From actions to events

Communicating through language and gesture

James Pustejovsky
University Waltham

In this paper, I argue that an important component of the language-ready brain is the ability to recognize and conceptualize events. By 'event', I mean any situation or activity in the world or our mental life, that we find salient enough to individuate as a thought or word. While this may sound either trivial or non-unique to humans, I hope to show that abstracting away events and their participants from the embodied flow of experience is a characteristic unique to humans. This ability is enabled, I will argue, by two critical competencies that act as scaffolds for language-ready thought in the prehuman brain. The first, as argued by Arbib (2006, 2012, 2016) and others, is a sophisticated system of gesture production and understanding in prehumans, which provided a template for schema-like sequencing and slot-filling of information units. The second involves the integration of multiple modalities of expression in the communicative act, in particular, the alignment of co-gestural speech and co-speech gesture. With such computational facilities, action-based gestures can be abstracted away from their associated objects and become full event representations. This view supports the MSH argument for the emergence of more complex linguistic expressions from initially holophrastic units. In particular, actions can be thought of as protoverbs, which through this process are abstracted to full event descriptions, i.e., verbs.

Keywords: gesture, language, action, affordances, events

1. Introduction

This paper addresses the general question of how language emerged from our prehuman ancestors, by focusing on the specific conceptual difference between actions and events: recognizing them, representing them, and communicating about them. This difference is neither trivial nor obvious. In articulating this distinction, however, it will become clear how important it is to distinguish these two concepts, particularly for the question of the emergence of language and the constitution of the language-ready brain.

Interaction Studies 19:1-2 (2018), **289–317**. DOI 10.1075/is.17045.pus
ISSN 1572–0373 / E-ISSN 1572–0381

The view that language initially evolved from the use of gesture in the common ancestors of primates and humans is not new, and has been argued from diverse communities, e.g., Armstrong et al. (1995), Arbib & Rizzolatti (1997), Rizzolatti and Arbib (1998), Corballis (2003). Specific to our concerns, Rizzolatti and Arbib (1998) and Arbib (2006, 2012), introduce and elaborate the *Mirror System Hypothesis*, which argues that language was made possible because of the mutual understanding (between conspecifics) for grasping activities, enabled by mirror neurons and systems "beyond the mirror" with which they are integrated (cf. Stout and Hecht, 2017). This ability was the scaffold with which the language-ready brain emerged in humans, to then enable a cultural (and, probably to a lesser extent, Baldwinian) evolution of language. Consider the hypothesized seven stages of language evolution, proposed in Arbib (2002).

(1) S1. Grasping.
S2. A mirror system for grasping shared with common ancestor of human and monkey.
S3. A simple imitation system for object-directed grasping, shared with common ancestor of human and chimpanzee.
S4. A complex imitation system for grasping (with variations in actions from repertoire).
S5. Proto-sign, a manual-based communication system, breaking through fixed repertoire of primate vocalizations to yield open repertoire.
S6. Proto-speech, resulting from ability to control mechanisms evolved for proto-sign coming to control vocal apparatus with increasing flexibility.
S7. Language, the change from action-object frames to verb-argument structures.

The operational characteristics for stages S5 and S6 have been researched considerably, and a clear distinction between them has emerged: manual dexterity appears to goes back to LCA-m, at least; its use for a small gestural repertoire is established for LCA-c, with an ability to generate a large open proto-lexicon of conventionalized proto-signs (Aboitiz, 2013, Fogassi et al., 2013, Arbib, 2013). Yet it is less clear how S6 would develop from S5. Proto-speech demands a comparable 'vocal dexterity' so that vocal gestures can play a comparable role to that of proto-signs. Further, it is not entirely clear what is entailed by the transition to S7, with the emergence of verb-argument structure and full compositional linguistic abilities, although Arbib's (2012) suggestion involving fractionation of holophrases is a start. I present two arguments in general support of this trajectory, focusing on S6 and the transition to S7. First I discuss the ways in which different gestures require quite distinct computational strategies, particularly when aligned with co-gestural speech (S5–S6). I then outline the requirements that inhere for S7 to emerge,

where action-denoting proto-verbs as used in S6 are further abstracted and able to describe events with variable participants.

Pre-humans could perform many of the same behaviors that humans can, including gesturing, referencing, encoding individuals and object classes, and action classification. Action gestures in pre-humans, it is argued, denoted 'action-object' pairs, but not events (Glenberg & Gallese, 2012). I claim that the differential between humans and pre-humans is the ability to abstract away from the situated context and arrive at what is effectively a verb. That is, the encoding of events with the underpsecification of the participants of the event is the functional equivalent of a verb. Pre-humans had what we can call a proto-verb. Along the lines of Arbib's (2002) notion of proto-words, a proto-verb is an encoding of an action-object pair (or simply, action).

I further argue that the exploitation of more than one modality greatly facilitates the *unanchoring* of reference (and meaning more generally) out of the common ground. Following Volterra et al (2005), Goldin-Meadow and Alibali (2013), and Gillespie-Lynch et al (2014), while gesture provides the scaffolding for language, there is growing evidence that it is the multimodal nature of communicative interaction that is crucial for the richness of human linguistic behavior. From co-situated embodied meaning, the use of multimodal signals, as expressed through co-speech gestures or co-gestural speech, enables the agent to align the co-situated references to multiple modal expressions. Co-gestural speech is important for enabling language meaning in the absence of a common ground.

The outline of the paper is as follows. First, we identify what elements of the domain of discourse are available to talk about, when two embodied conspecifics are interacting to communicate with one another. Then we we review what communicative acts there are, and which acts are expressible with the modalities of speech and gesture. In order to better understand the presuppositions of stages S5 and S6, we examine what gestures and co-gestural speech are able to denote within a shared context. Next, we argue that, while gestures are able to denote actions or action-object pairings, they are unable to describe events. This is a S7 capability, which emerges when an expressive gestural system is aligned with co-gestural speech, when multimodality allows for both embodied reference and subsequent "displaced meaning" with no common ground, where the reference to a spoken token is not present in the context.

2. Communication with a common ground

In this section, I discuss the situational and cognitive factors that determine the range of what can be communicated between two individuals, primate or human. Two conspecifics who share an experience, such as witnessing a natural event,

hearing a clap of thunder, or feeling the earth tremor, are jointly co-perceiving an event. Hence, they are *co-situated* and *co-perceptive*. If these two beings are communicating in order to carry out a shared task, such as building a structure, moving objects, or clearing a space, then they can be considered "agents", who are not only coperceiving the present situation and subsequent situations as they change, but are also acting, together or individually, as a result of communicative interactions. Hence, what is being shared in the latter is considerably richer and more complex in character. Namely, there is agreement, acceptance, or recognition of a common goal between the agents, what can be called *co-intent*. These combined factors constitute the major aspects of what has been called *common ground*: namely, co-situatedness, co-perception, and co-intent. The theory of common ground has a rich and diverse literature concerning what is shared or presupposed in human communication (Clark et al., 1991; Gilbert, 1992; Grice, 1981; Stalnaker, 2002; Asher, 1998; Tomasello and Carpenter, 2007). When engaged in accomplishing a task jointly, agents share one additional anchoring strategy that greatly enhances the expressiveness of common ground: namely, the ability to *co-attend*. Because of the inherently directed nature of attention and co-attention, we will speak of *shared situated references* in the discussion that follows. This ability will emerge as central to determining the meanings of communicative acts produced by participants in shared events.

With the presence of a common ground during shared experiences, embodied communication assumes that conspecifics have an ability to understand one another in a shared context, through the use of co-situational and co-perceptual anchors, along with a means for identifying such anchors, such as gesture, gaze, intonation, and language. This was clearly the preconditions for any early communicative interaction in pre-humans. By studying the constitution and configuration of this common ground in embodied communication, we can better understand the emergence of *displaced reference* in communicative acts, where there is no common ground. For humans, the absence of true common ground is the norm in most communicative interactions, and words and phrases in language act as proxies to objects and events that are not present in the shared context.

Clark et al. (1983) make clear the role that common ground plays in determining the reference of demonstrative expressions in dialogue. The thesis of *composite signals* (Engle and Clark, 1995, Clark, 1996) argues that speech and gesture are not separate "channels" of communication that are integrated, but are signaling strategies that are created as a composite in the act.

Events as we experience them are distinct from the way we refer to them with language. The mechanisms in language allow us to package, quantify, measure, and order our experiences, creating rich conceptual reifications and semantic differentiations. The surface realization of this ability is mostly manifest through our linguistic utterances, but is also witnessed through gestures. By examining

the nature of the common ground assumed in communication, we can study the conceptual expressiveness of these systems.

The inventory of communicative acts employed in human interactions is small, considering the range of ideas communicated by speakers. It is generally agreed that all communicative acts (*C*) fall into one of the basic types below (Austin, 1962, Searle, 1969, Bach, 2003).

(2) COMMUNICATIVE ACTS:
   a. Informative (telling, including consent and dissent)
   b. Interrogative (requesting)
   c. Imperative (command)
   d. Promising (obliging)
   e. Warning (cautioning)
   f. Inviting
   g. Greeting

These can be classified into two categories: *atomic* (warn, invite, greet); and *complex* (inform, question, command, promise). Atomic acts are directly interpretable and reference the agents only, while complex acts are operations over expressions, which are then interpretable. For example, I can greet you, and that is directly interpreted by you as 'greeting', and so forth. A question, however, is an interrogation regarding some state of affairs or event, which you must interpret, independent of having interpreted my utterance as a question. We will consider a communicative act, C_a, performed by an agent, *a*, to be a tuple of expressions from the diverse modalities available to an agent, involved in conveying information to another agent. For our present discussion, let us restrict this to the modalities of a linguistic utterance, *S* (speech), and a gesture, *G*. There are three possible configurations in performing a *C*:

(3) a. $C_a = (G)$
   b. $C_a = (S)$
   c. $C_a = (S, G)$

For each of the speech act configurations in (3), we wish to determine which of the communicative acts in (2) are expressible. For example, assuming a common ground as defined above, which acts are expressible through gesture alone, through language alone, and through the pairing of gesture and language? Since atomic *C*s seldom involve situated reference other than the indexicality of the agents involved (saying hello or goodbye), we will not discuss them here, but rather focus on complex *C*s.

We will introduce a means of keeping track of what common ground parameters are available to the agents involved in an interaction, whether communicating or merely experiencing an event. This will be called a *common ground structure(CGS)*, and we adopt the convention from Discourse Representation

Theory (DRT) (Kamp and Reyle, 1993) and SDRT (Asher and Lascarides, 2003, 2008) for wrapping the content of such a data structure in a container using a box notation (an alternative proposal would be to use record types, cf. Cooper and Ginzburg, 2015). A CGS captures the parameters that are relevant to a communicative act, as experienced by the embodied participants. We can think of it as a context container, which makes explicit what is available in the shared domain of communication for reference or presupposition. Inside this container, any form of a communicative act, *C*, be it gesture, an utterance, a co-speech gesture, or a co-gestural utterance, can be placed. We will identify four common ground parameters:

(4) a. **A**: The conspecific agents engaged in communication;
b. **B**: The shared belief space;
c. **P**: The objects and relations that are jointly perceived in the environment;
d. $\mathcal{E}$: The embedding space that both conspecifics embody in the communication.

This can be represented formally as in (5), where an agent, a_i, makes a communicative act through gesture, *G*, in a common ground consisting of the parameters specified above.

(5) a.

A:a_1, a_2	**B**:Δ	**P**:b	$\mathcal{E}$
G_{a_1}			

b.

A:a_1, a_2	**B**:Δ	**P**:b	$\mathcal{E}$
S_{a_1} = "You$_{a_2}$ see it$_b$"			

For example, (5a) specifies that two conspecific agents, a_1 and a_2, co-inhabiting an embedding space, $\mathcal{E}$, within which the experience is embodied, share a set of beliefs, Δ, where they can both see the object, *b*. Given this representation, the gesture is now situated to refer to objects and knowledge within the common ground, CGS. In (5b), the linguistic expression, S_{a1}, is grounded relative to the parameters of common ground, where the indexical *you* will denote the conspecific agent, a_2, and the pronoun *it* will denote the object, *b*.

3. The structure of actions and events

In this section, we address the question of what distinguishes actions from events, conceptually. The goal is to articulate the computational presuppositions behind Bickerton's (1990) notion of protolanguage and Arbib's (2012) theory of

holophrases. We start from Arbib's (2006, 2008) argument that such an artifact emerges from the ability to imitate an inventory of practical actions, when coupled with speech tokens constituting alignments to these actions. Before answering these questions, it will be important to distinguish what kinds of events there are, as packaged by language. The position taken here, consistent with the notion of protolanguage emerging from gestural abilities, is that the way in which humans carve up the flow of experience – Whitehead's (1919) "experiential duration" – is more expressive than prehuman cognitive abilities in one major respect: humans distinguish between "actions with objects" and "open relations", that are abstracted away from any particular object. This frees up the relational aspect of the event, where the participants are no longer fixed components of an action, but are underspecified variables in an event term. Here we show how holophrases can correspond to action-object pairs, what I will simply term *actions* in this paper. Then, later they expand to more underspecified constructions that are closer to *event descriptions*, i.e., verbs. Verbs refer to 'unanchored' properties and relations, that can associate with arbitrary participants, unlike fixed action-object pairings.

To begin, let us review some of the conventional terminology for how events are classified in linguistics. Informally, an event is any situation or happening denoted by a linguistic predicate. While the "argument structure" represents the participants of the situation (i.e., who does what to whom), it says nothing about the temporal properties of events: if something is an event, it must take place in and through time. There are at least two kinds of time-related information that are needed to interpret an utterance: these are known as *tense* and *aspect*. When we describe a situation, it is important for us to know when something happens. In many languages this information is grammaticalized as verbal tense, a linguistic category that locates events in time and relative to other events and time points, usually through tense morphemes and auxiliary verbs.

There is another facet of the temporal dimension of events having to do with their internal temporal structure, called *lexical aspect*. Different lexical aspects can be encoded as distinct event types. The most widely used classification of event types is that of Vendler (1967), who identified four basic event types, shown below (Pustejovsky, 1995; Pustejovsky & Moszkowicz, 2011).

(6) a. **State**: an attribute or property of an object measured over some period of time, and not involving any change; *love, be hungry*.
b. **Activity**: durative and dynamic, they express change of some attribute over an object; e.g., *move, walk, follow*.
c. **Accomplishment**: durative and dynamic like activities, but they are directed and have a natural culmination: *eat a banana, build a house, bake a fish*.
d. **Achievement**: dynamic but the change of state that they involve is instantaneous: e.g., *arrive, sit down, die, faint*.

It is important to point out that the above event type distinctions are made on the basis of studying properties of contemporary natural languages, both individually and typologically (i.e., crosslinguistically). As a result, this distinction is not directly relevant to how prehumans (or primates) may have characterized situations as they experience them, into distinct event classes.

In fact, it is perhaps more plausible that prehumans categorized their experience pragmatically rather than semantically, in terms of personal states and interpersonal interactions. An example of such a distinction is hypothesized below.

(7) PREHUMAN EVENT TYPE DISTINCTIONS:
   a. **State**: an attribute or property of an object measured over some period of time, and not involving any change, e.g., *happy, hungry*.
   b. **Action**: an action-object pairing; identified as embodied routines associated with a specific object, as performed by an individual, similar to an affordance, e.g., *grab banana, slam nut*.
   c. **Action-Result**: an (action-object)-state pairing; this encodes causation where the consequence of an action applied to an object is a specific result, e.g., *slam nut open*.

What is distinct in this classification is the notion of an *action* as a behavior associated with a specific object, which is related to the concept of an affordance (Gibson, 1979; Pustejovsky, 2013; Pustejovsky, & Krishnaswamy, 2016). This is a correlation between an agent who acts on a specific object with a systematic or prototypical effect. For example, when encountering an object, an individual knows the set of actions that can be performed with it. These include actions such as *grasp, move, hold, turn, open, throw, pick up*, and so forth. Because actions are indexed by means of specific objects, they are action-object pairs:

(8) ACTION-OBJECT pairs:
   a. grab-apple
   b. throw-banana
   c. open-mouth

The final hypothesized event type introduced above is a construction composed of an action-object pair along with its result, i.e., an ACTION-RESULT. Primates (and most likely prehumans) understand causation, particularly that resulting from specific actions. While holophrases would likely correspond to actions (action-object pairs), a template-like construction might be associated with this last event type in prehumans.

(9) ACTION-RESULT template:
   a. [slam-nut]-open
   b. [pull-banana]-down

With this review of event types, we now turn to the communicative expressiveness of the different configurations. We begin with gesture, followed by gesture combined with a co-gestural speech modality.

4. The interpretation of gesture

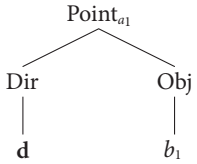
In this section, we examine some of the semantic details of gesturing, in order to better articulate the computational presuppositions assumed by Arbib's stages S4–S5, such that we can transition to S6. As mentioned before, every atomic communicative acts can be expressed by using gesture: i.e., warning, inviting, and greeting. Now we ask what kinds of complex *C*s are possible through gesture alone. Within the context of a common ground, an agent utilizing gesture should be able to express limited instantiations of any of the complex communicative acts, with one major restriction: both the act type and the content (what the event it is referring to) are bound to objects that have values within the embodied common ground. Following Kendon (2004) and Lascarides and Stone (2006, 2009), gestures can be described as simple grammar schemas, consisting of distinct sub-gestural phases, where *Stroke* is the content-baring phase of the gesture.

(10) $G \rightarrow$ (*Prep*) (*Pre_stroke Hold*) *Stroke Retract*

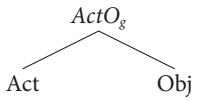
Two gesture types are typically distinguished (Kendon, 1995, Abner et al, 2015): (a) *interactive* (those that manage the communicative dialogue); and (b) *representational* (those that communicate content). Within the latter category, several subtypes are distinguished, including the following: deictic; depicting properties; iconic; metaphoric; and conventionalized gestures (thumbs up).

Gestures can denote a range of primitive action types, including: *grasp, hold, pick up, move, throw, pull, push, separate*, and *put together*. Procedurally, there are a number of ways of conveying the intent to carry out one of these actions, but they all involve two characteristics: (a) the action object is an embodied reference in the common ground; and (b) the gesture sequence must be interpreted dynamically, in order to correctly compute the end state of the event. For the present discussion, there are two kinds of gestures we wish to identify: establishing a reference; and depicting an action-object pair. While the phase-oriented substructure of these gestures is fairly simple, the interpretation of the main phase, *Stroke*, is not trivial. In order to reflect aspects of the meaning associated with a gesture, we will employ an "interpreted tree structure" notation, where a node is decomposed into its functional semantics. This is illustrated for the two gesture classes just mentioned in (11).

(11) a. **Deixis:** $Point_g \rightarrow Dir\ Obj$



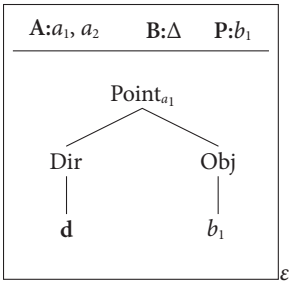
b. **Action:** $ActO_g \rightarrow Act\ Obj$



The interpreted structure in (11a) indicates that the gesture $Point_g$ functionally consists of a deictic orientation, *Dir*, and the referenced or denoting entity, *Obj*. In (11b), the action gesture type, A_g, functionally consists of an action-object pairing, where the action is applied to the object in some prototypical manner.

Consider now how such fairly simple gestures can be embedded into and interpreted within the common ground structure introduced previously in Section 2. Imagine a scenario where two conspecifics, a_1 and a_2, occupy a shared embodied space, $\mathcal{E}$. Agent a_1 points towards the object b_1, indicated through an orientational demonstratum, **d**.

(12)



The structure in (12) captures the embodiment provided by the common ground, CGS, including the domain of objects, $\{b_1\}$, its properties (part of Δ in **B**), and the experienced present of events as they unfold. We assume these are co-perceived (**P**), and the agents are co-situated (they share $\mathcal{E}$). As gesture is intended for visual interpretation, it is directly interpretable by the interlocutor in the context if and only if the value is clearly evident in the common ground, most likely through visual inspection. Directional or orientational information conveyed in a gesture identifies a distinct object or area of $\mathcal{E}$, by directing attention to the *End* of the designated vector cone of the pointing. Hence, the interpretation of the gesture, *Point* relative to the common ground is shown below.

(13) $[\![point]\!] = [\![End\ (cone(\mathbf{d}))]\!]$

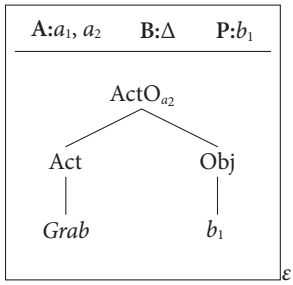
We say that the interpretation function, ⟦.⟧, fully determines the value of the deixis in the context. This denotation (a specific object, for example) can be situated as a response to a question, the posing of a question (regarding some attribute of this object), or wrapped within an imperative concerning this object (pick up this object). These are possible interpretations because of the common ground created by co-situated and co-perceptive embodiment. Hence, such a gesture can inform ('This is the object I intend you to look at') or interrogate ('Is this the object you intended?').

Now consider the interpretation of the action gesture, *ActO*. Actions include a number of distinct action-object pairs, each of which may have an associated gesture, for example:

(14) a. $Grab_g \rightarrow Act\ Obj$
b. $Push_g \rightarrow Act\ Obj$
c. $Throw_g \rightarrow Act\ Obj$
d. $Move_g \rightarrow Act\ Obj\ Loc$

Notice that these gestures are uninterpretable without the context of the common ground and a designation of who is performing the gesture. To see this, let us assume the same CGS as above in (12), where agent a_2 intends to have a_1 grab object b_1.

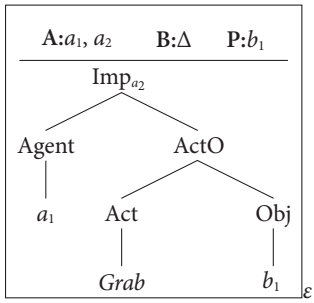
(15)



ε

What we failed to encode was the fact that action gestures are typically interpreted as imperatives (commands). So the interpretation of any action-object pair directed towards an interlocutor in the common ground would typically be taken as a command for the other agent to carry out this action on the salient object. Given this assumption, the *grab*-action can be interpreted as follows:

(16)



Hence, it is the structure of the common ground, together with the convention that gestures are directed at the interlocutor as an imperative, that identified a_1's co-embodied partner, a_2, as the Agent of the action to be performed.

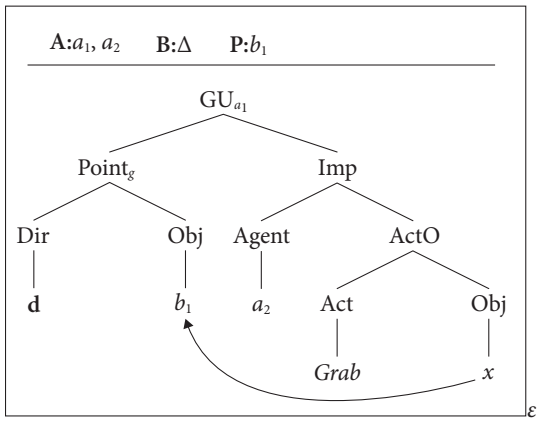
5. Gesture sequences

The ability to create sequences of representational gestures, even a two-gesture utterance, enables a number of possible enrichments to communication. This is assuming that it is accompanied by an associated sophistication in binding and tracking the referents mentioned in the gestures. Assume the same CGS as before, where agent a_1 first points at object b_1, and then gestures to grab it. This corresponds to the simple gesture sequence in (17), loosely following the notation proposed in Fricke (2013), where *GU* is a gesture utterance or gesture sequence.

(17) $GU \rightarrow Point_g\ Af_g$

Embedding this sequence (with *Grab*) in the common ground from (16) gives the following CGS.

(18)



This can be glossed as intending the following expression:

(19) Agent a_1: "That object b_1 grab b_1."

There are three computations of interest in this gesture sequence example: (a) the deixis is interpreted as b_1; (b) the imperative interpretation of the action in the context binds a_2 to the Agent value; and (c) the object value in the action-object *grab*-action is bound to b_1 from the previous deixis gesture.

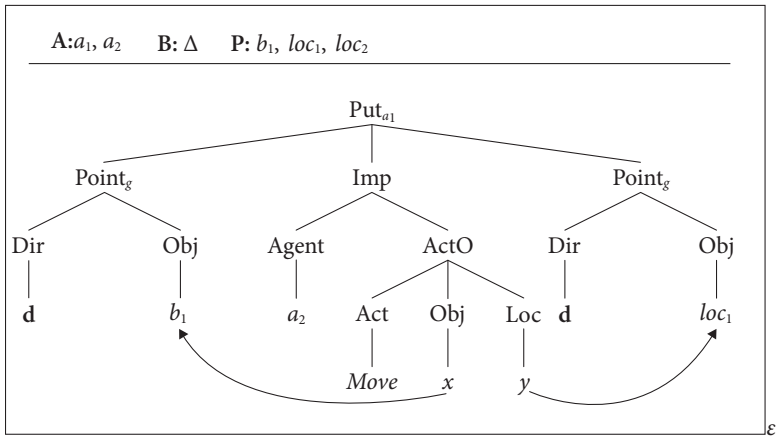
Consider now a slightly more complicated example, with three gestures, forming the operational equivalent of the imperative *put*.

(20) $Put_g \rightarrow Point_g\ Move_g\ Point_g$

We assume the common ground from before, but in order to motivate the need for a distinguishing deixis, we introduce two locations within the embodied space, $\mathcal{E}$, namely loc_1 and loc_2. The intended meaning for the *Put* gesture sequence is as follows:

(21) Agent a_1: "That object b_1 move b_1 to there, the location loc_1."

(22)



The semantics of the *Put* gesture sequence involves the same binding as encountered in the twogesture sequence in (18): namely, the object being moved is first bound by the deictic gesture to a common ground referent, b_1, and then subsequently this binding serves as the antecedent to the missing object in the following action-object.

An additional computation is needed, however, to establish the referent for where the object b_1 should be moved, since this is left underspecified by the *Move* gesture. The second pointing gesture identifies the goal location, loc_1, and serves as the value for the action introduced in the second gesture.

6. Multimodal communication: Gesture-speech ensembles

In this section, we determine what the formal requirements are for modeling stage S6, where multiple modalities are exploited to achieve communicative capabilities beyond the scope of the single modality gesture language we explored in the previous section. We explore the consequences of introducing co-gestural speech to gestures within the common ground.

When two or more modalities are employed for the anchoring of object or event reference within a common ground, they act to jointly contribute constraints towards the interpretation. Reference within the common ground is embodied, and joint reference provides multiple contexts and evidences to ground and reference an individual or an event that the individual is participating in. The joint reference of an object will be called an *ensemble*, and it is this structure, which enables the eventual *displacement* of meaning from the common ground of co-situated communication. An ensemble is an array of distinct modal expressions, temporally aligned within the common ground, jointly working to reference an object or event in the situation.

A multimodal communicative act, C, consists of a sequence of gesture-language ensembles, (g_i, s_i), where an ensemble is temporally aligned in the common ground:

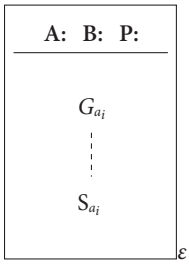
(23) $C = (g_1, s_1); \ldots; (g_i, s_i); \ldots; (g_n, s_n)$.

Following Arbib (2012) and Gillespie-Lynch et al. (2014), gesture drives the structure of the alignment, rather than any syntax associated with speech sequences. An ensemble can also be seen as an array, with aligned elements from each modality. Not every modality need be present in each alignment frame (column). We distinguish two different configurations of an ensemble in alignment: $\begin{bmatrix} G \\ S \end{bmatrix}$, gesture with co-gestural speech; and $\begin{bmatrix} S \\ G \end{bmatrix}$, co-speech gesture with language. The top row of the ensemble will determine the bulk of the interpretation between the two modalities, a point we come to later. For the present paper, we are most interested in co-gestural speech, since both the sequencing structure of the communicative act, and the underlying semantics, is being driven by the gestures used.

(24) **Co-gestural Speech Ensemble**: multimodal communication with Gesture, G, and Speech, S:
$$\begin{bmatrix} G & g_1 & g_i & g_n \\ S & s_1 & s_i & s_n \end{bmatrix}$$

We will bind co-gestural speech to specific gestures in the communicative act, within a common ground, CGS. A dashed line indicates that a co-gestural speech element, S, is aligned with a particular gesture, G.

(25)



We consider three scenarios utilizing a co-gestural language ensemble, with increasingly expressive language capabilities:

(26) a. Demonstratives: lexical items denoting objects (*this/that*) and locations (*here/there*);
b. Sortal classifiers: common nouns (*block, banana*);
c. Event classifiers: actions represented by verb forms (*move, put, grab*).

The goal is to explore how different gesture types, when aligned with specific co-gestural speech tokens, can form ensembles that are inherently more expressive than the gestures alone.

6.1 Co-gestural demonstratives

When one human gestures to another by pointing within a common ground, it is safe to assume that it can be interpreted as *deixis*, as presented in the previous section:

(27) **Deixis**: $Point_g \rightarrow Dir\ Obj$

It has been demonstrated, however, that such an assumption is not possible with chimpanzees, and arguably with prehumans, prior to a language-ready brain. As argued by Pika and Mitani (2009), Arbib et al. (2008), and Bohn et al. (2016), pointing in primates is not a pure referencedenoting gesture, but is conceptually bound to activities associated with body (or object) areas or objects within the region designated by the demonstratum. Nevertheless, for simplicity, I will assume the operational semantics of the pointing gesture introduced above.

Consider the scenario from the CGS in (18) above, where the two-gesture sequence, *Point-Grab* is accompanied by a co-gestural demonstrative that aligns with the *Stroke* phase of *Point*. Co-gestural expressions will naturally be simple, since they are effectively "annotating" the gestural expression and its interpretation. Hence, they can be seen as gesturally-aligned holophrastic utterances, corresponding to Arbib's (2008) notion of protowords. The simple grammar associated with co-gestural demonstratives is given in (28). We distinguish demonstratives by semantic type: i.e., objects or locations.

(28) a. $S \rightarrow Dem_O \mid Dem_L$
b. **Object**: $Dem_O \rightarrow$ **this** | **that**
c. **Location**: $Dem_L \rightarrow$ **here** | **there**

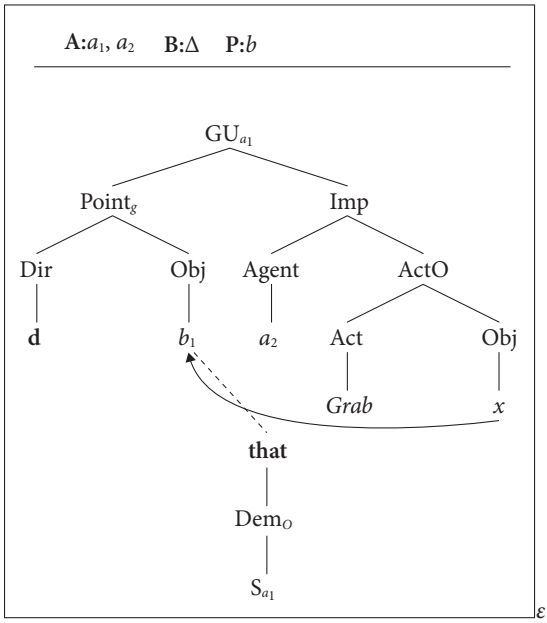
Consider the ensemble in (29), where a demonstrative, **that**, accompanies the deictic gesture, followed by an action gesture of *Grab*, without an aligned co-gestural speech. The intended interpretation for (29) is shown in (30).

(29) $\begin{bmatrix} G & Point_g & Grab_g \\ S & \textbf{that} & \text{---} \end{bmatrix}$

(30) Agent a_1: "**That** object b_1 grab b_1."

When the information in this ensemble is embedded within the CGS, as in (31), we see explicitly how the co-gestural demonstrative, **that**, is indirectly linked to the common ground object, b_1, through the referencing of the deixis gesture, *Point*.

(31)



Let us return to the more complex gesture sequence we encountered in (22), *Point-Move-Point*, annotated with two co-gestural demonstratives aligned with the two deictic gestures. In addition to the object demonstrative, **that**, let us introduce the location token **there**. We will need to allow our simple protolanguage to iterate simple holophrases, as in (32), where a speech utterance can consist of one or more demonstratives.

(32) $S \rightarrow Dem_O \mid Dem_L \, (Dem_O \mid Dem_L)^*$

In reality, however, this is more expressive than we need, and also fails to mirror the interpreted semantics underlying the gestures that the speech units are aligning with. That is, the gesture sequence suggests a pattern where an object token and location token are paired, or two locations are paired, or two objects are paired. Consider the intended expressions below:

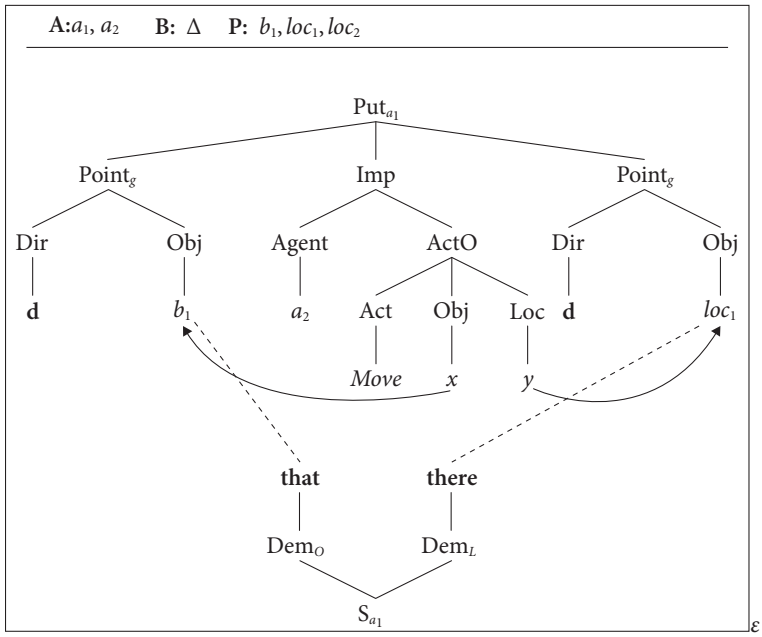
(33) a. "That object move to that location" ($S \rightarrow$ **that**$_O$ **there**$_L$)
b. "'From that location move (object) to that location" ($S \rightarrow$ **there**$_L$ **there**$_L$)
c. "That object switch with that object" ($S \rightarrow$ **that**$_O$ **that**$_O$)

For this reason, it is better to assume that there is no language-only grammar at this stage in prehuman communication: rather, we should think of the linguistic components as *conceptually* co-gestural expressions, where the semantic template driving the communication is carried by the interpretation of the gesture. In any case, the three-gesture expression in (22) has the following ensemble structure.

(34) $\begin{bmatrix} G & Point_g & Move_g & Point_g \\ S & \textbf{that} & \text{---} & \textbf{there} \end{bmatrix}$

The common ground embedding of this ensemble is illustrated below in (35).

(35)



6.2 More expressive co-gestural ensembles

We briefly consider the enrichment of the co-gestural channel to include both common noun (sortal) descriptions, as well as action or event descriptions.

Sortal terms classify the kind of thing being identified or designated with a gesture. This is useful, of course, since a deictic gesture may unambiguously identify an object within the common ground, but often there is uncertainty or ambiguity in the interpreted value of a deixis. The richer the object domain being included in the CGS of the common ground, the less informative a purely orientation-based referencing strategy will be. Adding sortal labels to the co-gestural speech inventory entails imposing tests to the local satisfaction conditions on the value returned by interpreting the pointing gesture, acting as a nominal classification within the common ground. Consider the following example. Even with an ambiguous reference from an ensemble of the deixis, *Point*, and the demonstrative, **that**, an additional co-gestural identifier classifying the object, such as **banana**, will determine reference sufficiently in more complex or noisy common grounds.

(36) a. $S \rightarrow Sort$
b. $Sort \rightarrow$ **banana** | **rock** | **block** | …

Perhaps more significantly, joint ensemble reference of a deixis with a sortal term, (g_i, $sort_j$), unlike a demonstrative, *Dem*, allows for *displaced* reference outside of the context of the common ground.

For example, a *Point* designating an object b_1 in CGS, when aligned with a sortal, **banana**, will form the ensemble, $\begin{bmatrix} G & Point \\ S & \textbf{banana} \end{bmatrix}$. We return to this topic in the next section.

Finally, consider co-gestural speech ensembles, where action-object pairs are explicitly labeled with action terms, such as **grab**. The inventory of holophrases must be enriched to allow for such expressions, as illustrated in (37).

(37) a. $S \rightarrow Action$
b. $Action \rightarrow$ **grab** | **move** | **put** | …

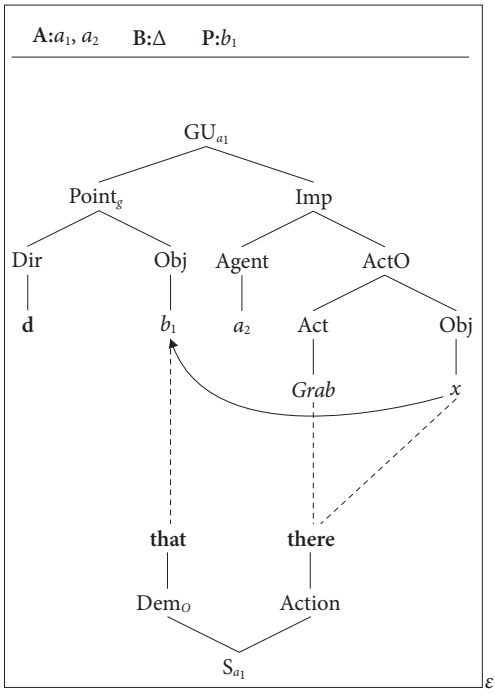
Because of their explicit function in the ensemble, we will call these *co-action classifiers*. Let us revisit the two-gesture expression from (29), with the modification that the ensemble has a cogestural token for each gesture, as glossed in (39).

(38) $\begin{bmatrix} G & Point_g & Grab_g \\ S & \textbf{that} & \textbf{grab} \end{bmatrix}$

(39) Agent a_1: "**That** object b_1 **grab** b_1."

Now consider how this ensemble is embedded into the common ground (cf. (40).

(40)



What is interesting in this representation is that the alignment between the action gesture, *Grab*, and the co-gestural speech, **grab**, illustrates that there is no transparent binding to anything denoting a verb: rather, the alignment is between an action-object pairing and what we could call a proto-verb. That is, the use of co-gestural speech for this case does not refer to an event, and the expression **grab** does not denote a verb. Rather, it is a saturated action-object unit, with no explicit arguments or event participants. Obviously, the deictic-referring demonstrative **that** could be seen as the direct object of *grab*, as a verb. But this would misrepresent both the role of deixis in the common ground, and the nature of actions as action-object units, that are conceptualized primarily as something that is done to objects, not as an activity by itself (Diessel, 2013).

7. Meaning in the absence of common groun

The goal of this paper has been to incrementally build up a coherent representation for encoding the slice of the world that is shared by two conspecifics, when communicating in a shared space. The assumption has been, that by providing a formal specification of common ground, it will be clearer what is at issue when researchers discuss notions of conceptual, communicative, and linguistic competence in pre-humans, primates, and humans. In other words, the aim has been to

provide part of the framework within which to evaluate the theoretical options describing how the language-ready brain emerged. To this end, this final section addresses the transition to Arbib's stage S7: i.e., from the co-gestural speech ensembles discussed in the previous section to the ability to express de-contextualized events without the need for a common ground that referentially grounds the participants in events. By focusing on the structure of embodied meaning, and the conditions under which communication can effectively be carried out within a common ground, we are now in a position to tackle "Hockett's hypothesis".

Hockett (1960) famously proposed that linguistic behavioral competence entails thirteen "design features", one of which he termed the criterion of *displacement*, the ability "to talk about things that are remote in space or time". While bold and insightful, this is Somewhat vague and without operational consequences, without a clear definition of 'remote', 'space', and 'time', relative to the communicative act. Bickerton (2009) and others have harvested related ideas from this theme, but what is still missing is a workable definition of "displacement". It has been the goal of the present paper to provide the platform for such a definition.

For Hockett to say that a linguistic expression can be "displaced", is to presuppose that the context within which the expression initially established its meaning (denotation), is absent from the context within which it can be used. Bickerton's adoption of the concept has brought it some recent currency, but it has had an influence for decades on linguistic researchers who are interested in more gradualists and naturalist approaches to language development, instead of Prometheus-oriented views of language emergence, involving a magical recursion-gene (Hauser, Chomsky & Fitch, 2002). In fact, the formal characterization of "common ground" developed in this paper has been an effort towards determining how exactly displacement might be modeled.

To be "displaced", entails you are displaced from something. This is in fact the embodied common ground we have discussed and modeled in the paper. Only when we have a well-defined enough data structure defining the constitution of the communicative context, from which expressions can be displaced, can we hope to understand what such a notion means. Because the common ground creates a multimodal environment within which gesturer-speaker and viewer-hearer interpret a gesture-utterance in an embodied space, the notion of *displacement* can be defined as any process which removes a linkage or binding from the meaning conveying symbols, when making a communicative act. (cf. Schlenker (2015), Strickland et al (2015))

The consequences of this view are fairly straightforward. Arbib et al. (2008), Gillespie-Lynch et al. (2014), and Arbib (2016) are correct to argue that gesture creates a scaffolding which is picked up and exapted by means of multimodal, integrated, communicative acts. These two pillars give rise to the *possibility* of displaced meaning in the absence of common ground, as exploited by language.

Hence, for Bickerton in particular, the displacement cart is put before the multimodal common ground horse, getting it completely backwards. This section will elaborate on this argument.

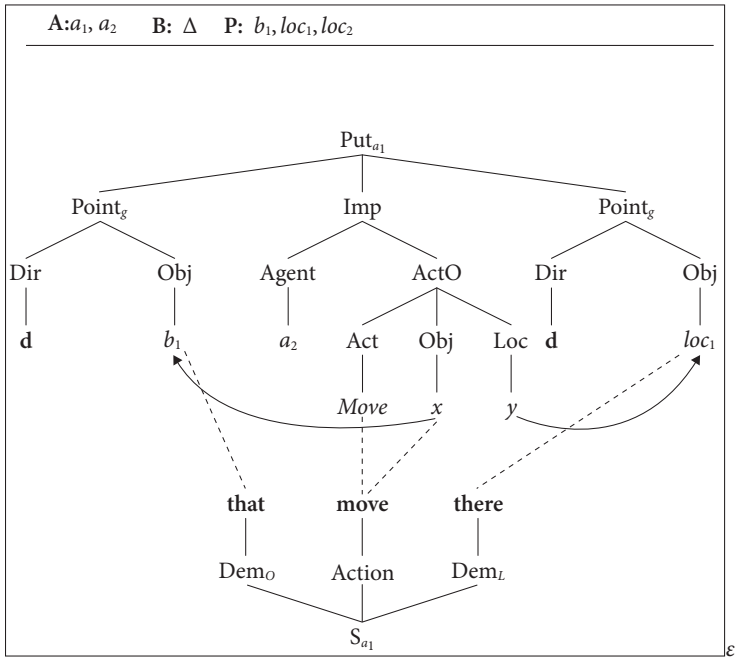
Let us assume that we have the three-gesture sequence encountered above, $Point_g Move_g Point_g$. This is accompanied by a co-gestural speech token for each gesture, with the intended meaning in (41), and the resulting ensemble shown in (42).

(41) Agent a_1: "**That** object b_1 **move** b_1 to **there**, location loc_1."

(42) $\begin{bmatrix} G & Point_g & Move_g & Point_g \\ S & \textbf{that} & \textbf{move} & \textbf{there} \end{bmatrix}$

As the common ground embedding for this sequence in (43) makes clear, co-gestural speech tokens are dependent on the gesture sequence structure for *interpretability*, and the common ground parameters for reference.

(43)



As we pointed out earlier, only when we understand (in some formal sense) how reference is captured or encoded, can we hope to model what sorts of processes of *displacement* are possible. These include any process removing a linkage or binding from the meaning conveyed by the common ground, as captured in the CGS. As an example, consider the ensemble in (42) without the action gesture of $Move_g$, but with the (previously go-gestural) speech token **move**. This changes the ensemble to that in (44).

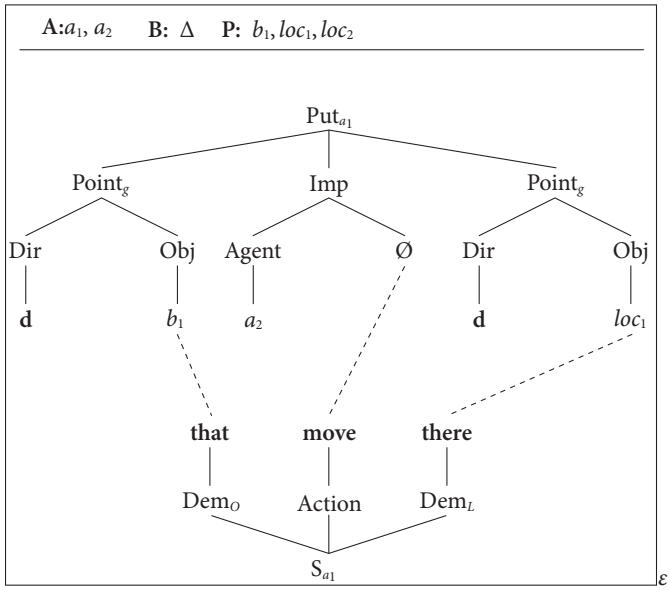
(44) $\begin{bmatrix} G & Point_g & \text{---} & Point_g \\ S & \textbf{that} & \textbf{move} & \textbf{there} \end{bmatrix}$

What is spoken is "**that move there**", with deictic gestures accompanying each demonstrative, **that** and **there**.

(45) Agent a_1: "**That** object b_1 **move there**, location loc_1."

This is a simple case of displacement involving the action-gesture-denoting language token, **move**. To illustrate this within the common ground, we will remove the $Move_g$ gesture from the gesture structure sequence, as illustrated in (46) below.

(46)



It is worth considering what is involved conceptually and computationally in this example. Recall from Section 6 where we stated that the top row of an ensemble acts to drive the interpretation of the act. This is why co-gestural speech is not identical to co-speech gesture. If we make this point explicit in the ensemble representation, then there is also an *interpretive level, I*, that carries the functional structure of the modality in the first row, in our case, gesture. Hence, we can more informatively represent the ensemble in (44) as follows:

(47) $\begin{bmatrix} G & Point_g & \text{---} & Point_g \\ I & [\text{Dir Obj}] & \text{---} & [\text{Dir Obj}] \\ S & \textbf{that} & \textbf{move} & \textbf{there} \end{bmatrix}$

Notice that there is no iconic action gesture to align with **move**, and hence to provide direct meaning. However, as Iverson et al. (1994) Iverson and Goldin-Meadow

(2005) show, co-gestural speech tokens can eventually be exploited as non-gestural tokens. This is also the view argued, somewhat more generally by Givòn (1995, 1998) and Armstrong and Wilcox (2007). Operationally, what this entails, for any single gesture-speech ensemble, is a transition to an ensemble where the co-gesture is absent, and then to a completely displaced use of the speech token as an independently content-bearing symbol. This is illustrated below.

(48) $\begin{bmatrix} G & Move_g \\ I & \text{[Act Obj Loc]} \\ S & \textbf{move} \end{bmatrix} \Rightarrow \begin{bmatrix} G & \text{---} \\ I & \text{[Act Obj Loc]} \\ S & \textbf{move} \end{bmatrix} \Rightarrow \begin{bmatrix} I & \text{[Act Obj Loc]} \\ S & \textbf{move} \end{bmatrix} = \begin{bmatrix} S & \textbf{move} \\ I & \text{[Act Obj Loc]} \end{bmatrix}$

When extended to multiple displacements, such as that in (49), where non-gestural speech is used in place of two gestures, Arbib's (2016) notion of construction becomes relevant.

(49) Agent a_1: "**Grab banana**."

The ensemble for this would be as shown in (50).

(50) $\begin{bmatrix} G & \text{---} & \text{---} \\ I & \text{[Act Obj]} & \text{[Dir Obj]} \\ S & \textbf{grab} & \textbf{banana} \end{bmatrix}$

This is a potential source for the transformation of an action to a full event, where the participant of the action is explicitly represented in the speech ensemble. That is, the linguistic token no longer is associated with an action-object pair, but denotes an action, since the interpretive level *I*, is still able to carry the functional information for how the two separate tokens combine, compositionally.

(51) $\begin{bmatrix} I & \text{[Act Obj]} & \text{[Dir Obj]} \\ S & \textbf{grab} & \textbf{banana} \end{bmatrix} \Rightarrow \begin{bmatrix} \text{I} & \text{[Act]} & \text{[Obj]} \\ \text{S} & \textbf{grab} & \textbf{banana} \end{bmatrix}$

This brings us finally to a freely occurring event interpretation for the token grab, from the actiondenoting interpretation it previously carried.

(52) grab(AGENT, OBJ)

Hence, displacement, enabled by the dual scaffolding of gesture and multimodal alignment, gave rise to the verb.

8. Toward a new road map

Let us now turn to the questions motivating this paper, namely: what are the differential capabilities between humans and pre-humans, and how is common

ground relevant to this question? More specifically, what conceptual abstraction in cognition was necessary in going from Arbib's stage S6 to S7, namely, human language competence?

To answer this, first consider two of the signature hallmarks of human language, compositionality and recursion. Recursion is the application of a rule that allows a sequence to contain an instance of itself, thus contributing to the generation of potentially infinite chains (as in embedding constructions). The chain can refer a plan with subplans, actions with subactions, events with subevents, thoughts with subthoughts, and of course, linguistic expressions with subexpressions. Compositionality dictates that a complex expression is determined by the meanings of its component parts and their syntactic arrangement. It is the key for understanding how we can build and understand a potentially infinite number of word combinations (including those we have never heard or used before) based on the knowledge of the lexical meaning of a finite set of words and phrases in our repertoire, and the syntactic rules that are used for combining them. From this perspective, compositionality can be regarded as the necessary link between lexical and sentential meamng.

We can see how these two properties are linked by contrasting the behavior of verbs and actionobject pairs, as discussed in previous sections above. Modern humans (stage S7) can understand and create arbitrarily complex event descriptions of the physical and mental world around them. This is facilitated by the *functional behavior* of verbs: they act as open (or underspecified) events, where their participants (arguments) are not determined (or anchored) until composed with other words to build a larger expression. Hence, the notion of *function application* (applying a verb to its argument) is both the core of compositionality as well as the facilitator of recursion.

For verbs to have this property, they must be able to recursively access the meaning of their arguments. As mentioned above, the absence of true common ground is the norm in most communicative interactions, and words and phrases in language act as proxies to objects and actions that are not present in the shared context. Function application can be seen as a recursively applied proxy: taking a proxy of an object (such as the missing banana), and returning a proxy event (I ate the banana). Hence, the power of functional abstraction, as demonstrated by verbs, is the ability to effectively *simulate* the absent common ground that originally provided denotations and references for the banana and my action with it. In contrast to verbs, actions (action-object pairs) are inherently *saturated* (rather than open) by reference to objects in the common ground. Hence, they do not act functionally (at least not in the same manner). Actions correspond to holophrastic proto-verbs, where the object to the action is anchored by means of and within the common ground.

As argued above, the ability to decompose an action-object pairing into further component parts (Arbib's fractionation) is scaffolded by two parameters: leveraging the sequence-based syntax from gesture while aligning such gestures with additional modal anchors, that can link (or index) a concept (action or thing) back to a shared experience. Hence, referential displacement is facilitated by the integration of multiple modalities in communication. The decomposition of gestural and linguistic holophrases into distinct components with identifiable semantics leads to general lexical enrichment, as well as structural (syntagmatic) enrichment to the communicative act (linguistic expression).

Language has evolved to reference any conceivable experience or thought. I have argued that the language-ready brain distinguishes itself with a robust ability to recognize and conceptualize events. Adopting Arbib's (2006) thesis that gesture provided the scaffolding for object referencing and simple action sequencing, I show how dual aligned modalities are necessary to situate co-gestural speech in a common ground. This provides the precondition with which processes of displacement can begin, which allow for unsituated and decontextualized expressions to still carry meaning. A gradual process of displaced reference was responsible for allowing a local, co-situated, embodied form of communication, to develop into the conceptual expressiveness of human language.

Acknowledgements

I would like to thank Michael Arbib for his detailed comments and suggestions for how to frame the paper. I would also like to thank Nikhil Krishnaswamy for useful discussion and input. This work was supported by Contract W911NF-15-C-0238 with the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO). Approved for Public Release, Distribution Unlimited. The views expressed herein are ours and do not reflect the official policy or position of the Department of Defense or the U.S. Government. The paper was prepared for a workshop funded by NSF Grant No. BCS-1343544 "INSPIRE Track 1: Action, Vision and Language, and their Brain Mechanisms in Evolutionary Relationship" (M.A. Arbib, Principal Investigator).

References

Aboitiz, F. (2013) "How did vocal behavior 'take over' the gestural communication system?." *Language and Cognition* 5, no. 2–3: 167–176. https://doi.org/10.1515/langcog-2013-0011

Abner, N., Cooperrider, K., & Goldin-Meadow, S. (2015). Gesture for linguists: A handy primer. *Language and linguistics compass*, 9(11), 437–451. https://doi.org/10.1111/lnc3.12168

Arbib, M. A. (2002) The Mirror System, Imitation, and the Evolution of Language. *Imitation in animals and artifacts*, 229.

Arbib, M. A. ed. (2006) *Action to language via the mirror neuron system*. Cambridge University Press.

Arbib, M. A., (2008). Holophrasis and the protolanguage spectrum. *Interaction Studies*, 9(1), pp.154–168. https://doi.org/10.1075/is.9.1.11arb

Arbib, M. A. (2008). From grasp to language: Embodied concepts and the challenge of abstraction. *Journal of Physiology-Paris*, 102(1–3), 4–20. https://doi.org/10.1016/j.jphysparis.2008.03.001

Arbib, M. A. (2012). *How the Brain Got Language: The Mirror System Hypothesis*. New York & Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199896684.001.0001

Arbib, M. A. (2013). Precis of how the brain got language: the mirror system hypothesis. *Language and Cognition*, 5(2–3), 107–131. https://doi.org/10.1515/langcog-2013-0007

Arbib, M. A. (2016). Towards a computational comparative neuroprimatology: framing the language-ready brain. *Physics of life reviews*, 16, 1–54.

Arbib, M. A., Liebal, K. and Pika, S., (2008). "Primate vocalization, gesture, and the evolution of human language." *Current anthropology* 49. 6, 1053–1076. https://doi.org/10.1086/593015

Arbib, M. A. & Rizzolatti, G. (1997). Neural expectations: a possible evolutionary path from manual skills to language. *Communication and Cognition*, 29, 393–424.

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

Armstrong, D. F., & Wilcox, S. (2007). *The gestural origin of language*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195163483.001.0001

Asher, N. (1998). Common ground, corrections and coordination. *Journal of Semantics*.

Asher, N. and Lascarides, A., (2003). *Logics of conversation*. Cambridge University Press.

Asher, N. and A. Lascarides. (2008). Commitments, beliefs and intentions in dialogue. *Proceedings of Londial*, 35–42.

Austin, J. L. (1962). *How to do things with words*. London: Oxford University Press

Bach, K. (2003). Speech acts and pragmatics. Blackwell. Guide to the philosophy of language, 147–167.

Bickerton, D. (1990) *Language and Species*, University of Chicago Press. Bickerton, Derek (2009). Adam's Tongue. Hill and Wang.

Bickerton, D. 2009. *Adam's tongue: how humans made language, how language made humans*. Macmillan.

Bohn, M., Call, J. and Tomasello, M., (2016). Comprehension of iconic gestures by chimpanzees and human children. *Journal of experimental child psychology*, 142, pp.1–17. https://doi.org/10.1016/j.jecp.2015.09.001

Cangelosi, A. (2010). Grounding language in action and perception: from cognitive agents to humanoid robots. *Physics of life reviews* 7(2), 139–151. https://doi.org/10.1016/j.plrev.2010.02.001

Clark, H. H. (1996). *Using language*. Cambridge University Press. https://doi.org/10.1017/CBO9780511620539

Clark, H. H., S. E. Brennan, et al. (1991). Grounding in communication. *Perspectives on socially shared cognition* 13(1991), 127–149. https://doi.org/10.1037/10096-006

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior*, 22(2), 245–258. https://doi.org/10.1016/S0022-5371(83)90189-5

Cooper, R. and Ginzburg, J. (2015). "Type theory with records for natural language semantics". *Handbook of Contemporary Semantic Theory*, pp.375–407. https://doi.org/10.1002/9781118882139.ch12

Corballis, Michael C., (2003) *From hand to mouth: The origins of language*. Princeton University Press.

Deacon, Terrence W. (1998). *The symbolic species: The co-evolution of language and the brain*. WW Norton & Company.

Diessel, H. (2013) "Where does language come from? Some reflections on the role of deictic gesture and demonstratives in the evolution of language." *Language and Cognition* 5.2–3: 239–249.

Engle, R.A. and Clark, H.H. 1995. Using composites of speech, gestures, diagrams and demonstrations in explanations of mechanical devices. In *American Association for Applied Linguistics Conference*, Long Beach, CA.

Fogassi, L., Coude, G. and Ferrari, P. F., (2013). The extended features of mirror neurons and the voluntary control of vocalization in the pathway to language. *Language and Cognition*, 5(2–3), pp.145–155. https://doi.org/10.1515/langcog-2013-0009

Fricke, E. (2013). Towards a unified grammar of gesture and speech: A multimodal approach. Body-language-communication. *An international handbook on multimodality in human interaction*, 733–754.

Gillespie-Lynch, K., Greenfield, P. M., Lyn, H., & Savage-Rumbaugh, S. (2014). Gestural and symbolic development among apes and humans: support for a multimodal theory of language evolution. *Frontiers in psychology*, 5. https://doi.org/10.3389/fpsyg.2014.01228

Gibson, J. J. (1979). *The ecological approach to visual perception: classic edition*. Psychology Press.

Gilbert, M. (1992). *On social facts*. Princeton University Press.

Givòn, T. (1995). *Functionalism and grammar*. John Benjamins Publishing. https://doi.org/10.1075/z.74

Givòn, T. (1998). On the co-evolution of language, mind, and brain. *Evolution of Communication*, 2(1), 45–116. https://doi.org/10.1075/eoc.2.1.04giv

Glenberg, A. M. & Gallese, V. (2012). Action-based language: a theory of language acquisition, comprehension, and production. *Cortex*, 48(7), 905–22. https://doi.org/10.1016/j.cortex.2011.04.010

Goldin-Meadow, S. and Alibali, M.W. 2013. Gesture's role in speaking, learning, and creating language. *Annual review of psychology*, 64, pp. 257-283.

Grice, H. P. (1981). Presupposition and conversational implicature. In C. Peter (ed.), *Radical pragmatics*. New York: Academic Press, pp. 183–198. Reprinted in Grice, H. P. (1989) *Studies in the way of words*. Cambridge: Harvard University Press, pp. 269–282.

Hsiao, K. -Y., S. Tellex, S. Vosoughi, R. Kubat, and D. Roy. (2008). Object schemas for grounding language in a responsive robot. *Connection Science* 20(4), 253–276. https://doi.org/10.1080/09540090802445113

Hauser, Marc D., Noam Chomsky, and W. Tecumseh Fitch. (2002) "The faculty of language: what is it, who has it, and how did it evolve?." *Science* 298.5598: 1569–1579. https://doi.org/10.1126/science.298.5598.1569

Hockett, C. F., (1960). The origin of speech. *Scientific American*, 203(3), pp.88–97. https://doi.org/10.1038/scientificamerican0960-88

Iverson, J. M., Capirci, O. and Caselli, M. C., (1994). From communication to language in two modalities. *Cognitive development*, 9(1), pp.23–43. https://doi.org/10.1016/0885-2014(94)90018-3

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological science*, 16(5), 367–371. https://doi.org/10.1111/j.0956-7976.2005.01542.x

Kamp, H. and Reyle, U. 1993. From discourse to logic: Introduction to model-theoretic semantics of natural language, formal logic and discourse representation. *Studies in Linguistics and Philosophy*. Kluwer.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *The relationship of verbal and nonverbal communication*, 207–228. The Hague: Mouton.

Kendon, A. (1995). Gestures as illocutionary and discourse structure markers in Southern Italian conversation. *Journal of pragmatics*, 23(3), pp. 247–279.

Lascarides, A. and M. Stone. (2006). Formal semantics for iconic gesture. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (BRANDIAL)*, pp. 64–71.

Lascarides, A. and M. Stone. (2009). A formal semantic analysis of gesture. *Journal of Semantics* 26(4), 393–449. https://doi.org/10.1093/jos/ffp004

Pika, S. and Mitani, J. C., (2009). The directed scratch: evidence for a referential gesture in chimpanzees. *The prehistory of language*, 1, pp.166–181. https://doi.org/10.1093/acprof:oso/9780199545872.003.0009

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press, Cambridge, MA.

Pustejovsky, J. (2013). Where things happen: On the semantics of event localization. In *Proceedings of isa-9: International workshop on semantic annotation*.

Pustejovsky, J. and N. Krishnaswamy. (2014). Generating simulations of motion events from verbal descriptions. In *Lexical and Computational Semantics* (*SEM 2014). ACL.

Pustejovsky, J. and J. L. Moszkowicz. (2011). The qualitative spatial dynamics of motion in language. *Spatial Cognition & Computation* 11(1), 15–44. https://doi.org/10.1080/13875868.2010.543497

Rizzolatti, G., and Arbib, M. (1998). Language within our grasp. *Trends in Neurosciences*, 21, 188–194. https://doi.org/10.1016/S0166-2236(98)01260-0

Schlenker, P. (2015). Visible meaning: Sign language and the foundations of semantics. Manuscript, Institut Jean-Nicod and New York University.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge university press. https://doi.org/10.1017/CBO9781139173438

Stalnaker, R. (2002). Common ground. *Linguistics and philosophy* 25(5), 701–721. https://doi.org/10.1023/A:1020867916902

Steels, L. (1999). The talking heads experiment: Vol. I. *Words and meaning (Special preedition)*. Brussels: Vrije Universiteit Brussel.

Steels, L. (2011). Modeling the cultural evolution of language. *Physics of Life Reviews*, 8(4), 339–356. https://doi.org/10.1016/j.plrev.2011.10.014

Steels, L. (ed), (2012). *Experiments in cultural language evolution*. Amsterdam: John Benjamins. https://doi.org/10.1075/ais.3

Stout, Dietrich, & Hecht, Erin E. (2017). Evolutionary neuroscience of cumulative culture. *Proceedings of the National Academy of Sciences*, 114(30), 7861–7868.

Strickland, B., Geraci, C., Chemla, E., Schlenker, P., Kelepir, M., and Pfau, R. (2015). Event representations constrain the structure of language: Sign language as a window into universally accessible linguistic biases. *Proceedings of the National Academy of Sciences*, https://doi.org/10.1073/pnas.1423080112

Tomasello, M. and M. Carpenter. (2007). Shared intentionality. *Developmental science* 10(1), 121?125. https://doi.org/10.1111/j.1467-7687.2007.00573.x

Vendler, Z. (1967). *Linguistics and Philosophy*. Ithaca: Cornell University Press.

Volterra, V., Caselli, M. C., Capirci, O., & Pizzuto, E. (2005). Gesture and the emergence and development of language. *Beyond nature-nurture: Essays in honor of Elizabeth Bates*, 3–40.

von Uexkll, J. (1957). A stroll through the worlds of animals and men: A picture book of invisible worlds, in Schiller, C. H. (ed), *Instinctive Behavior: The Development of a Modern Concept*. New York: International Universities Press, 5–80.

Wang, I., M. Ben Fraj, P. Narayana, D. Patil, G. Mulay, R. Bangar, R. Beveridge, B. Draper, and J. Ruiz. (2017). Eggnog: A continuous, multimodal data set of naturally occurring gestures with ground truth labels. In Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition.

Whitehead, A. N. (1919), *An Enquiry concerning the Principles of Natural Knowledge*, Cambridge: Cambridge University Press.

Ziemke, T. & Sharkey, N. E. (2001). A stroll through the worlds of robots and animals: Applying Jakob von Uexkll's theory of meaning to adaptive robots and artificial life. *Semiotica*, 2001(134), 701–746. https://doi.org/10.1515/semi.2001.050

Address for correspondence

James Pustejovsky
Brandeis University
415 South Street, MS-018
Waltham, MA 02454 USA
USA

jamesp@brandeis.edu