

DAVID MCNEILL, FRANCIS QUEK,
KARL-ERIK MCCULLOUGH, SUSAN DUNCAN,
ROBERT BRYLL, XIN-FENG MA,
AND RASHID ANSARI

DYNAMIC IMAGERY IN SPEECH AND GESTURE

1. INTRODUCTION

Someone begins to describe an event and almost immediately her hands start to fly. The movements seem involuntary and indeed unconscious, yet they take place vigorously and abundantly. Why is this happening? Whatever the reason, our person is not alone. Popular beliefs notwithstanding, every culture produces gestures. Gesturing is a phenomenon that passes almost without notice but it is omnipresent. If you watch someone speaking, in almost any language, and under nearly all circumstances, you will see what appears to be a compulsion to move the head, hands and arms in conjunction with speech. Speech, we know, is the actuality of language. But what are these gestures? They are not compensations for missing words or inarticulate speech – if anything, gestures are positively related to fluency and complexity of speech – the more articulate the speech, the more gesture.

Such gestures have been called gesticulations by Kendon (1988). They need to be distinguished from culturally codified gestures such as the ‘OK’ sign or waving goodbye, known as emblems. Gesticulations are characterized by an obligatory accompaniment of speech, a lack of language-defining properties, idiosyncratic form-meaning pairings, and a precise synchronization of meaning presentations in gestures with co-expressive speech segments. This stable of characteristics shows that gestures and speech are systematically organized in relation to one another. The gestures are meaningful. They form meaningful, non-redundant combinations with the speech segments with which they synchronize, despite the fact that they are idiosyncratic and ephemeral. Language actually includes motion of the limbs along with the usual linguistic components - words, phrases, etc.

The argument of this paper is that these gestures are part of our thinking process, processes that take place automatically as the mind engages itself with language (cf. McNeill, 1992; McNeill & Duncan, 2000). Not any thinking process but thinking that emerges with and is evoked through language. Dan Slobin in 1987 introduced a technical term for this cognitive mode – “thinking for speaking”. Gestures are irresistible because thinking for speaking is irresistible.

Our goal is to elaborate this argument and explore its plausibility and depth, making use of new techniques of motion analysis.

To illustrate the kind of everyday gesture that we mean, consider the following example. A speaker recounting a cartoon story lifts her hand upward with the meaning that a character is climbing up. Her gesture is not a codified cultural gesture such as the 'OK' sign. It is imagery created on the fly at the moment of speaking and is part of the speaker's meaning at this moment. The rising hand embodies upwardness and shares this meaning with the most co-expressive parts of the accompanying utterance, "[and he climbs **up the pipe**]."¹ By 'image' we mean a semiotic object with the following crucial property: the meanings of the parts are determined by the meaning of the whole. This is called the global property (McNeill, 1992). The hand is a character, it is not a hand, because it is part of a gesture whose overall meaning is a character rising upward. The global property is the opposite of the compositional property found in linguistic objects. A gesture is top-down while a sentence is bottom-up. A sentence is composed out of independently meaningful parts, words or morphemes. The meaning of the word 'hand' does not change according to the overall meaning of the sentence. The meaning of the hand in a gesture does change depending on the overall meaning of the gesture and can mean a hand or something else altogether in another gesture. The temporal alignment of imagery and speech is crucial, for it suggests that upness is being thought of in *two cognitive frameworks simultaneously*. One framework is instantaneous, visuospatial and actional, and is characterized by an imagistic mode of cognition visible in the gesture. The other is linear, segmented and language-like, and is actualized in the speech. These modes are *simultaneous*; it is not that one leads to the other. The image of upness was co-present with the segments "up" and "the pipe." Thinking takes place in both frameworks at once. Simultaneity is important because it suggests constraints on the mechanisms of speaking and thinking beyond the obvious requirements of communication.

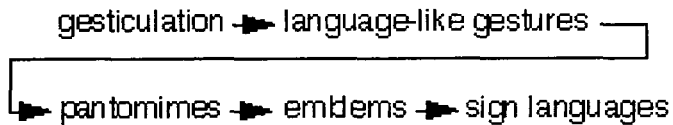


Figure. Kendon's continuum of gestures (reproduced from *Hand and Mind*, McNeill, 1992).

A key question is the source of constraint on gesture form and timing. Kendon (1988) described a taxonomy of gesture summarized and organized as a continuum in Figure 1. At one end of the continuum is *gesticulation*, which describes the free-form gesturing which typically accompany verbal discourse. At the other end of this continuum are the sign languages of the deaf (such as American Sign Language, or ASL), which are characterized by complete lexical and grammatical specification. In between, we have 'language-like gestures,' in which the speaker inserts a gesture in the place of a syntactic unit during speech; 'pantomimes,' in which the gesturer creates an iconic and motion facsimile of the referent; and 'emblems,' which are

codified gestural expressions that are not governed by any formal grammar (such as OK or the 'thumbs-up' gesture to signify 'all is well'). This paper focuses on the 'gesticulation' end of this continuum.

We learn from gesture that when the mind engages language it also engages instantaneous imagery as an integral part. Gesture and speech are not subservient to one another. One is not an afterthought to enrich or augment the other. Instead, they proceed together from the same 'idea units'. Contrary to a text-based tradition that has dominated linguistics and psycholinguistics from their historical beginnings, language contains analogic, imagistic elements that are as defining of language capacity as the time-honored analytic elements of words, phrases, clauses, sentences, and texts. The postulate that language and imagery are inseparable explains the power of gestures to open a window onto the mind and to reveal cross-linguistic differences in thinking for speaking. This is an insight that cannot be deduced from text-based data and requires theoretical concepts of language in new domains.

With gesture, we are invited inside the mental life of another person, not only because gesture offers 'evidence' of mind, but because the gesture, as such, is one of the very aspects of the mind. Some imagery is purely mental and without any apparent bodily enactment, but with gesture the image receives a material existence and gains thereby in actuality as well as observability. To put these points together – that language contains imagery as an integral part, and that gesture is the most developed form of it – we say that gesture is part of the actualization of language, along with speech itself, and is no less a crucial part of this actualization.

We seek an understanding of gesture and speech in natural human conversation. We believe that an understanding of the constants and principles of speech-gesture cohesion is essential to its application in human computer interaction involving both modalities. Several observations are necessary to frame the premise of such conversational interaction. While gesture may be extended to include head and eye gestures, facial gestures and body motion, we shall explore here only the relationship of hand gestures and speech. Conversational gestures differ from manipulative movement in several significant ways. First, because the intent of the latter is for manipulation, there is no guarantee that the salient features of the hands are visible. Second, the dynamics of hand movement in manipulative gestures differ significantly from conversational gestures. Third, manipulative gestures may typically be aided by tactile and force feedback from the object (virtual or real) being manipulated, while conversational gestures are typically performed without such constraints.

We shall present an underlying psycholinguistic model by which gesture and speech entities may be integrated, describe our research method that encompasses processing of the video data and psycholinguistic analysis of the underlying discourse, and present the results of our analysis. Our purpose, in this paper, is to motivate a perspective of conversation interaction, and to show that the cues for such interaction are accessible in video data. In this study, our data comes from a single video camera, and we consider only gestural motions in the camera's image plane.

2. PSYCHOLINGUISTIC BASIS

Gesture and speech clearly belong to different modalities of expression but they are linked on several levels and work together to present the same semantic idea units. The two modalities are not redundant; they are 'co-expressive,' meaning that they arise from a shared semantic source but are able to express different aspects of it, overlapping this source in their own ways. A simple example will illustrate. In the living space text we present below, the speaker describes at one point entering a house with the clause, "when you open the doors." At the same time she performs a two-handed anti-symmetric gesture in which her hands, upright and palms facing forward, move left to right several times. Gesture and speech arise from the same semantic source but are non-redundant; each modality expresses its own part of the shared constellation. Speech describes an action performed in relation to it, gesture shows the shape and extent of the doors and that there are two of them rather than one; thus speech and gesture are co-expressive. Since gesture and speech proceed from the same semantic source, one might expect that the semantic structure of the resulting discourse to be accessible through both the gestural and speech channels.

2.1. *The Catchment Concept*

The 'catchment' concept provides a locus along which gestural entities may be viewed to provide access to the discourse structure. A catchment is a term for discourse units inferred on the basis of gesture information. Catchments are recognized when gesture features recur in at least two (not necessarily consecutive) gestures. The logic is that imagery generates the gesture features; recurrent imagery suggests a common discourse theme. A catchment is a kind of thread of consistent visuospatial imagery running through a discourse segment. The catchment is a kind of thread of visuospatial imagery through a discourse that reveals the separate parts cohering into larger discourse units. By discovering a given speaker's catchments, we can see what for this speaker goes together into larger discourse units – what meanings are seen as similar or related and grouped together, and what meanings are isolated and thus seen by the speaker as having distinct or less related meanings. Consider one of the most basic gesture features, handedness.

Gestures can be made with one hand (1H) or two (2H); if 1H, they can be made with the left hand (LH) or the right (RH); if 2H, the hands can move and/or be positioned in mirror images or with one hand taking an active role and the other a more passive 'platform' role. Noting groups of gestures that have the same values of handedness can identify catchments. We can add other features such as shape, location in space, and trajectory (curved, straight, spiral, etc.), and consider all of these as also defining possible catchments. A given catchment could, for example, be defined by the recurrent use of the same trajectory and space with variations of hand shapes. This would suggest a larger discourse unit within which meanings are contrasted. Individuals differ in how they link up the world into related and unrelated components, and catchments give us a way of detecting these individual characteristics or cognitive styles.

3. EXPERIMENTAL METHOD

Hand gestures are seen in abundance when people describe spatially organized information. In our gesture and speech elicitation experiment, subjects are asked to describe their living quarters to an interlocutor. This conversation is recorded on a Hi-8 tape. Figure 2 is a frame from the experimental sequence that is presented here. Two independent sets of analyses are performed on the video and audio data. The first set entails the processing of the video data to obtain the motion traces of both of



Figure 2. A frame of the video data collected in our gesture elicitation experiment.

the subject's hands. The synchronized audio data are also analyzed to extract the fundamental frequency signal and speech power amplitude (in terms of the RMS value of the audio signal). The second set of analyses entails expert transcription of the speech and gesture data. This transcription is done by carefully transcribing and analyzing the Hi-8 video tape using a frame-accurate video player to correlate the speech with the gestural entities. We also perform a higher-level analysis using the transcribed text alone. Finally, the results of the psycholinguistic analyses are compared against the features computed in the video and audio data. The purpose of this comparison is to identify the cues accessible in the gestural and audio data that correlate well with the expert psycholinguistic analysis. We shall discuss each step in turn.

3.1. Extraction of Hand Motion Traces for Monocular Video

The overarching goal of our technical work is to build data and computational bridges among the disciplines represented in this paper – linguistics, psycholinguistics, and machine vision. The research proceeds along two paths toward two sets of complementary objectives. First, we research and develop the enabling technology for providing computation and access to the audio/video data and the associated derived information that support the needs of the science project. This includes the identification of the primitives that have to be computed, and the

Multimodality in Language and Speech Systems

Granström, B.; House, D.; Karlsson, I. (Eds.)

2002, X, 243 p., Hardcover

ISBN: 978-1-4020-0635-7