Multimodal-first or pantomime-first?
Communicating events through pantomime with and without vocalization

Jordan Zlatev, Sławomir Wacewicz*, Przemyslaw Zywiczynski, Joost van de Weijer

* corresponding author

Jordan Zlatev
Centre for Languages and Literature
Lund University
Box 201
221 00 Lund, Sweden
jordan.zlatev@semiotik.lu.se

Sławomir Wacewicz
Center for Language Evolution Studies; Department of English
Nicolaus Copernicus University
Bojarskiego 1 (Room C.3.32)
87–100 Toruń, Poland
wacewicz@umk.pl

Przemysław Żywiczyński
Center for Language Evolution Studies; Department of English
Nicolaus Copernicus University
Bojarskiego 1 (Room C.3.32)
87–100 Toruń, Poland
przemek@umk.pl

Joost van de Weijer
Centre for Languages and Literature
Lund University
Box 201

221 00 Lund, Sweden
joost.van_de_weijer@ling.lu.se

Abstract

A persistent controversy in language evolution research has been whether language emerged in the gestural-visual or in the vocal-auditory modality. A “dialectic” solution to this age-old debate has now been gaining ground: language was fully multimodal from the start, and remains so to this day. In this paper, we show this solution to be too simplistic and outline a more specific theoretical proposal, which we designate as *pantomime-first*. To decide between the multimodal-first and pantomime-first alternatives, we review several lines of interdisciplinary evidence and complement it with a cognitive-semiotic experiment. In the study, the participants saw – and then matched to hand-drawn images – recordings of short transitive events enacted by 4 actors in two conditions: visual (only body movement), and multimodal (body movement accompanied by nonlinguistic vocalization). Significantly, the matching accuracy was greater in the visual than the multimodal condition, though a follow-up experiment revealed that the emotional profiles of the events enacted in the multimodal condition could be reliably detected from the sound alone. We see these results as supporting the proposed pantomime-first scenario.

Keywords: bodily mimesis, pantomime, multimodal origins, multimodality, language origins, language evolution, gesture

1. Introduction

From the early speculations on language origins involving philosophers like Condillac and Herder (cf. Johansson, 2005), until recent more or less empirically grounded theories of language evolution, there has been a persistent controversy on whether language emerged in the gestural-visual or in the vocal-auditory modality (cf. Fitch, 2010; Zywiczynski & Wacewicz, 2015). Since the intensification of research in language evolution in the last decades of the previous century, theories of "gestural primacy" have been strongly influential. Hewes (1973) summarized much of the evidence available at the time in their favor, highlighting the relative success of teaching manual signs to non-human apes in contrast to the failures to teach them speech (e.g. Gardner & Gardner, 1969). Later comparative research indicated greater flexibility and intentionality in gestural than vocal signals in apes' spontaneous communication (Call & Tomasello, 2007; Pika, 2008; Tomasello 2008), which was complemented with neuropsychological arguments pertaining to handedness and lateralization (Corballis, 2003) and mirror neuron systems (Arbib, 2005). Support for gestural primacy also came from the growing realization that the signed languages of the deaf constitute full-fledged languages (Stokoe, 1991; Armstrong, Stokoe & Wilcox, 1995).

On the opposing front, "speech-first" theorists take the present-day dominance of speech and some apparent adaptations for it in the vocal apparatus as arguments that language must always have existed in the vocal-auditory modality (e.g. Dunbar, 1996; MacNeilage, 2008). Speech-first proponents tend to emphasize the pre-linguistic role of vocalization, e.g. by pointing to its bonding function (Dunbar, 1996). Furthermore, the alleged rigidity of ape vocalizations, claimed as an argument for "gesture-first" (e.g. Tomasello, 2008), is currently under reconsideration (cf. See, 2014; Section 2.3).

As might have been expected, a compromise solution to the controversy has eventually emerged, and has been gaining ground in recent years: language was neither purely vocal nor gestural from the start, but multimodal (Taglialatela, Russell, Schaeffer, & Hopkins, 2011; McNeill, 2012; Kendon, 2014b; Perlman, this issue). In this article, we argue that despite the many merits of multimodality, such a "multimodal-first" resolution of the debate is problematic. In Section 2, we point to three different sets of problems: definitional, theoretical and empirical; and propose corresponding refinements. On this basis, we outline a theoretical proposal that is arguably distinct from both gesture-first and multimodal-first theories, and which we designate as *pantomime-first*. Section 3 describes an empirical study that contrasts this theory and the multimodal-first in terms of predictions, and as we show, offers at least some indirect support for our proposal. In Section 4, we sum up the argument and conclude that the multimodality of present-day language is better regarded as an outcome of language evolution than its starting point.

2. A "multimodal" origin of language?

The notion of *multimodality* has recently gained considerable popularity in the language sciences (Vigliocco, Perniss, & Vinson, 2014). In particular, the appreciation of the multimodal nature of conversation has been productive for our understanding of the nature of present-day language (Kendon, 2014a) and has correspondingly enriched the field of language origins.[1] While we accept the fruitfulness of the general notion of multimodality, we here wish to highlight some of its limitations. In particular, rather than treating it as a "magic bullet" solution to such classical issues as the speech-first vs. gesture-first debate in language origins, multimodality is better regarded as a point of departure that needs to be further developed both conceptually and empirically. Our argument is that for so-called "multimodal theories of language origins" to be legitimate contenders, they need to rely on clear definitions and theoretical constructs, which generate falsifiable predictions that may be tested in empirical studies (Wacewicz, Zywiczynski, & Orzechowski, 2016). Accordingly, in this section we review various meanings of the term "multimodality" (2.1) and on this basis take a fresh look at the theoretical positions that appeal to this concept (2.2) and link these to corresponding empirical evidence (2.3).

2.1 Definitional issues: What is linguistic multimodality?

The first difficulty with the notion of multimodality is conceptual. For example, the statement that "language is multimodal" is in several ways ambiguous. In one sense, the claim is quite uncontroversial: human face-to-face communication is typically perceived

[1] See the special issue of *Language & Communication* "The multimodal origins of human linguistic communication" [in press].

through at least two sensory modalities, vision and audition/hearing. Blind people, and especially blind children, also combine hearing and touch (Landau & Gleitman, 2009).

In a second sense, communication involves a number of different productive modes/resources such as gaze, pointing, iconic gestures, conventional gestures, bodily postures, prosody, and spoken words. It should be noted that there is no simple correspondence between these "modes" or "modalities" and the sensory modalities. Gestures are, for example, typically perceived visually, but in some cases, (also) through hearing (e.g. hand-clapping) or touch (e.g. back-slapping) and may be more appropriately regarded as *cross-modal* (see 2.2).

Proponents of "language as a multimodal phenomenon", most often have neither of the above two senses in mind but a third, according to which speech/language and gestures figure as the two (major) equipollent semiotic modalities: "Speech and gesture are part and parcel of the same system and together constitute a tightly integrated unit, thus underscoring the need for a multimodal approach" (Vigliocco, Perniss, & Vinson, 2014: 2). While it remains debatable what defines "the same system", as we discuss in 2.3, many researchers could accept this proposal and yet claim a basic asymmetry with respect to where the bulk of the message is expressed, aligning language mostly with speaking and referring to gestures as "co-speech" rather than to speech as "co-gestural".

The challenge for language origins research is thus to explain both the close connection between speech and gestures in verbal interaction, on the one hand, and their division of labor on the other, rather than to fuse them in an undifferentiated notion of "multimodal language". As we show next, it is possible to do this in several ways, and these need to be kept distinct.

2.2 Theoretical issues: What are the alternatives?

Most current language evolution theories acknowledge multimodality in the senses outlined above. For example, many gesture-first theorists openly admit that "there never was a time when visible gestures were unaccompanied by vocalizations" (Armstrong & Wilcox, 2007: 68). On the other hand, in a well-known speech-first theory, MacNeilage (2008: 283) accepts more or less the reverse with respect to gestures. Therefore, we are dealing with a difference in emphasis rather than in kind, and we can distinguish (logically) theories that claim that at its dawn, language was expressed and perceived predominantly in (a) the bodily-visual modality, (b) the spoken-auditory modality, or (c) equally distributed (in some sense that remains to be determined) in both modalities. Leaving (b) out of the picture for present purposes, we therefore need to distinguish theories of type (a) from type (c).

Starting from the latter kind, which we call multimodal-first, perhaps the best-known example is McNeill's Growth Point theory (McNeill, 2005, 2012). According to this, speech and gesture form a single dialectic unit, with speech responsible for the propositional content, and gesture for the "imagistic" side of language. This constitutes a single psychological system that is essential for human language and cognition. In an early and still influential paper, McNeill (1985: 351) argued for this thesis on the basis of evidence, some of which has been questioned, suggesting that:

> gestures … occur only during speech, are synchronized with linguistic units, are parallel in semantic and pragmatic function to the synchronized linguistic units, perform text functions like speech, dissolve like speech in aphasia, and develop together with speech in children.

McNeill (2012) proposes that the Growth-Point system was fully established relatively late in evolution: "This development was an exclusively human phenomenon and was completed in *H. sapiens* about 200-100 kya" (ibid: 112).

A rather different multimodal-first theory is that of Kendon (2014b), who proposes that the tight multi-modal (or "poly-modal") integration we observe today – "the 'natural' state of spoken language is a speech-kinesis ensemble" (ibid: 76) – is the ancestor of evolutionarily ancient practical skills based on hand-mouth coordination like feeding. Such actions were only eventually transformed into communicative signals, leading Kendon to the conclusion that "the early steps of language evolution also consisted in multi-modal signals, instead of being predominantly hand-based or vocalization-based" (Aboitiz, 2012: 8, cited by Kendon 2014b: 69).

In contrast, theories that emphasize the visual modality tend to view bodily and vocal signals as integrated but distinct semiotic resources. The two modalities would naturally have interacted in evolution, but it is in the bodily-gestural system that the breakthrough to human specificity is assumed to have taken place. For example, Levinson and Holler (2014: 5, our emphasis) underscore “the gradual co-evolution of vocal language with a pre-existing gestural mode of communication”. As this “may have taken place over nearly a million years, … the different modalities are intertwined”. Or as expressed even more explicitly by Collins (2013: 136, our emphasis): “… human primates must have been at first better at transmitting information through gesture than through voice … only gradually shifting from a code that foregrounded gesture to one that foregrounded voice.” We should note there is no mention of speech “supplementing” gesture in such scenarios, which is the usual criticism from multimodal-first theorists like McNeill.

While such approaches are usually subsumed under the label “gesture first”, we believe that this is rather misleading. For one thing, the term “gesture” itself is ambiguous and hard to define in a clear, theory-independent way (Andrén, 2010), prompting, e.g., (Kendon, 2012) to search for more neutral alternatives like “visible bodily action”. Furthermore, the scenarios outlined by researchers such as Collins or Levinson and Holler are consistent with the hypothesis that the uniquely human cognitive-semiotic breakthrough consisted in an adaptation for bodily mimesis, allowing “conscious, self-initiated, representational acts that are intentional but not linguistic” (Donald, 1991:168). As argued by Zlatev (2014), one advantage of such a theory is that it does not suffer from the problems of traditional gesture-first models (e.g. Hewes, 1973), as it does not preclude early vocalizations and does not assume a “gestural proto-language”. Again, there is no need for any “supplantation” or a “modality switch” (e.g. Burling, 2005), as such a theory postulates a gradual and partial increase in the dependence on the vocal modality.

The type of communicative system that corresponds to, and arises from, the cognitive capacity of bodily mimesis is that of *pantomime*, defined by Zywiczynski, Wacewicz, and Sibierska (2016) as:

> a non-verbal, mimetic and non-conventionalised means of communication, which is executed primarily in the visual channel by coordinated movements of the whole body, but which may incorporate other semiotic resources, most importantly non-linguistic vocalisations [; and which may]

> holistically refer to a potentially unlimited repertoire of events, or sequences of events, displaced from the here and now.

This characterization in part overlaps with the domain covered by the general term "gesture", but is both more specific and less burdened by assumptions emanating from particular theoretical models, as indicated earlier.

Pantomime is posited as a vital stage in the evolutionary emergence of language not only by the mimesis theorists but also by other influential researchers in the field, most importantly Arbib (2005, 2012) and Tomasello (2008), but not by, e.g., McNeill (2012).[2] Crucial for the context at hand is the role of multimodality in pantomime. The status of pantomime as a mimetic mode of communication requires “a cross-modal mapping between exteroception (i.e. perception of the environment, normally dominated by vision) and proprioception (perception of one’s own body, normally through kinaesthetic sense)” (Zlatev, 2008: 19). In this way, pantomime depends on a cognitive infrastructure that operates across sensory modalities, and on being able to combine different communicative modalities in the construal and expression of meaning (Zywiczynski, Wacewicz, & Sibierska, 2016).

At the same time, pantomime is characterized by the dominance of the bodily-visual modality, due to its greater potential for iconic (resemblance-based) representations (Brown, 2012; Zlatev, 2014). Hence in the case of pantomime, the role of modalities that are not bodily-visual, and specifically the input of non-linguistic vocalization, is not predicted to increase the effectiveness of bodily visual communication in referential terms (though vocalization may well serve other functions, such as emotional expression or attention-getting).

Thus, rather than the traditional debate between gesture-first and speech-first, we are led to a more subtle, but still theoretically and empirically significant opposition between the multimodal-first and the pantomime-first theory of language origins. The

[2] It is characteristic that when addressing language evolution, McNeill (2012) rejects the relevance of pantomime, claiming that even if it was present in the communication of early hominins, it constituted an evolutionary dead-end and should not be seen as a precursor of language.

former, in either Kendon's or McNeill's version, would claim that bodily-visual and vocal-auditory signals were fully integrated at least from the onset of language, and likely even earlier. Such a theory predicts (a) no advantage for any of the modalities in the communication of non-human apes, (b) joint neural control of vocalization/speech and visible bodily communicative action, (c) inseparable connections between speech and gesture/action in language use, (d) co-emergence in child development and (e) more successful communication using multimodal performance than only bodily-visual pantomime in experimental studies.

In contrast, the pantomime-first theory would be more consistent with findings of (a') greater flexibility in bodily over vocal communication in apes, (b') a degree of dissociation in motor control of speech and pantomime, (c') partial independence between speech and gesture in use, (d') gradual integration of the modalities in development, and (e') either no advantage for multimodal performances, or actually advantages for silent pantomimes in experimental settings.

In the following sub-section, we review some of the evidence with respect to these points, focusing on the last one, which serves as the springboard for our own study described in section 3.

2.3 Empirical issues: How do we decide?

Some of the arguments presented in favor of gestural primacy may be extended to the pantomime-first proposal, while others need reconsidering. Starting with primatology, the relatively high degree of flexibility in the gestures of non-human apes, compared to the relative rigidity of their vocalizations was generally accepted a decade ago: "primate gestures are individually learned and flexibly produced communicative acts … vocal displays are mostly unlearned, genetically fixed, emotionally urgent, involuntary, inflexible [and] … broadcast mostly indiscriminately" (Tomasello, 2008: 54). More recent findings can rather be seen as evidence for multimodal-first. For example, wild chimpanzees seem to fall back upon the visual modality when more secrecy is needed (Hobaiter & Byrne, 2012), but use the vocal-auditory modality for attracting attention, in a selective, and (apparently) intentionally communicative manner (Schel et al., 2013). Chimpanzees and bonobos combine their calls in sequences and modify them based on the composition of the audience ("audience effects") or even on their state of knowledge, since calls are apparently used to inform naive individuals (Clay & Zuberbühler, 2014). Still, one may argue that the degree of plasticity of the bodily-visual modality is greater than that of the vocal modality, as admitted also by See (2014: 211): "the only truly salient differences between the vocalization and gestures of the chimpanzees is that the latter possess a more open-ended plasticity in production". Thus we could say that the jury is still out concerning this type of evidence.

When it comes to evidence from neuroscience, the findings are similarly rather inconclusive. On the one hand are the influential findings concerning the presence of mirror neuron systems in monkey, ape and human brains, and the fact that these systems have gradually expanded in evolution from initial action-recognition, to first allow (complex) imitation and then pantomime (Arbib, 2005; Zlatev, 2008). An argument for pantomime-first is that such cortical systems largely overlap with the "language areas" Broca (BA 44, 45) and Wernicke (BA 22, 39, 40), while monkey vocalizations are supported by the anterior cingulate cortex (ACC), closely associated with affect (Paus, 2001).

On the other hand, the human ACC apparently plays a role in speech as well, while attention-getting vocalizations in chimpanzees also activate their homolog to Broca's area (Taglialatela, Russell, Schaeffer, & Hopkins, 2011). While such evidence can be taken as supportive for multimodal-first, it does not show that there can be no dissociations between speech and gesture/pantomime, as assumed by the "single system" of McNeill (see 2.2). In fact, there is evidence that aphasics who do not also suffer from apraxia (motor impairment) retain the ability to communicate through a form of pantomime, i.e. iconic gesturing produced in the absence of speech (Kemmerer, Chandrasekaran, & Tranel, 2007: 70)

The precise relationship between speech and gestures is one of the key issues in the interdisciplinary field of gesture studies. As noted, both McNeill (1985, 2012) and Kendon (2004, 2014b) argue for a close link and a broad notion of language that includes gestures, but their theories are substantially different. McNeill's Growth Point theory presupposes that such a link is indispensable for the existence of both speech and gesture, whereas for Kendon their integration is an interactional achievement and not the outcome of a psychological mechanism. Research has shown that while gestures indeed share semantic properties with what is being said and differ across languages, speakers also use gestures to represent objects and events in ways that go beyond language, such as left-right directionality that is not encoded verbally. In sum, many results support a model of "two qualitatively different representations [which] are adjusted with respect to each other and co-evolve" (Kita & Özyurek, 2003: 30). This does not arbiter between our two models, but at least does not prejudge in favor of multimodal-first.

The developmental evidence likewise supports an analysis in terms of two interacting systems rather than a completely inseparable Speech-Gesture bond. There is indeed close interaction between gesture and speech in language development (Volterra, Caselli, Caprici, & Pizzuto, 2005) but deictic (pointing) gestures and at least some representational gestures emerge prior to speech, and they play an essential role in the development of language (Lock & Zukow-Goldring, 2010). Indeed, there is evidence for "a gradual specialization from unimodal forms of communication, less demanding in cognitive, social and semiotic terms, to multimodal patterns involving the coordination of specific gestures and vocalizations" (Murillo & Belinchón, 2012: 31). With all due precautions in extending this analysis from ontogeny to phylogeny, we can see that it is more in line with the pantomime-first model.

The final source of evidence to review – and the one most relevant for the study described in the next section – comes from semiotics: the interdisciplinary field investigating different systems of meaning-making, such as visual representations, speech and music (e.g. Sonesson, 1997; Chandler, 2007). While traditional semiotics was mostly theoretical, modern approaches of experimental semiotics (Galantucci & Garrod, 2010) and cognitive semiotics (Zlatev, 2012) are considerably more empirical. [3]

[3] Experimental semiotics "focuses on the experimental investigation of novel forms of human communication […] which people develop when they cannot use pre-established communication systems" (Galantucci & Garrod, 2011). As the participants in such studies are modern humans equipped with present-day brains and speaking present-day languages, and because the laboratories are not meant to re-create create ancestral environments, such experiments cannot directly adjudicate between competing theories of language origins. Still, they can cover one of the many inferential steps required for such far-reaching conclusions and answer more narrowly defined questions, such as what conditions (including semantic domains, signal modalities, and other parameters such as feedback) are more likely to result in communicative success.

Two recent studies (Fay, Arbib, & Garrod, 2013; Fay, Lister, Ellison, & Goldin-Meadow, 2014) used the method known as a “referential game”, asking pairs of participants to communicate to one another 24 different concepts represented by written words, divided into the categories Emotion, Action and Object. The “director” was to communicate those meanings to the “matcher” without using language by one of three means: (a) vocalization, (b) gesture consisting of silent bodily re-enactment, including facial expressions or (c) a combination of both. We can note that the authors’ use of the term “gesture” corresponds to what we would call a reduced (since the directors were sitting, and communicating simple concepts rather than events) form of pantomime, i.e. the use of whole body bodily mimesis for event representation (see Section 2.2). The results showed that in all cases matching was above chance, and that for the Emotion class, the vocalization-only group managed fairly well (ca. 70%). However, gesture/pantomime with or without vocalization had a clear advantage, leading the authors to conclude that “gesture outperforms non-linguistic vocalization because it lends itself more naturally to the production of motivated [i.e. iconic] signs” (Fay, Arbib, & Garrod, 2013: 1). Interestingly, the multimodal condition of gesture combined with vocalization did not result in greater communicative success than silent gesture in both experiments; In one case, communicating Action items in the multimodal condition was slightly but significantly less effective than silent gesture alone (Fay, Lister, Ellison, & Goldin-Meadow, 2014).

In a differently designed study aiming to compare the ability of gesture/pantomime and vocalization to express novel meanings, Brown (2012) created 39 video-clips in which an actor said a word in Japanese (a) without any gesture, (b) together with an iconic gesture or (c) together with a so-called “adaptor”, in which he touched some part of his body, without this having any connection to the meaning of the word. Participants with no knowledge of Japanese were asked to play a “foreign language guessing game” where they were shown the different video-clips together with four pictures which represented different objects, one of which was the “correct” choice. Feedback was provided after each guess. The three conditions were compared for how well participants remembered the new “meanings” after a period of time. As expected, condition (c) was the most difficult, but (somewhat surprisingly), there was no significant difference between (a) vocalization only and (b) vocalization plus pantomime.

Thus, in two very different studies in experimental semiotics, multimodal communication did not give rise to higher communicative or learning success than gesture alone (Fay et al., 2013; Fay et al., 2014) or vocalization alone (Brown, 2012). As long as the two modalities were not incongruent, there was no disadvantage for multimodality in Brown's study, though in the study of Fay et al. (2014), in one case participants performed slightly but significantly worse in the multimodal than in the silent gesture/pantomime condition. We take these findings to be rather troublesome for the multimodal-first approach.

However, there are two methodological aspects that could have contributed to the relatively low degree of success in the multimodal conditions of such experiments. The first is that the bodily-visual representations used were not full-fledged pantomimes, as participants were either sitting when communicating or else instructed to perform rather reduced manual gestures. This could in principle have prevented vocalization and gesture from naturally blending so as to aid communication effectiveness. The second point is that in these studies – as well as in many others in experimental semiotics – the task was to communicate simple lexical concepts, i.e. corresponding to individual "word-size" entities like MAN or RUN. However, from a communicative perspective, the primary unit of language is not a single lexical concept but rather an utterance, speech act, or move in a "language game" (Wittgenstein, 1953; Zlatev, 1997). Again, the combination of both modalities could possibly be exploited more effectively in a task that involved the communication of whole propositions by full-fledged pantomimes rather than isolated gestures.

In the following section, we describe a study with an adapted design to address precisely these two aspects and thus give the multimodal-first theory a more leveled competing ground with the pantomime-first theory that we endorse.

3. Pantomiming events with and without vocalization: an experimental study

3.1 Rationale and hypotheses

We designed a study in which participants had to interpret events that were communicated to them without the use of language. Four actors were video-recorded as they re-enacted/pantomimed simple transitive events selected from the matrix of 20 cartoon-like pictures shown in Figure 1. Each actor did this once with and once without vocalization. The video-recordings were shown to 44 participants who matched the events back to the corresponding drawing on the same picture matrix. We operationalized communicative success as the proportion of drawings “guessed” correctly, in line with the way in which this is usually done in the “referential games” of experimental semiotics (see Section 2.3).

This setup extends previous experimental-semiotic studies in three ways. First, the actors had to re-enact and interpret composite events rather than single lexical concepts. Second, inspired by Brown’s study, the participants responded non-verbally, that is, by selecting an image and not a written word or sentence. The avoidance of language makes the task more ecologically realistic as a model of a precursory stage to language. Finally, we made sure that the communicators used their whole bodies, and not just their hands, when they re-enacted the events, and that these performances were visible to the audience.

In line with the definitions of the multimodal-first and pantomime-first theories, the empirical findings from experimental semiotics, and the methodological issues presented in Section 2, we could formulate the following predictions, emanating from the respective theory.

- *Multimodal-first*: There will be greater communicative success in the multimodal (visual + vocal) condition. If so, previous findings supportive of the primacy of unimodal gestural communication may indeed have resulted from design limitations.
- *Pantomime-first*: There will be no difference, or greater communicative success in the visual-only condition. If so, existing support for gesture-first theories could be generalized to mimetic-pantomimic theories of language origins.

3.2 Materials

Aiming for full semantic systematicity, which is one of the hallmarks of language and which provides the basis for the ability to construct complex meanings from more simple ones (Zlatev, 2008; Dor, 2015), the experiment concerned the communication of simple transitive events, each decomposable into three elements: an Action (KISS, WAVE, SLAP, PUSH), an Agent (MAN, WOMAN, BOY, GIRL) and a Patient (MAN, WOMAN, BOY, GIRL). For each of the four Actions, five unique Agent-Patient combinations were used, resulting in 20 unique transitive events. An artist drew these as simple cartoon-like pictures, with different images of Agents and Patients in each picture. The drawings were placed pseudo-randomly on a 5x4 matrix, which was printed out on an A4 sheet of paper, shown in Figure 1.

<INSERT FIGURE 1 ABOUT HERE>

Figure 1. The 20 Agent-Action-Patient events, used by the actors in producing the pantomimes, and by the participants in "guessing" which event was being pantomimed

Two of the actions represented positive valence (KISS, WAVE) and the other two, negative valence (SLAP, PUSH). While the contrasts between the four actions were expected to be fairly easy to represent and clear to the participants, those between Agents, as well as between Patients (MAN, WOMAN, BOY, GIRL), were expected to lead to "semantic crowding", and high *confusability* within and between the Agent and Patient categories (cf. e.g. Meir et al., 2017). As explained earlier, our goal with this aspect of the design was to make the task more difficult than in earlier studies to avoid a ceiling effect.

Two female and two male amateur actors from the Spinning Globe Students' Theatrical Association at Nicolaus Copernicus University in Toruń were recruited and provided with informed consent forms, whereby they agreed to be recorded and for the videos to be used in our study. Each actor was given the matrix of event-pictures and asked to "act out", i.e. represent by means of whole-body pantomime, 16 randomly selected non-repeating pictures, in two experimental conditions:

- Visual (VIS): in this condition, the actors were instructed to be silent.

- Multimodal (MULT): the actors were instructed, and reminded, to communicatively use non-linguistic vocalizations.

In both conditions, the use of language or other conventional signs (e.g. emblematic gestures such as the OK-gesture) was forbidden. Two of the actors (one male, the other female) re-enacted the events first with vocalization, and then without it. The other two actors did this in reverse order. This resulted in eight *rounds* of re-enactments of 16 events per round and thus a stimulus set of 128 re-enactments (4 rounds x 2 conditions x 16 events = 128), see Figure 2 below.

The task was made communicative by asking the actors to pantomime the events to a “matcher” positioned directly behind the camera. The matcher was given the matrix of event-pictures and was requested to guess which item was being enacted. There were 8 matchers, so that each round started with a naive matcher. The actors were instructed to perform in such a way as to aid the matchers in “guessing”. For each event, the actor continued performing for a maximum of about 20 seconds or until the matchers indicated that they had made their choice. The actors took short breaks between the individual items as well as a longer break between the conditions. See supplementary online material for an example of an enacted event (Item 14: boy-push-man, recorded by Actor 4 in the MULT condition).

The performances were video-recorded in HD quality with a consumer-level Panasonic HC-V700 video camera mounted on a tripod. This resulted in 128 event representations (4 actors x 2 conditions x 16 items), with an average length of 14 seconds (min. 4s, max. 25s, SD=4.997). The VIS clips were longer on average (14.02 seconds) than the MULT clips (13.09), which was taken into account in the analysis (see Table 1). Note that due to the randomization, the events acted out by the four actors in the two conditions did not completely overlap. Two video files were created, each consisting of half of the 128 video-recorded event re-enactments (32 MULT, 32 VIS).

The films were converted to MP4 video format (video: 1280x720 pixels, 1631 kbps, 29 fps; audio: 140 kbps, 48kHz, stereo). In the stimulus video files, the individual clips followed one after another continuously, though each clip was preceded by a 5-second fade-in screen showing the clip number. We then presented the re-enactments to two groups of participants as shown in Figure 2.

<INSERT FIGURE 2 ABOUT HERE>

Figure 2. The composition of stimulus video files for the two groups, balancing for order of condition (MULT/VIS) in performance and actor gender. Each cell constitutes a round of 16 enactments, with a separate matcher for the production of each round.

3.3 Participants and procedure

Participants in the study were 44 volunteers, 29 female and 15 male students of English at Nicolaus Copernicus University in Toruń, all native speakers of Polish. Written consent for participation was obtained before the experiment began. The participants were divided into two equally-sized groups. Group 1 was shown stimulus file 1 and group 2 was shown stimulus file 2 (see Figure 2), making sure that each participant was exposed to only one condition (MULT or VIS) per actor. After being instructed, the participants received response sheets and picture-matrices and were given three minutes to familiarize themselves with the drawings.

The videos were displayed in a large lecture hall with a ceiling-mounted projector (Sanyo PLC WTC 500+PLC-LNS-S11) on a large projection screen (Adeo Professional 406 Alumid, 406x277 cm). We used the audio system installed in the room (PLS-700 loop amplifier, 33 in-ceiling speakers). The experimenters made sure that audio/video material was clearly audible and visible to all the participants. After each item, the playback was stopped and the participants had 15 seconds to note down on the response sheet the number of the picture that they thought corresponded to the event they saw. Effectively, this was a 20-alternative forced choice task (chance level 5%). There was a 3-minute break halfway through the study, i.e. after the second round, that is item 32.

3.4 Results

A total of 2816 (64 items x 44 participants) responses were collected, of which 15 were not identifiable and therefore excluded. The results, as well as post-study interviews with both actors and participants, showed that the task was indeed quite difficult, with participants identifying the target event correctly roughly half the time (mean = 49.5%;

min. 30%, max. 64%, SD=7.8%). As shown in Figure 3, the overall results showed that not only did the combination of pantomime and vocalizations not facilitate correct identification, but that the overall proportion of correct results was lower for MULT (mean = 46.3%, SD=10.9%) than for the VIS condition (mean = 52.6%, SD=12.6%).

We analyzed the results of the experiment using a mixed-effects logistic regression with actor (1-4), event (1-20) and participant (1-44) as random effects and condition (MULT or VIS), clip duration (4 - 25 seconds), trial number and recording session (i.e. the order in which the two conditions were recorded by each actor, see Figure 2) as fixed effects. In the analysis, the clip durations were centered at the minimum value, and the trial number was rescaled to values ranging from 0 to 1, so that the regression coefficient would give an estimate of the expected increase in correct responses from the beginning to the end of the experiment. The regression results are shown in Table 1.

Table 1. Regression table (fixed effects), main study

	estimate	standard error	*z*	*p*
intercept	-1.333	0.565	-2.357	0.018
condition VIS vs MULT	-0.249	0.086	-2.902	0.004
clip duration	0.004	0.011	0.318	0.750
trial number	2.545	0.676	3.762	0.000
recording session	0.332	0.086	3.857	0.000

Most importantly for the purpose of the current study, the condition effect was significant, with responses given to clips in the MULT condition being significantly less often interpreted correctly than those given to clips in the VIS condition. Additionally, we found significant trial and recording session effects. Respectively, these two effects indicate that the participants on average improved their responses during the course of the experiment, and that, overall, there were more correct responses to clips that were recorded during the second session than to those recorded during the first session (see Figure 2). The effect of clip duration, finally, was not significant.

<INSERT FIGURE 3 ABOUT HERE>

Figure 3. Proportions correct responses for the conditions MULT and VIS. Error bars represent 95% confidence intervals.

To analyze the source of the errors, we re-coded all the participants' responses into their underlying Agent-Action-Patient structure (e.g. drawing number 2, see Figure 1 was Agent=Man; Action=Kiss; Patient=Girl) to see to what extent the participants' responses overlapped with the target (correct) ones. Overall, we found that Action was incorrectly identified in only 13% of the responses. In comparison, the participants failed to correctly identify the Agent and Patient in 39.5% and 41.2% of their responses, respectively. This could be explained both by the greater "semantic crowding" of the Agent/Patient categories compared to Action (see 3.2) and by the fact that pantomime, which is essentially a form action re-enactment (see 2.2), is inherently more suited for the representation of actions than for example the age and gender of people.

In sum, the results of the study are much more consistent with the prediction from the pantomime-first model than that from the multimodal-first model.

3.5 Discussion

On the face of it, the results may appear counterintuitive, as it is hard to see why *more* information should lead to *less* communicative success. We examined possible alternative explanations, including some that could be due to artifacts of the design. One such was the order in which the conditions were presented to the actors, the audience or both. To recall, Actors 1 and 2 were recorded in the MULT condition before they were recorded in the VIS condition. For Actors 3 and 4 this order was reversed (see Figure 2). As shown in Figure 4, correct performance in the VIS condition was higher than in the MULT condition for Actors 1 and 2, which could indeed be a possible effect of practice, which had a significant effect on the results, as shown above. However, this was also the case for Actor 4, who should have been more experienced with the task in the MULT condition. Thus, difference in practice cannot explain the results.

<INSERT FIGURE 4 ABOUT HERE>

Figure 4. Proportions correct responses for the four actors in the two conditions. Actors 1 and 2 performed the pantomimes in the order MULT before VIS, while actors 3 and 4 did so in the reverse order.

Another possible explanation was that vocalizations were simply distractive for the audience, analogous to the way in which adaptor-gestures were in the third condition of the Brown (2012) study. In that case, removing the sound from the MULT condition should result in higher performance. We tested this additional hypothesis[4] in a follow-up study with 20 new participants (14 female, 6 male), using only the events recorded in the MULT condition for each of the four actors.[5] This time, however, playback was performed without sound, so that even though the actors did use vocalization in their pantomimes, the participants could not hear this (though they could, in principle, be influenced by the orofacial expressions generated in producing these vocalizations).

The proportion of correct responses was 45.1% (min. 34%, max. 58%, SD=6.6%), which was, in fact, somewhat lower than the proportion of correct responses to the MULT clips with sound in the main experiment. The combined results of the main experiment and the follow-up study were re-analyzed using mixed-effects logistic regression, using the same predictors as before. Table 2 shows the results.

Table 2. Regression table (fixed effects), for follow-up study with sound removed from the clips for the MULT condition

[4] It may be noted that this is not inconsistent with the pantomime-first theory as such, but it does not follow from it either.

[5] Out of the original 21 participants we excluded 1 person, who had a very low correct response rate (z-score = -2.9), suggesting a lack of motivation and semi-random answering.

	estimate	standard error	*z*	*p*
intercept	-1.356	0.461	-2.940	0.003
condition SIL vs MULT	0.068	0.107	0.640	0.522
condition SIL vs VIS	0.324	0.108	2.999	0.003
clip duration	0.001	0.009	0.105	0.916
trial number	1.953	0.550	3.516	0.000
recording session	0.381	0.074	5.115	0.000

There was no significant difference in guesser performance between the MULT clips without sound in the follow-up study and the original MULT clips with sound. As in the main study, there were significantly more correct responses to the VIS clips than to the SILENT (MULT clips with sound removed). As before, the effects of trial number and recording session were significant, but the effect of clip duration was not. Accordingly, the additional hypothesis that it was the "visual + auditory" nature of the stimulus per se that resulted in the lower performance in the MULT relative to VIS condition could be ruled out.

A third possibility was that the effect was in production rather than reception, i.e. the MULT condition was not so much harder to understand by the participants, as it was harder to produce by the actors. This possible explanation finds additional support from the post-study interviews with the actors, who indicated that it was difficult to vocalize but not use words and that having to suppress speech was distracting. It would be necessary to study the "quality" of the re-enactments in the two conditions in detail before this explanation could be corroborated.

To summarize, it is hard to pinpoint a single factor that may explain why the MULT condition gave rise to lower communicative success than silent pantomime. Most likely it is a combination of factors such as the higher iconicity of the bodily-visual representations, and the possibly distractive role of non-linguistic vocalizations. In any case, the results show that multimodality is not always beneficial for communication, and thus provide indirect support for the pantomime-first theory of language origins.

However, the post-study interviews presented a puzzle: the majority of the participants gave an affirmative answer to the question whether the vocalizations of the actors were “helpful” for guessing at the correct pictures. How can we reconcile these judgments with the lower proportion of correct responses in the MULT condition? To help answer this, we conducted an additional follow-up study. 23 new participants (15 female, 8 male) heard only the audio tracks from the 64 MULT clips and evaluated the valence of the emotion that they thought was present in that audio item. They did that by indicating their choice on a 5-point Likert scale (-2 = clearly negative, -1 = somewhat negative, 0 = neutral, 1 = somewhat positive, 2 = clearly positive).

We expected that the actors’ vocalizations would carry detectable emotional information contrasting the positive (KISS, WAVE) and negative (SLAP, PUSH) events (see Materials). Consequently, we expected the audio tracks of the clips for negative Actions to be evaluated significantly lower for the emotional score than the events with positive Actions. Thus, for each of the 64 MULT clips in this follow-up study, we computed the “emotional score” as a mean of all participants’ Likert scores for the emotion heard by them in the audio track for that clip. On average, the audio tracks from clips depicting a positive Action had an emotional score of 0.26 (*N*=31, SD=0.9), whereas in the negative group the score was -0.37 (*N*=33, SD=0.98). An independent samples *t*-test revealed this difference to be statistically significant (*N*=64, *t*(62)=2.663, *p*=0.01).

Thus, the actors vocalized in a way that was consistent with the emotional valence of the represented action. This could in principle have helped to identify the type of action that was being communicated, which in turn could be reflected in the participants’ judgments of the “helpfulness” of the vocalizations. Still, as the majority of the misidentifications concerned the “semantically crowded” and mutually confusable Agent and Patient categories, the availability of extra information on the action concerned would not have led to a higher rate of response correctness in the main study.

4. Conclusions

As we pointed out in the introduction, the age-old debate between gesture-first and speech-first theories of language origins has recently been regarded by many as resolved in a dialectic fashion: not either–or, but both! The story is simple: gestures and vocalizations, in multimodal combination, gave rise to language. The main point of the present article is that such a resolution would be indeed too simple. Not only does it hide definitional issues concerning the different senses of “multimodality”, but it disregards differences between different kinds of multimodal-fist theories (e.g. those of McNeill and Kendon), as well as the pantomimic approach that we have explored in this article. At least some of the evidence that we reviewed in Section 2 supported the pantomime-first theory, claiming that the breakthrough towards language occurred in the bodily-visual modality (due to its greater iconic potential), with vocalizations playing initially a more modest “para-gestural” (Collins, 2013) role.

Several studies in experimental semiotics were interpreted as evidence for this scenario, but we suggested that they may have set their communicative ribbons too low, as pantomime involves the communication of whole events rather than single concepts, as well as using a rather general notion of “gesture”. The study that we designed and carried out was intended to make up for this gap. We reasoned that these changes towards increased communicative realism (i.e. whole-body pantomime instead of manual-only gesture; communicating complex events rather than single concepts; using nonlinguistic rather than linguistic representations) could bring out the advantage of the combination of modalities more adequately, and thus offer support for the multimodal-first scenario.

Our findings undermined this hypothesis, as the multimodal condition was not only not better than silent pantomime under such more realistic conditions, but significantly worse. Some of the difficulty may have been on the side of production rather than reception, as modern speakers appear to find it difficult to “vocalize” without speaking, while they find it less unnatural to perform pantomime without speaking. But this latter fact itself supports relative independence between the bodily-visual and vocal-auditory modalities, rather than a single, undividable link.

Therefore, in combination with other results from evolutionary modeling (e.g. Brown, 2012), our findings support, however indirectly, a *pantomime-first* scenario in language origins, where vocalization did not play an essential part in the communication of propositional information, in particular with respect to multiple semantic roles (i.e. “who did what to whom”). On the other hand, our study showed that vocalization is clearly important for the communication of emotion (Mithen, 2005). In assuming that early hominins, who had already parted evolutionary paths with (the ancestors of) chimpanzees, first communicated with whole-body pantomime, there is no reason to suppose that this would have been fully “silent”. With time, increasing cultural complexity and a growing “symbolic landscape” (Dor, 2015) to communicate about, it is only natural that such para-gestural vocalizations would have been recruited for symbolic communication, based predominantly, though not exclusively, on conventional rather than iconic expression-content relations (Brown, 2012).

In other words, the multimodality of present-day linguistic communication, with a privileged role for speech (in hearing individuals) is most likely a consequence of the long process of language evolution, rather than a feature that was present in its onset.

Acknowledgements

We are grateful to the following people for their help at different stages of the project: Marta Sibierska, Aleksandra Borowska, Jacek Stopa, Jacek Wilczyński, Aleksandra Wojtasik, Marta Boruta, Monika Boruta, Marek Placiński, Krzysztof Kaszyński, Natalia Pałka, Katarzyna Rogalska, Maciej Pokornowski, Anna Jaskólska, Gelena Mikitało, Peer Christensen, Jarosław Hetman, Waldemar Skrzypczak. We also wish to express our thanks to Michael Tomasello and Manuel Bohn for suggestions and to three anonymous reviewers for their constructive critical comments, which led to a thorough revision of the first version of this text.

References

Aboitiz, F. (2012). Gestures, vocalizations, and memory in language origins. *Front. Evol. Neurosci*, *4*(2).

Andrén, M. (2010). *Children's gestures between 18 and 30 months*. Lund: Media Tryck.

Arbib, M. (2005). From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. *Behavioral and brain sciences*, 28, 105-168.

Arbib, M. (2012). *How the brain got language*. Oxford: Oxford University Press.

Armstrong, D. F., Stokoe, W. C., & Wilcox, S. E. (1995). *Gesture and the nature of language*. Cambridge: Cambridge University Press.

Armstrong, D. F., & Wilcox, S. E. (2007). *The gestural origin of language*. Oxford: Oxford University Press.

Brown, J. E. (2012). The evolution of symbolic communication: An embodied perspective. PhD Thesis, University of Edinburgh.

Burling, R. (2005). *The talking ape: How language evolved*. Oxford: Oxford University Press.

Call, J., & Tomasello, M. (2007). *The gestural communication of apes and monkeys*. London: Lawrence Erlbaum.

Chandler, D. (2007). *Semiotics: The Basics, Second edition*. London: Routledge.

Clay, Z. & Zuberbühler, K. (2014). Vocal communication and social awareness in chimpanzees and bonobos: insights from studies of vocal communication. In D. Dor, C. Knight, & J. Lewis (Eds.), *The Social Origins of Language* (pp. 141-156). Oxford: Oxford University Press.

Collins, C. (2013). *Paleopoetics: The evolution of the literary imagination*. New York: Columbia University Press.

Corballis, M. C. (2003). From mouth to hand: gesture, speech, and the evolution of right-handedness. *Behavioral and Brain Sciences* 26 (2), 199–208.

Donald, M. (1991*). Origins of the modern mind: Three stages in the evolution of human culture*. Cambridge, MA: Harvard University Press.

Dor, D. (2015). *The instruction of imagination: Language as a social communication technology*. Oxford: Oxford University Press.

Dunbar, R. (1996). *Grooming, gossip and the evolution of language*. London: Faber & Faber.

Fay, N., Arbib, M., & Garrod, D. (2013). How to bootstrapp a human communication system. *Cognitive Science*, 37, 1356-1367.

Fay, N., Lister, C. J., Ellison, T. M. & Goldin-Meadow, S. (2014). Creating a communication system from scratch: gesture beats vocalization hands down. *Frontiers in Psychology* 5 (354).

Fitch, W. T. (2010). *The evolution of language*. Cambidge: Cambridge University Press.

Galantucci, B., & Garrod, S. (2010). Experimental semiotics: A new approach for studying the emergence and the evolution of human communication. *Interaction Studies*, 11, 1–13.

Galantucci, B., & Garrod, S. (2011). Experimental semiotics: a review. *Frontiers in human neuroscience*, *5*(11).

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

Hewes, G. W. (1973). Primate communication and the gestural origin of language. *Current Anthropology* 14.1/2, 5–24.

Hobaiter, C. & Byrne, R. W. (2012). Gesture use in Consortship: wild chimpanzees' use of gesture for an 'evolutionarily urgent' purpose. In S. Pika & K. Liebal (Eds.), *Developments in primate gesture research* (pp.127-144). Amsterdam: John Benjamins Press.

Johansson, S. (2005). *Origins of language: Constraints on hypotheses*. Amsterdam: Benjamins.

Kemmerer, D., Chandrasekaran, B., & Tranel, D. (2007). A case of impaired verbalization but preserved gesticulation of motion events. *Cognitive neuropsychology*, 24(1), 70-114.

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

Kendon, A. (2012). Language and kinesic complexity. *Gesture*, 12, 308-326.

Kendon, A. (2014a). Semiotic diversity in utterance production and the concept pf "language". *Phil. Trans. R. Soc. B,* 369(1651), 20130293.

Kendon, A. (2014b). The 'poly-modalic' nature of utterances and its implication. In D. Dor, C. Knight, & D. Lewis (Eds.), *The social origins of language* (pp. 67-76). Oxford: OUP.

Kita, S., & Özyurek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48, 16–32.

Landau, B., & Gleitman, L. R. (2009). *Language and experience: Evidence from the blind child.* Cambridge, MA: Harvard University Press.

Levinson, S. C. & Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society of London. Series B,* Biological Sciences, 369, 201303302.

Lock, A., & Zukow-Goldring, P. (2010). Preverbal communication. *The Wiley-Blackwell Handbook of Infant Development* (pp. 394-425). Oxford: Willey-Blackwell.

MacNeilage, P. F. (2008). *The origin of speech.* Oxford: Oxford University Press.

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review,* 92, 350-371.

McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago Press.

McNeill, D. (2012). *How language began: Gesture and speech in human evolution.* Cambridge: Cambridge University Press.

Meir, I., Aronoff, M., Börstell, C., Hwang, S., Ilkbasaran, D., Kastner, I., Lepic, R., Ben Basat, A. L., Padden, C. & Sandler, W. 2017. The effect of being human and the basis of grammatical word order: Insights from novel communication systems and young sign languages. *Cognition,* 158, 189-207.

Mithen, S. (2005). *The singing Neanderthals: the origins of music, language, mind and body*. London: Weidenfeld and Nicholson.

Murillo, E., & Belinchón, M. (2012). Gestural-vocal coordination. *Gesture,* 12, 16-39.

Paus, T. (2001). Primate anterior cingulate cortex: where motor control, drive and cognition interface. *Nature Reviews Neuroscience*, 2(6), 417-424.

Pika, S. (2008). What is the nature of gestural communication in great apes? *The shared mind: Perspectives on intersubjectivity* (pp. 165-186). Amsterdam: Benjamins.

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(10), e76674.

See, A. (2014). Reevaluating chimpanzee vocal signals: Toward a multimodal account of the origins of human communication. In M. Pina & N. Gontier (Eds.), *The evolution of social communication in primates* (pp. 195-215). New York: Springer.

Sonesson, G. (1997). The ecological foundations of iconicity. *Semiotics around the world: Synthesis in diversity* (pp. 739–742). Berlin: Mouton de Gruyter.

Stokoe, W. C. (1991). Semantic phonology. *Sign Language Studies* 71, 107–114.

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.

Tomasello, M. (2008). *The origins of human communication*. Cambridge, MA: MIT Press.

Vigliocco, G., Perniss, P., & Vinson, D. (2014). Language as a multimodal phenomenon: implications for language learning, processing and evolution. *Phil. Trans. R. Soc. B*, 369(1651), 20130292.

Volterra, V., Caselli, M., Caprici, O., & Pizzuto, E. (2005). Gesture and the emergence and development of language. *Beyond nature-nurture: Essays in honor of Elisabeth Bates* (pp. 3-40). Mahwah, NJ: Lawrence Erlbaum.

Wacewicz, S., Zywiczynski, P., & Orzechowski, S. (2016). Visible movements of the orofacial area. Evidence for gestural or multimodal theories of language evolution? *Gesture* 15(2), 251–284

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Blackwell.

Zlatev, J. (1997). *Situated embodiment: Studies in the emergence of spatial meaning*: Stockholm: Gotab.

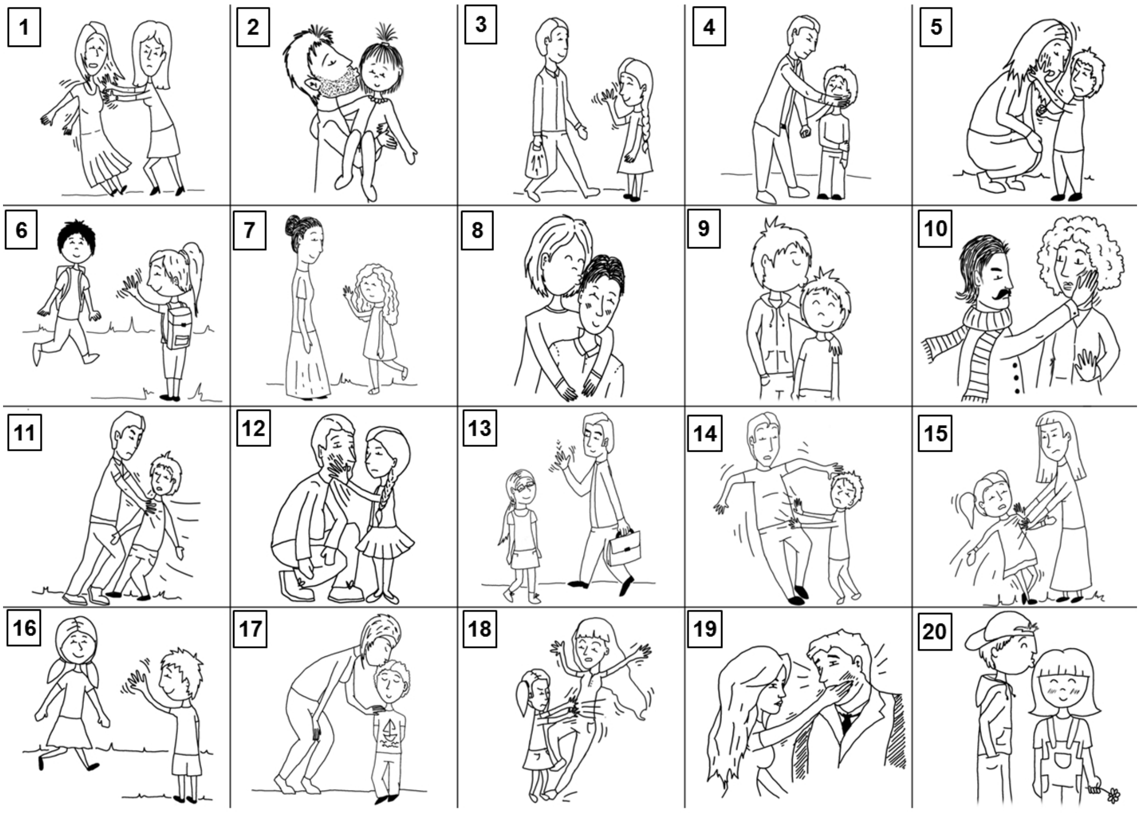
Zlatev, J. (2008). From proto-mimesis to language: Evidence from primatology and social neuroscience. *Journal of Physiology – Paris*, 102, 137-152.

Zlatev, J. (2012). Cognitive semiotics: An emerging field for the transdisciplinary study of meaning. *Public Journal of Semiotics,* 4, 2-24.

Zlatev, J. (2014). Bodily mimesis and the transition to speech. In M. Pina & N. Gontier (Eds.), *The evolution of social communication in primates* (pp. 165-178). New York: Springer.

Zywiczynski, P., Wacewicz, S. (2015). *Ewolucja języka. W stronę hipotez gesturalnych*. Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.

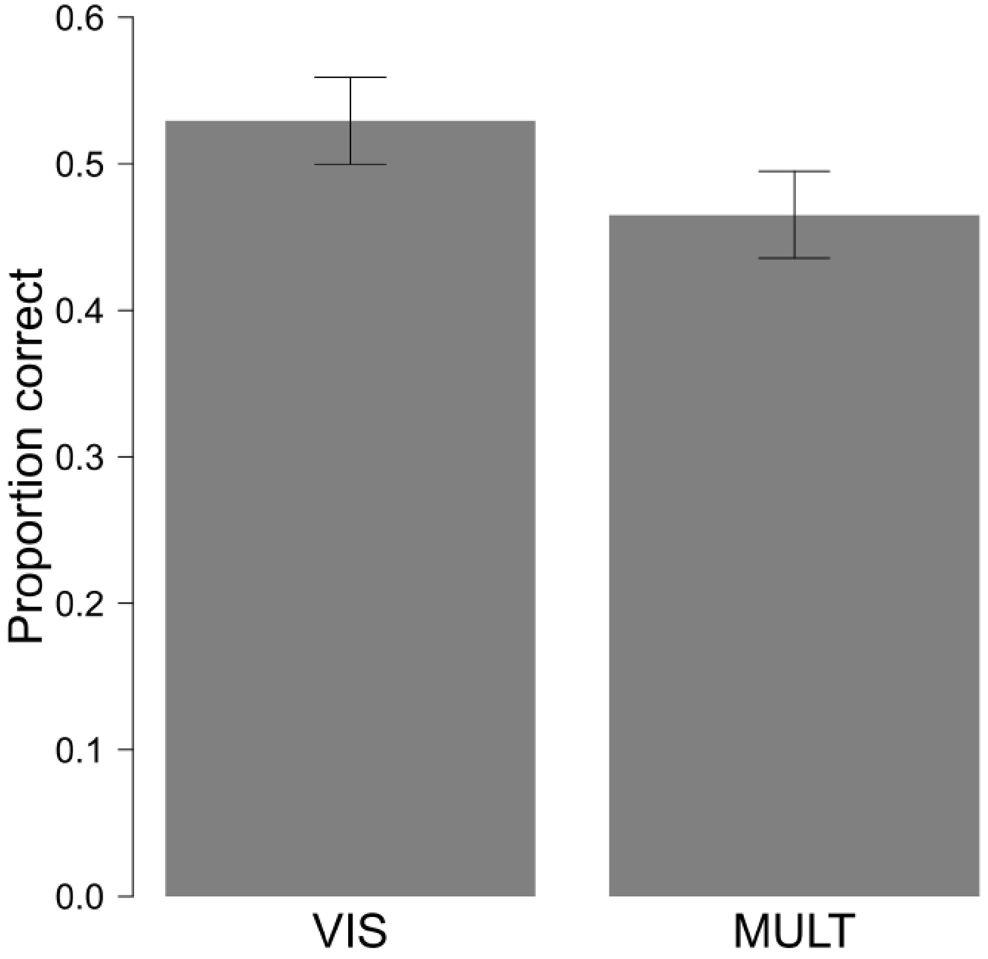
Zywiczynski, P., Wacewicz, S., & Sibierska, M. (2016). Defining pantomime for language evolution research. Topoi. *An International Review of Philosophy* [in press]. doi:10.1007/s11245-016-9425-9

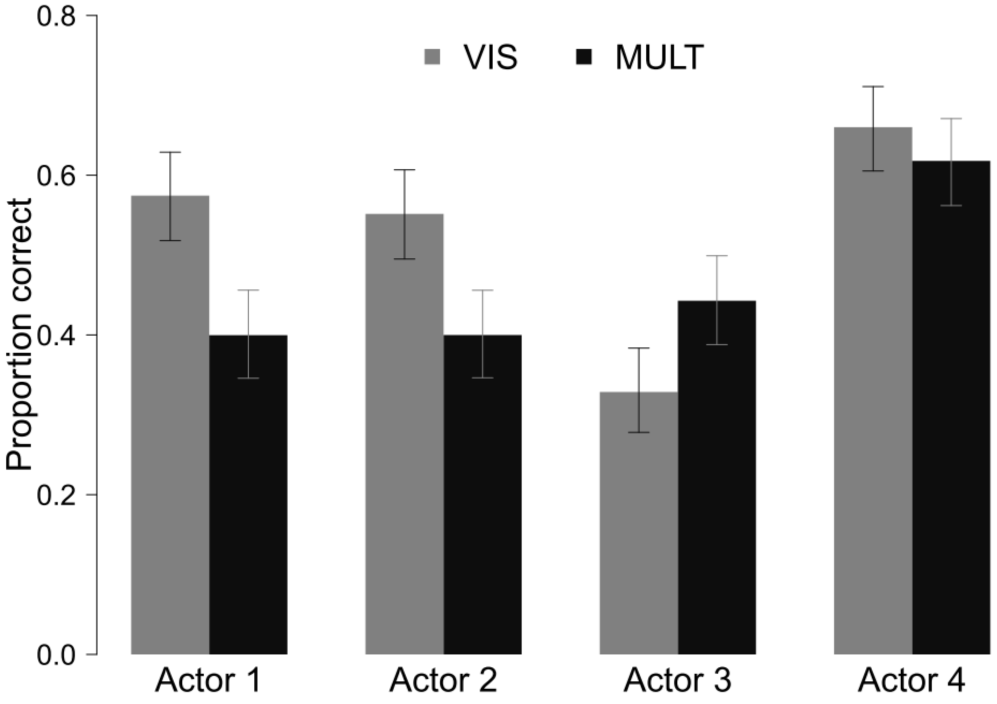


ACTOR ID	Condition (Round)	Condition (Round)
Actor 1 (male)	▨ **MULT (1)**	**VIS (1)**
Actor 2 (female)	**MULT (2)**	▨ **VIS (2)**
Actor 3 (male)	▨ **VIS (3)**	**MULT (3)**
Actor 4 (female)	**VIS (4)**	▨ **MULT (4)**

Stimulus videos

□ **Group 1** ▨ **Group 2**


Proportion correct
0.6
0.5
0.4
0.3
0.2
0.1
0.0
VIS
MULT


■ VIS
■ MULT
Proportion correct
0.0
0.2
0.4
0.6
0.8
Actor 1
Actor 2
Actor 3
Actor 4