003), or verbal cues (Arnold et al., 2013). Lastly, one study does focus on the interaction between eye-gazing and verbal cues (the topic of the present investigation) by manipulating both as independent variables (Macdonald and Tatler, 2013). However, their study required the interference of a confederate in order to modulate the degree to which verbal cues were ambiguous. The results strongly suggest that gaze-cues can be used as a spatial pointer when verbal cues do not contain all the necessary information to unambiguously constrain attention to the intended referent. Even though all of these studies provided insights into the effectivity of gaze-cues as a referential semiotic resource, or into the interaction between gaze-cues and vocal-verbal cues, none of them manipulates gaze-cues as the independent variable.

1.4. Hypotheses

The present study set out to experimentally investigate the interaction between eye-gaze cues and vocalizations in a naturalistic cooperative task by manipulating access to eye-gaze alone as the independent variable. This allows us to test the assumption that *gaze-cues provide important spatial-referential information that facilitates coordination in cooperative tasks in peripersonal space when hands are not readily available*.

If gaze-cues play a critical role in the establishment of multimodal reference, we should observe interruption of the communicative efficacy when these cues are disrupted. This motivates Hypothesis 1.

Hypothesis 1. *Joint performance in a spatial coordination task will be lower if gaze-cues are unavailable relative to a situation where participant have access to these cues.*

This degradation of performance could follow from instances of ambiguous spatial reference or misunderstandings when eye cues are not available as communicative pointers to referents. This leads to Hypothesis 2:

Hypothesis 2. *The frequency of episodes of conversational repair will be higher in conditions without eye cues than in conditions where these cues are available.*

However, we are often in situations where we need to communicate without full visual access to each other's bodily cues (e.g. when talking on the telephone or between different rooms). We could therefore imagine that people, when deprived of a particular modality for multimodal reference, will flexibly compensate simply by greater reliance on other referential strategies. We might thus anticipate that joint performance in the task will remain the constant between conditions of available or unavailable eye cues, because participants compensate by flexibly shifting their strategy of reference. This leads to Hypothesis 3:

Hypothesis 3. *When participants are deprived of access to eye information they will rely on purely verbal strategies for instance in terms of verbal expressions that locate objects relative to landmarks (rather than expressions of deictics such as "over there", "this one" etc).*

2. Methods

2.1. Participants

Twenty dyads ($n = 40$, 17 females/23 males, mean age 24, sd 4.26) participated in the experiment in return for a monetary reward ($\approx$€ 10). Participants were recruited through the participant pool at the Cognition and Behavior Lab at Aarhus University and gave informed written consent following the guidelines of the local research ethical committee. Six dyads (out of the twenty) consisted of participants who knew each other prior to the experiment (this factor is accommodated in the analysis). All participants were native Danish speakers with the exception of two participants who were non-native but fully fluent Danish speakers and were paired with natives. Unlike previous investigations (Kobayashi and Koshima, 1997; 2001), we decided that the variable of interest was the contrast between the iris and sclera disregarding surrounding skin, so we did not control for the ethnicity of our participants.

2.2. Materials and procedure

The task of the participants was to collaborate to place and move a number of objects (cups) at particular spots on a table. Participants stood across from each other, separated by the table. One participant acted as the director and the other as the matcher. The director held a tablet (11.6″ screen) that displayed maps of target configurations of the objects for each trial and instructed the matcher who then would move the objects to the designated positions on the table. The maps were displayed to the director in coherence with her own point of view. The trial ended when all objects were placed according to the mapped configuration. Participants swapped roles every 4 min, until the completion of the session after four rounds (totaling 16 min of interaction for each dyad). Both participants could talk freely, but the directors were instructed not to use their hands for anything else than shifting to the next map on the tablet (no finger/hand pointing allowed). The matchers were instructed not to hover their hands over the cups nor spaces in order to avoid using their hands as pointers, either intentionally or accidentally. All participants were instructed to stand in a central position relative to the table for the duration of

the experiment. In the cases where these instructions were not followed by either the director or the matcher, we coded the utterances as invalid and excluded them from the analyses. A mat on the table had markings of target spaces distributed in an irregular pattern (thus making it hard to solve the task by reliance on a simple coordinate system to refer to positions). In the experimental condition, participants wore UV swimming goggles that allowed them to see out of the goggles normally, but did not allow others to see their eyes. That is, they could not rely on eye-gazing cues to facilitate the coordination task. This condition will heretofore be referred to as the "no-gaze" condition. In the control condition, participants had full access to their partner's' gaze direction (no goggles). This will be referred to as the "gaze" condition. The sessions were recorded with a GoPro HERO 2 video camera and Sennheiser wireless lavalier microphones. A schematic representation of the setup can be seen in Fig. 1 below. Fig. 2 shows the setup from two different angles, in the two different conditions.

2.3. Analysis

Responses to the experimental manipulation were assessed with reference to three complementary dependent variables. 1) Performance: A performance measure for each dyad was derived by counting the number of trials (=object configurations) the dyad completed within the designated time (16 min). 2) Repair: videos from the experiment were coded for instances of conversational repair. Following the well-established literature in conversation analysis, repair was defined as any expression that subsequently gave rise to instances of repetition of the same instruction, clarification, or reformulation of the instruction using a different strategy (Goodwin and Heritage, 1990; see Table 1 for an example). 3) Strategy: we also coded the strategies participants used when referring to cups or positions at the table. These were divided into two types, depending on the nature of the modalities involved: a) *verbal-visual*: any instance in which the *matcher* chose a cup, or space, based on multimodal referential cues relying on deictic references such as "that one", "over there", "this cup", and; b) *verbal only*: regardless of

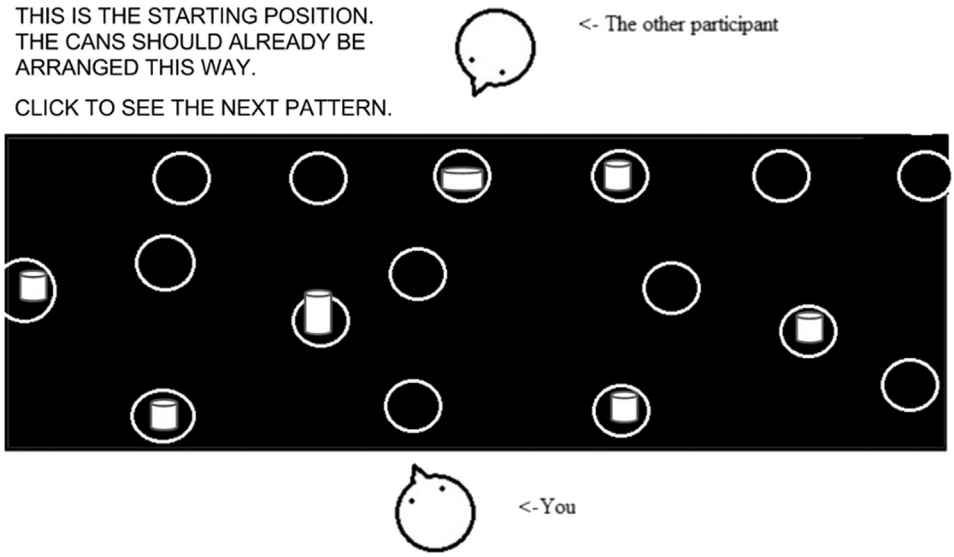

Fig. 1. Schematic representation of the setup, taken from the instructions as displayed to the director in the tablet. The graphic represents a bird-view of the participants and the table between them. Participants took turns telling each other were to place different plastic cups in empty spaces. Here translated into English, but originally in Danish.

Fig. 2. Screenshots from the footage in both conditions ("gaze" to the left; "no-gaze" to the right). Error bars express 95% confidence intervals.

Table 1
Actual examples from the three categories (repair, visual–verbal, and verbal only) taken from the transcriptions. Below are their approximate translations in English.

Repair	Visual–verbal	Verbal only
DA: *til hvilken side?*	DA: *må jeg sige "den der"?*	DA: *og så skal du ta den tættest på vinduerne*
EN: *to which side?*	EN: *can I say "that one"?*	EN: *and then you should take the one closest to the windows*

whether the director provided visual cues, or not, the minimally required information for the *matcher* to infer the right cup, or position, was completely and unambiguously inferable from the director's speech content (see Table 1 for examples of the two categories). While one expert coder coded the full video material, approximately 10% of video material was coded independently also by a second coder in order to assess intercoder reliability. This subset corresponded to 264 instances (speech turns) of assigned strategy. Intercoder reliability was calculated in percent and Cohen's Kappa (Cohen, 1960) using the IRR package in RStudio 0.98.978 (RStudio Team, 2015).

We used a multiple linear regression approach to test hypothesized dependencies between our variables of interest relying again on the statistical software RStudio. In all analyses we included an extra variable as fixed effect in order to control for potential unsystematic variance due to whether dyad members knew each other in advance. This variable will be referred to as 'known' in the following sections. The first analysis tested Hypothesis 1: whether the experimental manipulation had an impact on task performance, that is, if access to eye-gazing cues facilitates coordination in the task thus helping participants to complete more trials in the designated time. That is, task performance was the outcome variable, while the predictor variables were *condition* and '*known*'. Notice that since performance is a count variable this was run as a Poisson model. The second analysis tested Hypothesis 2: that the experimental manipulation would make spatial references ambiguous and thus interrupt the coordination giving rise to more instances of conversational repair. The probability of a speech turn being followed by an instance of *repair* was the outcome variable, while the predictor variables was still condition and 'known'. The third analysis tested Hypothesis 3: that the experimental manipulation would give rise to compensatory strategies such as greater reliance on verbal descriptions not involving multimodal deictic expressions. The outcome was thus the dyad-wise probability of using either the verbal-visual or the verbal only strategy. The predictors were again the condition and 'known'. It follows from Hypothesis 3 that performance in the two conditions could be contingent on different strategies, that is, while in the gaze condition a visual–verbal strategy could be very effective, it would not work in the no-gaze condition. This motivated the prediction that condition and strategy might interact in predicting performance. This was tested using a Poisson model with performance as dependent variable and strategy and condition as independent variables.

3. Results

Intercoder reliability of the "strategy" variable between the two independent coders was calculated to 77%, k = 0.38, corresponding to "fair agreement" (Landis and Koch, 1977). *Performance*: dyads in the "no-gaze" condition performed worse on average (M = 31.66, SD = 7.75) than dyads in the control condition (M = 38.77, SD = 6.78) as measured by number of trials completed. This effect was significant, $\beta = -0.19$ (SE = 0.09), $z(17) = 2.01$, $p = .044$, lending support for Hypothesis 1 (see Fig. 3a). The known variable was not a significant predictor of this effect and did not interact with condition.

Repair: dyads in the "no-gaze" condition experienced a higher probability of episodes of conversational repair (M = 0.30, SD = 0.06) than dyads in the control condition (M = 0.23, SD = 0.05). Again this effect was significant supporting Hypothesis 2:

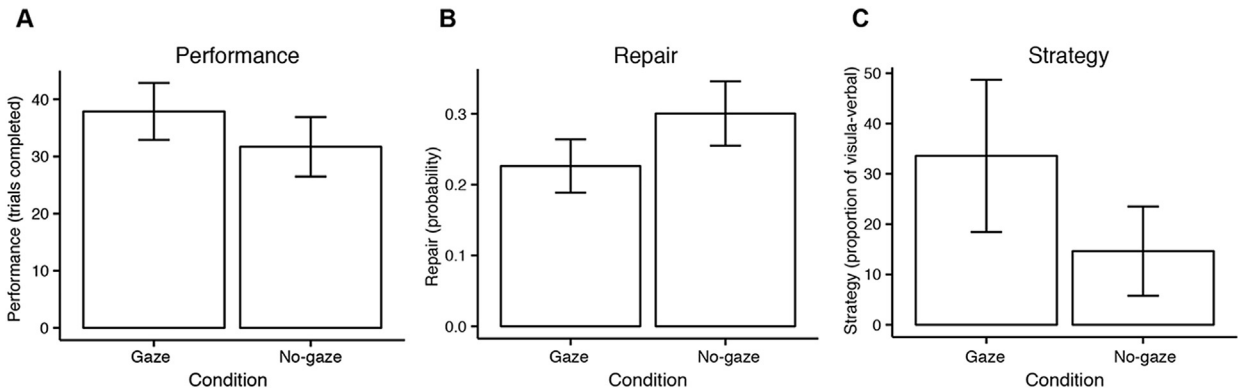

Fig. 3. Results. Effect of the two conditions, that is the "gaze" group (who had unrestricted access to gaze-cues), and the "no-gaze" group (whose access to gaze-cues was obstructed by goggles) on the three dependent variables. A: The relation between condition and performance. B: The relation between condition and the probability of experiencing episodes of conversational repair. C: The relation between condition and relative frequency of strategies (the proportion of visual–verbal relative to verbal-only).

$ß = 0.07$ (SE = 0.03), $t(17) = 2.41$, $p = .02$ (see Fig. 3b). There was no main or interaction effects of the 'known' variable. *Strategy*: dyads in the "no-gaze" condition had a significantly lower proportion of the 'verbal-visual' strategy (M = 14.65, SD = 12.40) than the dyads in the control condition (M = 33.58, SD = 21.13), lending support for Hypothesis 3: $ß = -15.24$ (SE = 7.06), $t(17) = -2.16$, $p = .046$ (see Fig. 3c). For this test, however, there was also a significant main effect of 'known': $ß = 18.47$ (SE = 7.71), $t(17) = 2.39$, $p = .028$. Lastly, condition and strategy were found to interact in predicting performance, $ß = 0.01$ (SE = 0.00), $z(17) = 2.73$, $p < .01$.

4. Discussion

The results lend support to the hypothesized role of eye-gaze in instantiating referentiality in cooperative tasks within peripersonal space. When eye-gaze was blocked (in the "no-gaze" condition), the efficiency of multimodal strategies diminished, resulting in lower performance. Furthermore, in the no-gaze condition participants more often entered into sequences of conversational repair – for instance prompting their partner to repeat reference to the intended cup (or slot). The absence of gaze-cues seemingly led to more ambiguity in reference which indicates that eye-gaze is used deictically to assist the establishment of reference. Lastly, our results suggest that participants in the no-gaze condition resorted to other communicative strategies than participants in the gaze condition, relying more on verbal expressions containing all the necessary referential information to successfully restrain attention to the intended referent. Interestingly, and partly against the predictions of Hypothesis 3, the increased reliance on purely verbal strategies in the no-gaze condition, however, was not sufficient to compensate for the lack of gaze information – i.e. not sufficient to keep performance on level with the "gaze" condition. These findings are in accordance with Macdonald and Tatler's (2013) suggestion that eye-gaze is used as a spatial pointer that interactants selectively recur to when verbal cues do not contain all the necessary information to unambiguously constrain attention to the intended referent. It is important to remark that we did not directly measure how much participants relied on eye-gaze across conditions – rather, we infer that the lack of access to eye-gaze in the "no gaze" condition motivated the differences across conditions. In the "no gaze" condition, participants could still make use of a variety of other visual cues other than eye-gaze, such as the nose, direction of the face, etc. Even though follow-ups to this study would ideally control for other means of effecting referentiality by orchestrating VAM and AAM that do not rely on eye-gaze, our instructions for participants (see Materials and procedure) and our coding strategy ensured that VAM were not instantiated with hands, arms, or body position.

An interesting side-finding in relation to participants' choice of strategy was that familiarity between the participants (whether they knew each other before) also significantly modulated this behavior. Monitoring the director's eye-gaze requires a certain degree of eye contact. We speculate that in cases where participants did not know each other in advance but were just introduced prior to the experiment, they could be more prone to gaze aversion (i.e. avoid the gaze of others, see Emery, 2000 for a review) and thus less inclined to engage in direct eye-contact. This would work against a strategy of conveying information by visual means through eye-gaze, forcing participants to convey more information through verbal cues. To effectively rely on a multimodal visual–verbal strategy, communicative partners thus might have to overcome their reluctance to engage in eye contact. Similarly, our dyads were composed by men and women. Since perceived gender has an effect on dominance and attractiveness, it is possible that there was an effect of perceived gender on the gazing behavior of our participants (Keating, 1985). However, we did not ask our participants about their sexual orientation so we were unable to assess these effects, if any.

More generally, our results can be taken as an indication that it would have been difficult – if not impossible – for other hominins to convey referential meaning for co-present entities while using their hands for other purposes, unless: a) they were able to efficiently coordinate VAM with AAM and b) they presented an ocular morphology that, like ours, eased gaze-following from conspecifics. It might well also have been the case that a de-pigmented sclera facilitated gaze-following, and that this would have in turn facilitated coordination between interactants, but it does not necessarily follow that gaze-cues would have been the only way they had to visually restrain attention. We propose that it is precisely the new affordances brought about by a de-pigmented sclera (with little or no adaptive value in what regards referentiality at the beginning of the development of the trait) that could have facilitated the enaction of certain cooperative practices especially ones calling for fast and efficient online coordination (like cooperative cooking, hunting, tool-making, etc.). These practices would in turn become increasingly important, eventually becoming an adaptive pressure in having an eye morphology that eased gaze-following.

4.1. The role of bodily constraints in conveying referential information

The participants in the no-gaze condition were all fluent speakers of Danish, and could partly make up for the loss of referential information conveyed visually (although they still suffered from relatively lower performance compared to the gaze condition) by verbally constraining attention to the intended referent, either by invoking its unique spatial disposition relative to the participants (ie., "the one *closest to you*"), relative to absolute landmarks of the setting ("the cup *at the bottom*"), or relative to other landmarks ("the one close *to the one you just moved*"). Sometimes, when available, participants would also recur to specifying salient aspects of the intended referent (i.e., "the *small* cup").

Our hominin ancestors, however, would not have had a fully developed vocalic-referential communicative system such as ours. It seems thus unlikely that they could have successfully coordinated their attention to the same entities in activities such

as this (cooperative tasks in peripersonal space) in the absence of VAM instantiated through hands for intentional pointing, or a conspicuous eye-morphology to enact reference as a by-product of attending to an entity. Learning about the origins and development of a de-pigmented sclera in our genus (Cerqueira et al., 2012) could thus tell us the range of situations in which our hominin ancestors were able to precisely convey referential information.

4.2. Multimodality at the bases of human linguistic communication

Protodeclaratives in apes, like deictics in humans, require the integration of visual and auditory cues in order to function as proper pointers. It seems like multimodal referentiality is, at least in a looser sense, a pervasive communicative strategy among great apes, with the case of humans showing the highest degree of flexibility (and, perhaps, volitional control) in articulating vocal-verbal cues, eventually culminating in the development of fully-fledged linguistic abilities. Crucially, and of relevance to the field of evolutionary linguistics, we propose that deictics themselves could have derived from a basal ability to indicate interest through vocalizations while visually pointing (through eye-gaze) at the referent that motivates such attention simply as a by-product of attending to it, because of the hyper-conspicuity afforded by a de-pigmented sclera. This could provide an explanation for the apparent gap between voluntary and involuntary communication that has so notoriously been borne as an essential difference between human and nonhuman communication systems.

Some of the most widely accepted language origins accounts have described linguistic interaction as emerging from a two-way communicative system, which had the function (mainly) of regulating the interaction between two agents, to a triadic, referential interaction in which attention is co-directed to an entity or event outside of the communicative dyad (Carpenter et al., 1998; Tomasello et al., 2005). Recent evidence strongly supports the idea that nonhuman great apes regularly engage in this kind of interactions when born in the wild (Sumatran orangutan: Gruber, 2014), in captivity (Western gorilla: Tanner and Byrne, 2010), and in enculturated settings (Lowland gorilla; Gómez, 2010) (but see also Tomonaga et al., 2004). An increasing quantity of evidence also points at the occurrence of gestures in sequences, resembling a proto-grammar (see Perlman and Tanner, 2016. However, the degree to which this is facilitated by their peculiar eye morphology remains to be investigated. Recent work on the flexibility of production and context dependence of monkey (Lemasson, 2003; Bouchet et al., 2013) and great ape (orangutans: Lameira et al., 2015) vocalizations challenge the traditional view that they are largely fixed, and invariable. Vocalizations emerge thus again as a likely important component of early protolinguistic forms in our hominin ancestors (Liebal et al., 2013). With this, a new picture emerges in which human communication arises as the result of the complex interplay between intentional and unintentional cues of different nature (Kendon, 2011), rather than a monolithic, linear progression of one, two, or any specific number of communicative resources throughout the evolution of our genus.

Funding

This work was supported by the Interacting Minds Centre at Aarhus, Denmark.

Acknowledgements

This project has been supported by seed funding 2015 from the Interacting Minds Centre, Aarhus University. The Cognition and Behavior Lab, Aarhus Universitet, provided material resources that were necessary for the experimental setup.

References

Allopenna, P.D., Magnuson, J.S., Tanenhaus, M.K., 1998. Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models* 1,* 2,* 3,* 4,* 5. J. Mem. Lang. 38 (4).
Altenmüller, E., Schmidt, S., Zimmermann, E. (Eds.), 2013. The Evolution of Emotional Communication: From Sounds in Nonhuman Mammals to Speech and Music in Man. OUP, Oxford.
Arnold, J.E., Kaiser, E., Kahn, J.M., Kim, L.K., 2013. Information structure: linguistic, cognitive, and processing approaches. WIREs Cogn. Sci. 4, 403–413. http://dx.doi.org/10.1002/wcs.1234.
Baldwin, D.A., 1993. Early referential understanding: infants' ability to recognize referential acts for what they are. Dev. Psychol. 29, 832–843.
Bavelas, J.B., Chovil, N., 2000. Visible acts of meaning: an integrated message model of language in face-to-face dialogue. J. Lang. Soc. Psychol. 19 (2), 163–194.
Bouchet, H., Laporte, M., Candiotti, A., Lemasson, A., 2013. Flexibilité vocale sous influences sociales chez les primates non-humains. Revue de primatologie (5).
Brennan, S.E., Chen, X., Dickinson, C.A., Neider, M.B., Zelinsky, G.J., 2008. Coordinating cognition: the costs and benefits of shared gaze during collaborative search. Cognition 106 (3), 1465–1477.
Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., Moore,, C., 1998. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. Monogr. Soc. Res Child Dev. i–174.
Cerqueira, C., Paixão-Côrtes, V.R., Zambra, F., Salzano, F.M., Hünemeier, T., Bortolini, M.C., 2012. Predicting Homo pigmentation phenotype through genomic data: from Neanderthal to James Watson. Am. J. Hum. Biol. 24 (5), 705–709.
Cherubini, M., De Oliveira, R., Oliver, N., Ferran, C., 2010. Gaze and Gestures in Telepresence: Multimodality, Embodiment, and Roles of Collaboration arXiv preprint arXiv:1001.3150.
Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46.
Corballis, Michael C., 2002. From Hand to Mouth: The Origins of Language. Princeton University Press.
Cowley, S.J., 2007. The cognitive dynamics of distributed language. Lang. Sci. 29 (5), 575–583.
Dediu, D., Levinson, S.C., 2013. On the antiquity of language: the reinterpretation of Neandertal linguistic capacities and its consequences. Front. Psychol. 4.
Dunham, P.J., Dunham, F., Curwin, A., 1993. Joint-attentional states and lexical acquisition at 18 months. Dev. Psychol. 29 (5), 827.

Emery, N.J., 2000. The eyes have it: the neuroethology, function and evolution of social gaze. Neurosci. Biobehav. Rev. 24, 581–604.
Gamer, M., Hecht, H., 2007. Are you looking at me? Measuring the cone of gaze. J. Exp. Psychol. Hum. Percept. Perform. 33 (3), 705–715. http://dx.doi.org/10.1037/0096-1523.33.3.705.
Gale, C., Monk, A.F., 2000. Where am I looking? The accuracy of video-mediated gaze awareness. Percept. Psychophys. 62 (3), 586–595.
Gómez, J.C., 1996. Ostensive behavior in great apes: the role of eye contact. In: Reaching into Thought: The Minds of the Great Apes, pp. 131–151.
Gómez, J.C., 2005. Species comparative studies and cognitive development. Trends cognitive Sci. 9 (3), 118–125.
Gómez, J.C., 2007. Pointing behaviors in apes and human infants: a balanced interpretation. Child. Dev. 78 (3), 729–734.
Gómez, J.C., 2010. The ontogeny of triadic cooperative interactions with humans in an infant gorilla. Interact. Stud. 11 (3), 353–379.
Goodwin, C., Goodwin, M.H., Brenneis, D., Thompson, S.A., Mann, W.C., Woolard, K., 1987. IPRA Papers in Pragmatics vol. 1, no. 1.
Goodwin, C., Heritage, J., 1990. Conversation analysis. Annu. Rev. Anthropol. 19, 283–307.
Griffin, Z.M., Bock, K., 2000. What the eyes say about speaking. Psychol. Sci. 11 (4).
Gruber, T., 2014. Wild-born orangutans (*Pongo abelii*) engage in triadic interactions during play. Int. J. Primatol. 35 (2), 411–424.
Hanna, J.E., Brennan, S.E., 2007. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. J. Mem. Lang. 57 (4), 596–615.
Hanna, J.E., Tanenhaus, M.K., Trueswell, J.C., 2003. The effects of common ground and perspective on domains of referential interpretation. J. Mem. Lang. 49 (1), 43–61.
Hockett, C.F., 1969. The origin of speech. Sci. Am. 203, 88–111.
Ishii, H., Kobayashi, M., May, 1992. Clearboard: a seamless medium for shared drawing and conversation with eye contact. In: Proc. of CHI'92, pp. 525–532. Monterey, CA, USA.
Johansson, S., 2015. Language abilities in Neanderthals. Annu. Rev. Linguist. 1 (1), 311–332.
Keating, C.F., 1985. Gender and the physiognomy of dominance and attractiveness. Soc. Psychol. Quart. 48 (1), 61–70.
Kendon, A., 2011. Gesture first or speech first in language origins. In: Deaf Around the World: The Impact of Language, pp. 251–267.
Kobayashi, H., Kohshima, S., 1997. Unique morphology of the human eye. Nature 387, 767–768.
Kobayashi, H., Kohshima, S., 2001. Unique morphology of the human eye and its adaptive meaning: comparative studies on external morphology of the primate eye. J. Hum. Evol. 40 (5), 419–435.
Lameira, A.R., Hardus, M.E., Bartlett, A.M., Shumaker, R.W., Wich, S.A., Menken, S.B., 2015. Speech-like rhythm in a voiced and voiceless orangutan call. PloS One 10 (1), e116136.
Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174.
Leavens, D.A., Russell, J.L., Hopkins, W.D., 2010. Multimodal communication by captive chimpanzees (*Pan troglodytes*). Anim. Cogn. 13 (1), 33–40.
Lemasson, A., 2003. Communication vocale et organisation sociale chez la mone de Campbell (Cercopithecus campbelli): partage vocal et relation sociales (Phd Université de Rennes1–France).
Levinson, S.C., 1983. Pragmatics (Cambridge textbooks in linguistics).
Levinson, S.C., Holler, J., 2014. The origin of human multi-modal communication. Philos. Trans. R. Soc. Lond. B Biol. Sci. 369 (1651), 20130302.
Liebal, K., Waller, B.M., Slocombe, K.E., Burrows, A.M., 2013. Primate Communication: A Multimodal Approach. Cambridge University Press.
Linell, P., 1982. The Written Language Bias in Linguistics. University of Linköping, Department of Communication Studies, Linköping, Sweden.
Lobmaier, J.S., Hartmann, M., Volz, A.J., Mast, F.W., 2013. Emotional expression affects the accuracy of gaze perception. Motivation Emot. 37 (1), 194–201.
Love, N., 2004. Cognition and the language myth. Lang. Sci. 26 (6), 525–544.
Macdonald, R.G., Tatler, B.W., 2013. Do as eye say: gaze cueing and language in a real-world social interaction. J. Vis. 13 (4), 6.
Mayhew, J.A., 2013. Attention cues in apes and their role in social play behavior of western lowland gorillas (*Gorilla gorilla gorilla*) (Doctoral dissertation, University of St Andrews).
Mayhew, J.A., Gómez, J.C., 2015. Gorillas with white sclera: a naturally occurring variation in a morphological trait linked to social cognitive functions. Am. J. Primatol. 77 (8), 869–877.
Mendizábal, I.M., Ferreras, J.L.A., 2009. El origen del lenguaje: la evidencia paleontológica. Munibe Antropologia-Arkeologia 60, 5–16.
Monk, A.F., Gale, C., 2002. A look is worth a thousand words: full gaze awareness in video-mediated conversation. Discourse Process. 33 (3), 257–278.
Mundy, P., Block, J., Delgado, C., Pomares, Y., Van Hecke, A.V., Parlade, M.V., 2007. Individual differences and the development of joint attention in infancy. Child. Dev. 78 (3), 938–954.
Partan, S., Marler, P., 1999. Communication goes multimodal. Science 283 (5406), 1272.
Peräkylä, A., 2008. Analyzing talk and text. In: Collecting and Interpreting Qualitative Materials, pp. 351–374.
Perea García, J.O., 2016. Quantifying ocular morphologies in extant primates for reliable interspecific comparisons. J. Lang. Evol. J. Lang. Evol. 2 (1).
Perlman, M., Tanner, J., 2016. Moving beyond 'meaning': Gorillas combine gestures into sequences for creative display. Lang. Commun..
Poggi, I., D'Errico, F., Spagnolo, A., 2009, February. The embodied morphemes of gaze. In: International Gesture Workshop. Springer, Berlin, Heidelberg, pp. 34–46.
Rączaszek-Leonardi, J., Dębska, A., Sochanowicz, A., 2014. Pooling the ground: understanding and coordination in collective sense making. Front. Psychol. 5.
Richardson, D.C., Dale, R., Kirkham, N.Z., 2007. The art of conversation is coordination. Psychol. Sci. 18 (5).
RStudio Team, 2015. RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. URL. http://www.rstudio.com/.
Ho, S., Foulsham, T., Kingstone, A., 2015. Speaking and listening with the eyes: gaze signaling during dyadic interactions. PloS One 10 (8), e0136905.
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 (4), e18852.
Tanner, J.E., Byrne, R.W., 2010. Triadic and collaborative play by gorillas in social games with objects. Anim. Cogn. 13 (4), 591–607.
Tomasello, M., 2000. First steps toward a usage-based theory of language acquisition. Cogn. Linguist. 11 (1/2), 61–82.
Tomasello, M., 2010. Origins of Human Communication. MIT Press.
Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H., 2005. Understanding and sharing intentions: the origins of cultural cognition. Behav. Brain Sci. 28 (05), 675–691.
Tomasello, M., Farrar, M.J., 1986. Joint attention and early language. Child. Dev. 57, 1454–1463.
Tomasello, M., Hare, B., Lehmann, H., Call, J., 2007. Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. J. Hum. Evol. 52 (3), 314–320.
Tomasello, M., 2009. The Cultural Origins of Human Cognition. Harvard University Press.
Tomonaga, M., Tanaka, M., Matsuzawa, T., Myowa-Yamakoshi, M.A.S.A.K.O., Kosugi, D., Mizuno, Y., …, Bard, K.A., 2004. Development of social cognition in infant chimpanzees (*Pan troglodytes*): face recognition, smiling, gaze, and the lack of triadic interactions. Jpn. Psychol. Res. 46 (3), 227–235.
Tylén, K., Fusaroli, R., Bundgaard, P.F., Østergaard, S., 2013. Making sense together: a dynamical account of linguistic meaning-making. Semiotica 2013 (194), 39–62.
Vertegaal, R., Slagter, R., Van der Veer, G., Nijholt, A., 2001, March. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, Chicago, pp. 301–308.