LANGUAGE AND COGNITIVE PROCESSES, 1998, *13* (2/3), 373–424

The Emergence of Propositions from the Co-ordination of Talk and Action in a Shared World

Brian Hazlehurst and Edwin Hutchins

Department of Cognitive Science, University of California, San Diego, USA

We present a connectionist model that demonstrates how propositional structure can emerge from the interactions among the members of a community of simple cognitive agents. We first describe a process in which agents coordinating their actions and verbal productions with each other in a shared world leads to the development of propositional structures. We then present a simulation model which implements this process for generating propositions from scratch.

We report and discuss the behaviour of the model in terms of its ability to produce three properties of propositions: (1) a coherent lexicon characterised by shared form-meaning mappings; (2) conventional structure in the sequences of forms; (3) the prediction of spatial facts. We show that these properties do not emerge when a single individual learns the task alone and conclude that the properties emerge from the demands of the communication task rather than from anything inside the individual agents. We then show that the shared structural principles can be described as a grammar, and discuss the implications of this demonstration for theories concerning the origins of the structure of language.

Requests for reprints should be addressed to B. Hazlehurst, Sapient Health Network, 720 SW Washington St., Suite 400, Portland, OR, USA.

The work reported in this paper was supported by grant No. IRI-9311496 from the Division of Information, Robotics, and Intelligent Systems, and the Cross Disciplinary Activities Office of the National Science Foundation. We are also grateful to three anonymous reviews for many constructive comments.

INTRODUCTION

Propositions: What are They, and Where Could They Come From?

As early as Plato, the question "What are propositions?" has been addressed in terms of sentences which have been understood to instantiate propositions. It has often been noted that these sentences contain a verb that provides a characterisation of a noun, and that these components are involved in some kind of "claim" made about the world by the sentence (cf. Prior, 1976). A prototypical example is the sentence "The cat is on the mat". In more general terms, such sentences have been described as claiming (asserting or denying) something about (predicating) some thing (an argument) in the world.

While such descriptions yield rich clues about the nature of propositions—suggesting, for instance, intimate relations to cognition and language—many puzzles remain unaddressed and unsolved. In this article we address the origins of propositions. Where do propositions come from? At least one major tradition (arguably, *the* major tradition) in cognitive science has taken the origins of propositions to reside in the innate constitution of mind. According to this view, propositions exist because the mind is constituted by a propositional language of thought. Sentences that instantiate propositions are simply behavioural manifestations of this presumably biological (but, in practice, theoretically inferred) fact.

In this article, we present an alternative characterisation of the origins of propositions and the nature of mind. In one sense, this alternative entails a return to the original observations about propositions already cited, that is, propositions "claim (assert or deny) something about (predicate) some thing (an argument) in the world". The alternative we present is built upon the premise that such "claiming" is performed by a situated speaker, and therefore is "about something in the world" as constituted by histories of action in a shared material world.

In particular, we take the world of interlocutors to be a place where acts of reference are mediated by language that has conventionalised functional consequences. The use of language is seen as a process involving the organisation of agent behaviour in a material world. Sentences that instantiate propositions are seen to emerge *from* this organisation, while simultaneously providing structure *for* that organisation. This system is made possible by the development of co-ordination between structure that is internal to an agent and the shared external structure created in situated language use.

We present a computer simulation to show how simple propositional sentences could arise in interactions among agents who must learn to co-ordinate their actions in a shared environment. First, we present a

working definition of propositions which can be decomposed into some minimal cognitive specifications. This is followed by an outline of a process which could produce propositions in interaction. Next, we discuss the implementation of a computer model for simulating this process and present results and analysis from one example simulation. Finally, we conclude with a general discussion of what the model demonstrates and what can be learned from it.

A Working Definition of Propositions

We begin with the claim that "propositions" can be viewed as conventional relationships among structures that represent states of the experienced world for interpreters who are members of a community of practice. Although propositions are generally given material form in sentences they are not the sentences themselves. Rather, propositions reside in the conventions or shared behavioural organisation that provides the capacities for producing sentences with certain properties. The relationships which are defined or given form by such conventions are of several kinds. We take the following three relations to be a minimal set required to explain the nature of propositions: (1) A coherent lexicon: The relations between the constituent elements of agents' sentences and their experience in the world; (2) syntax: The relations among the constituent elements of the sentences (i.e. the principles that govern the sequential ordering of tokens in the sentences); and (3) semantics: The relations between sentences and practices of the community in which sentences are employed to do some kind of cognitive work.

The sentences "The cat is on the mat" and "Above[ball, table]" instantiate propositions. These sentences employ conventional arrangements of constituent tokens. The tokens stand in conventional relationship to a shared experience of the environment. In addition, such sentences are *functionally related* to the way the objects in the world are (or could be) situated. The perceptions that ground the functional relations between sentences and states of the world are mediated by socially constructed experience; that is, propositional sentences function to make known or understood to some agents something about some thing. Propositional sentences make claims about a world that is known to agents through their shared experience in it.

The sentence, "The cat is on the mat" is a sequence of public structures, which functions to represent a state of the environment in which the object known as "cat" is perceived or imagined by the interpreter of the sentence to be spatially related to or "on" the object known as "mat". What the objective relationship between sentence and world is depends on the functional properties of agents who employ that sentence as a part of living in that

world. There is no experience-independent set of criteria which completely define the relationship between the material structures given in sentences and those given in the world. The only way to understand or explain that relationship is by reference to the invariants of agent perception and action potential in the world as well as principles of inter-agent communication, social dependence, and interactional history. These invariants constrain how the world (through learning) can be known to agents, and how sentences mediate that knowing.

Propositions, therefore, are conventional ways of structuring representations which capture (and create) relevant properties of the socio-historically constructed and natural world. Clearly, much of human culture consists of these kinds of structures which serve as resources for organising behaviour. Although our model presents a vastly simplified case compared to any human language, culture, or situation, it is none the less our hope that the model sheds light upon the functional properties that might be operative in the human case.

A Process for Generating Propositions from Scratch

The simulation described here demonstrates how individuals and communities of individuals—artificial cognitive agents—might create conventions entailing compositions of public representational structures that predicate something about the world they share. Each agent is composed of neural network modules and is endowed with functional behaviours that stand for the most rudimentary perceptual, motor, social, and verbal capacities. The agents are members of a population we will call a "community" because they interact with each other in a shared environment. In the simulation, pairs of agents engage each other in interactions we will call "discourse", which engenders co-ordinated action in a problematically shared world of visual perception. The actions in discourse which require co-ordination are shifts in focus of attention enacted in order to bring shared attention to a target object that is intended by one agent (the "speaker") yet unknown to the other agent (the "listener"). In the process of co-ordinated parsing of the visual field (in service of the built-in goal), interlocutors achieve shared understanding about objects and about the spatial arrangements of objects in the environment. The major claim of the article is that agent-generated verbal structures that are produced in co-ordination with the joint shifts in focus of attention become sentences that embody propositions about space and about the arrangement of objects in space.

The sequentially ordered, agent-created public structures,"sentences", are expressions whose constituent forms, ordering principles, and meanings

become shared. They come to function to describe (or make known) structure in the world. The structure in the world is given in visual scenes which contain simple arrangements of objects on a square lattice. When agents interact as speaker–listener pairs in discourse, the scene constitutes a shared visual field which yields information about objects, as well as one's own and the other agent's focus of attention (Fig. 1).

Each agent can only attend to a small part of the scene (a single location in the visual field containing a single object), and this attending is available to both agents as a public "finger" or focus of attention. In order to collectively focus upon some object (intended by the speaker and unknown to the listener) within this shared visual field, agents must maintain joint attention while converging on the target object. Agents employ verbal productions to facilitate this co-ordination. This "facilitation" is at first not helpful since the development of useful word forms and meanings (and the mappings between these) takes place within these same encounters in the world. Early in the simulation the structure of verbal productions is simply given by the (partially) random initial conditions of the agents, and this is not helpful since this structure is meaningless for the task at hand. Over time, as a product of learning to co-ordinate their discourse in the contexts of many different scenes and intended target objects, sentences come to represent shared understandings about the world and how to locate objects within it.

In order to understand the emergence of propositions in the model, we will need to understand the bases for the emergence of the three kinds of relations given in our earlier definition of propositions. These were: (1) The relations between the constituent elements of agents' expressions and their experience in the world; (2) the relations among the constituent elements of the sentences (i.e. the sequential ordering principles entailed by the sentences); and (3) the relations between sentences and practices of the community in which sentences are employed to do some kind of cognitive work. We now briefly discuss the development of each of these kinds of relation and the basis for their emergence in the simulation, before turning to a description of the implementation of the simulation model.

Relations Between the Constituent Elements of Agents' Expressions and Their Experience in the World. The problem here is to develop referring structures for which the denotating function is shared across the population. It has been demonstrated elsewhere that communities of artificial agents (connectionist networks) can invent such a shared lexicon. The elements of the emergent lexicon, form-meaning pairs, appropriately distinguish visual phenomena for members of a community who share visual experiences in communicational encounters (Hutchins & Hazlehurst, 1995).

The lexicon emerges through repeated interaction between randomly chosen pairs of agents sharing randomly chosen visual scenes. In each

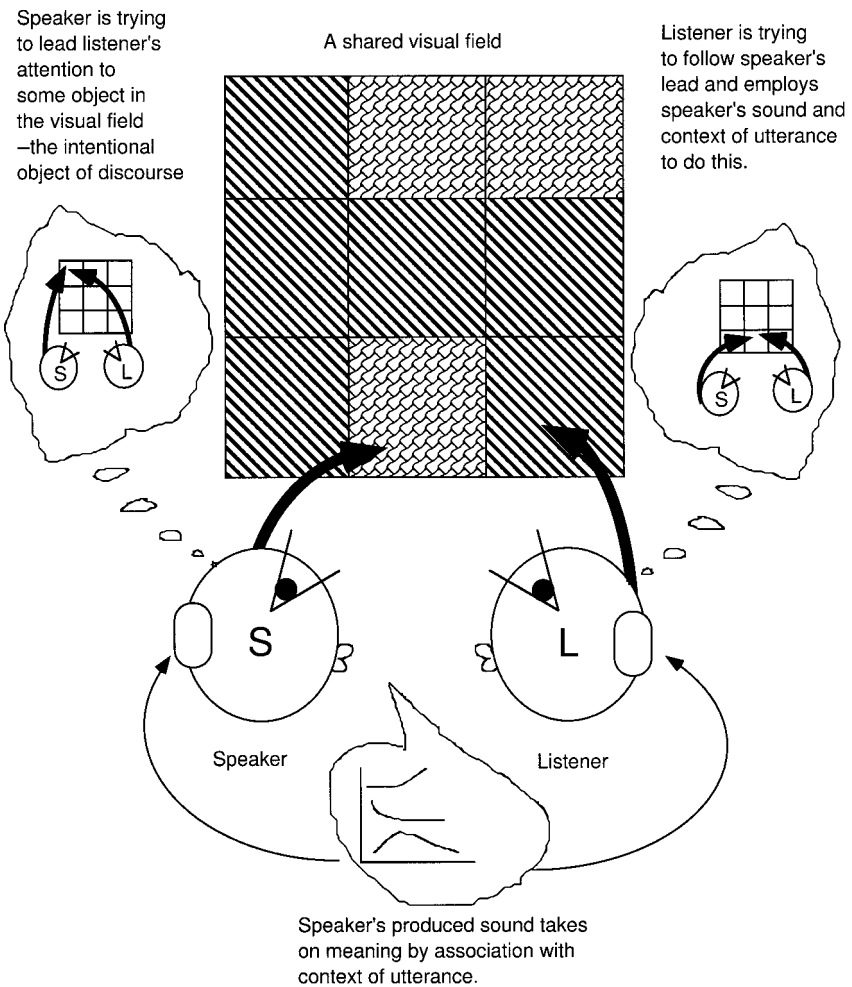

FIG. 1. General outline of agent interaction.

Two agents engage the shared task of locating an intentional object of discourse held privately in one agent's (the speaker's) mind. This requires negotiating the contents of a shared visual field in interaction. This negotiation entails the co-ordination of agent attention shifts (modelled as public finger moves) and sound productions. After many such interactions, in which agents get experience as both speakers and listeners, conventions arise for accomplishing the task. The conventions developed entail a coherent lexicon and principles for sequential construction of tokens from the lexicon.

interaction, each agent produces a word (a public form, or verbal structure) that represents its current understanding of the scene taken as a visual classification problem. At first these verbal productions are meaningless as they are products of (partially) random initial conditions—that is, agents' internal organisations are highly random at the beginning of the simulation. In interaction, each agent's verbal production must simultaneously classify the scene (as part of a self-organising process known as "auto-association" in connectionist terminology) and do so in a way that is compatible with the partner agent's classification. Over time, the communities productions will, under certain conditions, converge upon a shared classificatory scheme—a shared lexicon. Sharing the lexicon entails sharing a set of form-meaning mappings that are in systematic correspondence with the experience of structure in the world. The existence of a shared lexicon constitutes specific relations between the constituent elements of agents' public expressions and their experience of visual phenomena originating in the world, as prescribed by the definition of propositions given earlier.

Relations Among the Constituent Elements of the Sentences. In the simulation model just described (Hutchins & Hazlehurst, 1995), it was assumed that complete visual scenes, the objects referred to by agents' public expressions, are unambiguously shared by interlocutors in interaction. In the current work we relax that constraint in order to pursue the very real cognitive and social problem of sharing a focus of attention. Interlocutors in the simulation reported later share a visual field (a scene) in interaction, but can only attend (independently) to small portions of that field at one time. As a product of co-ordinating joint attention, interlocutors develop descriptions of scenes that appropriately site objects within them. The organisation of elements in descriptions is constrained by the spatial relations among objects in the scene and the negotiation of conventional preferences for attending to spatial relations among objects.

This added complexity in the model creates a condition in which terms must refer to objects in the world and the sequential ordering of the terms must reflect the sequence of shifts in focus of attention. Sequential organisation becomes an issue because agents employ the lexicon over the course of an extended interaction during which joint attention shifts from object to object in co-ordinated search for the object intended by the speaker. At each time step (during which joint attention is either focused on an object or shifting between objects) a verbal token is produced by the speaker. Collected over the course of the interaction, the sequential production of speaker's tokens constitutes a sentence. The principles which structure this sequential organisation are not given a priori, but rather emerge in the course of negotiating the dynamics of joint attention. These principles participate in determining the relations among constituent

elements of sentences, a kind of relation prescribed by the definition of propositions already given.

Relations Between Sentences and Practices of the Community. A second condition created by the introduction of a problematically shared world into the model is that something other than direct perception of simple objects in the environment must be the basis for grounding (at least some) terms in the lexicon. In particular, the spatial relations between objects in the world are not specified by the properties of those objects themselves. For example, the basis for perceiving one object as "on" another is not given in the perceptions of objects themselves but, rather, by their positioning relative to each other and relative to a frame of reference. As such, the exact same arrangement of objects in space can be the basis for any of a large number of different relations. Therefore, the particular relation chosen to predicate an arrangement must be imposed by the speaker and, if it is to be understood correctly, also by the listener.

In the simulation model, the agents encode these abstract relations by using internal structure, which participates in the shifting of attention, to generate external structure. These external structures come to be reliably associated with particular spatial patterns of shifts in attention, and in this way come to have meanings that might be glossed as "up", "down", "left", and "right". We give these glosses with the realisation that their "meanings" can only be established on the basis of their use. We do not know that the term we labelled "up" would not be better glossed "above". When employed as a constituent of a larger sentence which places the term in the context of other terms referring to objects in space, the term could come to mean something like "above", or "on". What we do know is that a particular term emerges which is reliably produced by the interactants when the focus of attention moves toward what we have designated the "top" of the visual display and that such a term can reliably direct the attention of the listener to the object above the current focus. Thus, the selection of a particular relation from among those given by the arrangement of objects in the world is accomplished by the interactants through the negotiation of attention shifting in discourse. This negotiation leads to sharing the verbal productions which refer to the shared shifts in attention, and to sharing the function of sentences as mediators of the process whereby target objects within the visual field are reached. This function is reasonably viewed as a kind of predication over the contents of the visual field and is constructed in the practices of the community, as prescribed by the definition of propositions.

IMPLEMENTATION OF A SIMULATION MODEL

Here we give a functional description of a simulation model of the development of propositions among a community of artificial cognitive agents. Following this description, we analyse the results of one simulation run.

Imagine a world in which pairs of cognitive agents drawn from a population come together in interactions requiring the negotiation of shared visual experience. Agents are equipped with sensory surfaces which provide a means for experiencing structure in the shared setting. Agents are also endowed with an ability to sense and produce two kinds of structure in the world—namely, sounds and body positioning (finger location or, equivalently in this simulation, direction of gaze)—each of which has a temporal component.

This world is shown in general outline in Fig. 1. The two agents have distinct roles as speaker and listener. The speaker has in mind some particular physical object that is available in the shared visual field. The speaker wants the listener to attend to the object. The listener has no such object in mind but wants to understand the speaker's directed activity and thus attempts to follow the speaker's actions with similar actions of its own. Over multiple interactions in multiple contexts, the population of agents develops structures—both internal and external—for accomplishing this task. We will argue that this process is capable of producing language-like structures that instantiate propositions and can be used to shape behaviour that appears rule-governed.

The World

The world of the simulation is composed of a population of *objects* and a population of *agents*.

Objects are known to agents through the activation of agents' visual sensory surfaces. This is implemented by encoding objects as vectors of real values representing distributions of light intensity on the visual sensory surface of agents. In the simulation reported later there are only two objects, represented by the scalar values 0.0 and 1.0, referred to in the discussion as Object0 and Object1.

Agents are members of a population. Each agent is composed of a modular, primarily connectionist, architecture. This architecture is the same for all individuals in the population. The behaviour of an agent is determined by the weights on connections in the modules that compose the agent. These weights are initially assigned small random values. Agent learning shapes behaviour by changing the values of weights as a consequence of experience. Depending on the initial weights and the pattern of experience, different agents may arrive at different weight structures that produce the same

functional behaviours. That is, development will produce functionally, but not physically, equivalent agents during the course of the simulation.

The Practice

Life in this world consists of *interactions* between pairs of agents. We therefore refer to the population as a *community* of agents, and to the simple activity of the group as a *practice*. Each interaction entails a series of *discourse turns* (sometimes called *time-steps*) which take place in an *environment*. The primary component of the environment is a visual *scene*, which is a 3×3 square arrangement of objects. Given that there are two kinds of objects in the world, and each object occupies one of the nine locations in the scene, there are 2^9 or 512 possible scenes. The scene projects onto the visual sensory surfaces of both agents of the pair in interaction. The scene of an interaction therefore constitutes a *shared visual field* (see Fig. 2).

The visible environment of an interaction is composed of a scene together with *engagement* of the scene by the pair of agents. An agent's engagement of the scene consists of a focus of attention, which is marked by the location of the agent's *finger*, within the scene. That is, the agents can see the shared visual field and they can see each others' fingers, which indicate their respective foci of attention in the scene.

An agent's finger serves a dual function. First, it occupies a location within a larger visual field that is of special interest to the owner of the finger. The contents of this location (in addition to the other contents of the scene) are projected onto the visual sensory surface of this agent. Second, an agent's finger provides information to partner agents about where the first agent is attending. This information about shared focus of attention enables agents to co-ordinate their actions in spite of asymmetric knowledge about the task. Learning in interaction produces internal structure in the individuals that can sustain the co-ordinated shifts in finger location as speaker and listener move their attention to the target location.

Agents are also capable of producing sounds, which, in the case of speakers' sounds, are emitted into the environment of the interaction and heard by the pair of interactants (Fig. 1). The speaker emits one sound per time-step. Concatenated over the course of the interaction, the sequence of sounds so produced is called a "string" or "sentence", and is said to be composed of "constituent tokens" or "words", which denote aspects of the contexts in which the tokens were produced.

A denotative function is a form-meaning mapping, a set of relationships between sounds and the interpretation of visual patterns. The agents develop form-meaning mappings between sounds and two kinds of visual patterns: The contents of the locations in the visual field (sounds associated

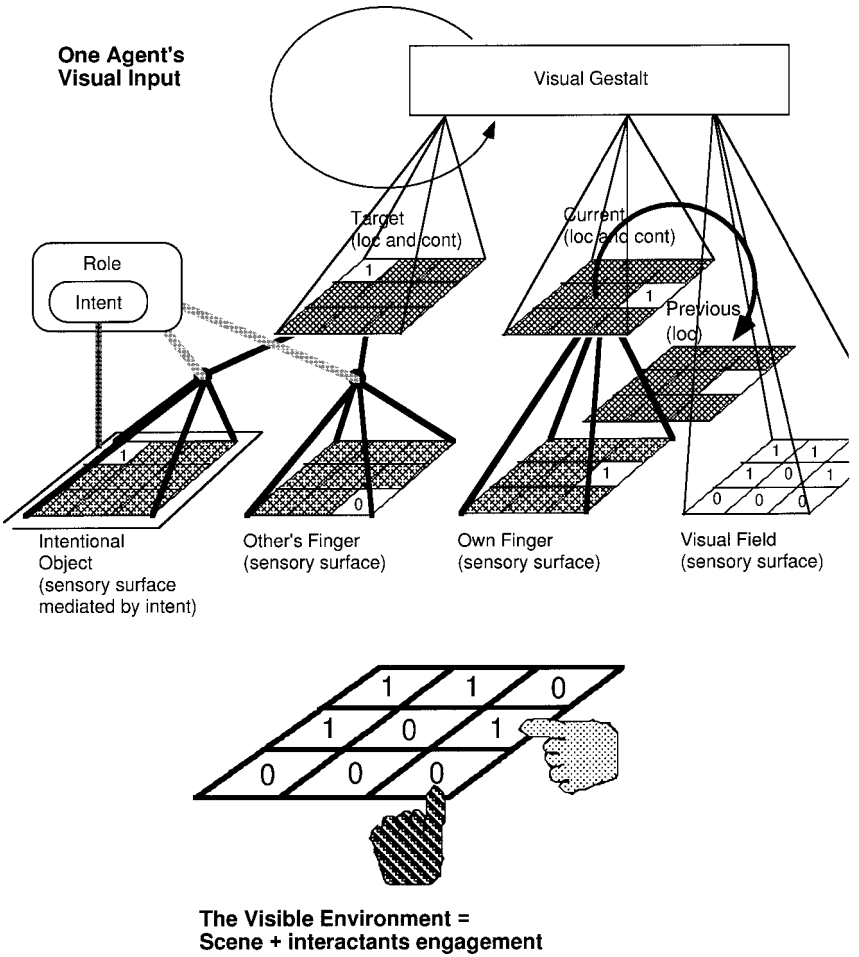

FIG. 2. An agent's Visual Input.

Each agent possesses four sensory surfaces for constructing visual information about the state of the environment. The diagram shows how this information appears on those surfaces and how that information propagates forward in the agent's visual channel. Thick arrows and lines represent a mechanism that simply copies information from one layer of connectionist units to another. Thin arrows and lines represent learnable connection weights that modulate the propagation of information from one layer of units to another, in the standard connectionist sense of feed forward of activation. The agent's "Role" is defined by the stance the agent takes in the interaction (as either "Speaker" or "Listener") and serves here to gate whether "Intentional Object" or "Other's Finger" will serve as "Target" information. "Intent" (activated when Role=Speaker) is the private internal state of the speaker, which defines the intentional object of discourse. The function of the layer marked "Previous" in the diagram is to provide a simple memory (employed for generating a finger movement teaching signal—see Fig. 3 and Table 1).

with objects), and changes in the location of agents' fingers in the visual field (sounds associated with actions).

The two agents of an interaction take on (by random assignment) asymmetric functional roles: One is *speaker*, one is *listener*. The *shared goal* of discourse is to reach the *intentional object* of the discourse which is privately held by the speaker. This object is a product of the speaker's *intent* (a private internal state specifying where the object is located) and the contents of the shared visual field (specifying which object appears there). Whereas the intentional object is always evident somewhere within the visual field, it is only available to the listener's attention if the listener happens to already be attending to it, or if the speaker can lead the listener's attention to the object (Figs. 1 and 2).

The objective of discourse is said to be "shared" because each agent is playing a distinct (yet mutually co-operative) part in the activity of locating the speaker's intentional object through the control of joint attention. In particular: (1) The speaker will break off the interaction whenever the pair become disco-ordinated—that is, if the two agents fail, at the beginning of a turn, to be attending to the same location; (2) the speaker directs the discourse in the sense that (a) it is the speaker's sounds which are heard by both agents, and (b) the speaker initiates each turn; (3) the listener employs the speaker's words and shift in attention as the target behaviour for itself. That is, the listener is explicitly trying to follow the speaker and simultaneusly to replicate the speaker's verbal productions.

These three built-in principles of agent activity, together with reinforcement learning conducted on each time-step (as discussed later), generate a system of behaviour in which speakers attempt to reach their intentional objects while constrained to shift attention in each context in a way that can be anticipated by the listener. In other words, speakers are trying to lead listeners to the objects they have in mind, and employ shared expectations about (a convention for) the way to do this. Such a system is capable of developing structures that solve the fundamental problem of co-ordinating action and generate propositions in the process.

In the remainder of this section we describe the simulation framework in detail. Each subheading below identifies the label of a procedure employed by the algorithm given at the end of the section.

Update_Sensory_Surfaces

Each agent possesses four *sensory surfaces* (Fig. 2) onto which visual information from the environment is mapped. Each surface contains a different kind of information about the environment. The Intentional Object surface (active for speaker only) specifies the location of the speaker's intent and the contents of that location in the visual field. The two

Finger surfaces specify own and other finger locations, and contents of the visual field at those locations. The Visual Field surface constructs a projection of the entire arrangement of objects in the environment, i.e. the scene. Collectively, these four sensory surfaces comprise a vehicle for the agent's *visual input* (Figs. 2 and 3).

Each surface is activated by inputs as follows. The inputs to the Intentional Object and two Finger surfaces are each coded by a two-element tuple. The first element of each tuple is a nine-place binary vector encoding a single location within the visual field (i.e. eight bits are off and one is on). The second element of each tuple is an encoding of the object within the visual field found at the location specified by the first element. In the simulation described here, the second element of each tuple is a 1 or 0 representing Object1 and Object0, respectively. The Visual Field is activated by a single vector of nine values (1s and 0s) representing the positioning of nine objects (of two types) within the environment.

The procedure Update–Sensory–Surfaces activates an agent's four visual sensory surfaces by copying these input representations onto the agent's input layers.

Feed-forward–Activation

Propagating activation from visual inputs through the weights that connect layers within the agent leads to the production of sounds. The sound produced by an agent is encoded by a vector of real values, meant to represent a point in acoustic space. In the simulation described here, this space is three-dimensional (i.e. coded by three real values). The sounds emitted by a speaker are heard by *both* speaker and listener without distortion (Fig. 3). The sounds produced by listener in the interaction are ignored. However, in order for the denotation function to become shared across the population, listeners' verbal productions are error-corrected—during the Learn procedure, as described later—in the direction specified by speakers' sounds on each time step.

Produce–Action

Propagation of activation forward again (now including the input of a sound representation at the agent's "ear"—see Fig. 3) results in the agent producing a motor action. Agents' fingers are under motor control, and agents must learn to traverse the visual field in co-ordination with each other. On each time-step of the interaction, each agent is capable of selecting (according to a stochastic process governed by agent age, as discussed later) to move its finger in one of four directions (up, down, left, and right) or selecting to leave its finger at its current location (i.e. selecting a "don't move" action). However, enactment of the selection is subject to

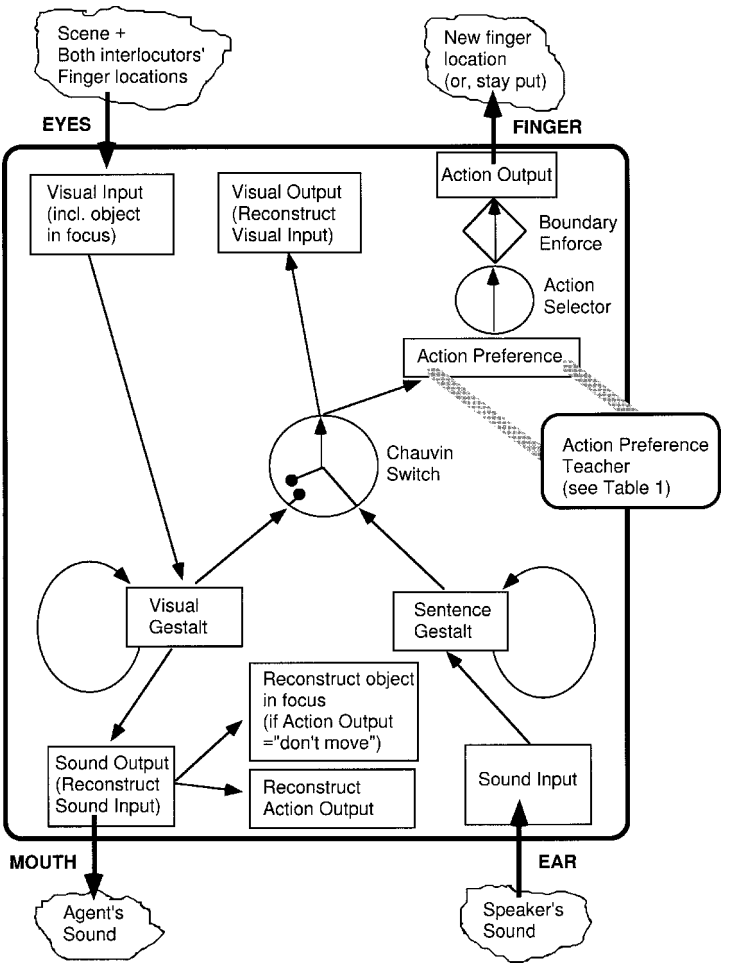

FIG. 3. Agent architecture and learning.

The processing components which constitute a single agent are enclosed in the large box. The nature of inputs and outputs to the agent are shown by connecting arrows which cross this boundary. Arrows within this boundary (within the agent but not enclosed by any other boundary) stand for sets of parameters (connectionist weights), which propagate activation between standard connectionist units organised into layers (shown with labelled rectangles). (Note: Not all layers and connections are shown. This is a slightly simplified diagram from the actual architecture, but it captures the essential features.) All shown connections are modifiable weights, and the architecture employs the back-propagation of error from output layers to adjust these parameters. Target outputs for this procedure are provided, in all but one case, by already available input or produced output activations. This can be seen by noting layers labelled with the word "Reconstruct". Each of these layers takes as a "target" for output activations produced or already available, as labelled. The difference between the "target" activations and the actual activations produced at these layers provides the signal for error correction through connection weight modifications (i.e. learning). Notice that the "Sound Output" layer is treated as both an "output" and a "hidden" layer. That is, error is derived by adding the difference between output and target activations to the error back-propagated from two output

enforcement of boundary conditions on the visual field—the selection of a shift which leads off, the visual field results in a "don't move" action being realised in the world (Fig. 3).

More formally, actions are produced in a two-step process that first generates a Selected-action and then a Realised-action that is enacted in the world. The process begins with the activations at the Action Preference layer (Fig. 3). There are five units in this layer, each corresponding to one of the five possible actions. Each unit's activation level is treated as a preference for choosing the corresponding action. Action selection is performed probabilistically (by the stochastic Action Selector, see Fig. 3) according to the following formula. Let M be the set {1,2,3,4,5} denoting all possible moves agents can ever make, and let $m = [m_1,m_2,m_3,m_4,m_5]$ be the vector of activation at the agent's Action Preference layer following feedforward of activation in the agent's network architecture. Then, for all $i \in M$, m_i denotes the agent's degree of preference for move i and P_i, the probability that i is the Selected-action, is given by the equation:

$$P_i = \frac{e^{(m_i/T)}}{\sum_j e^{(m_j/T)}}, \text{ for all j in M}$$

T is a parameter (a positive real value $\leqslant 1.0$, called the "Temperature") which controls the randomness of the selection made by the stochastic selector. The higher the Temperature, the more likely actions not preferred by the agent will be selected. The lower the Temperature, the more likely actions preferred by the agent will be selected. T changes during the lifetime of the agent—it is a developmental parameter—starting out high (to allow random traversing of the visual field) and gradually decreasing with agent "age", as preferred parsings of the visual field are learned. Agent age is measured in terms of experience speaking in interactions. T varies as a function of agent age in discrete steps of age τ, which is fixed for all agents. In particular, given a starting temperature (identical for all agents) of T_0, and some agent ψ whose age (number of interactions as speaker) is α_ψ, then the temperature for this agent as a function of its age is given by:

---

layers. A single exception to where "target" activations for learning come from in the system, is the explicit teaching function which generates an error signal for the agent's action preferences. The nature of this function is described in Table 1. Action preferences are produced deterministically by standard feed-forward of activation through the shown sets of weights. Action selection is then made according to a stochastic process, as described in the main text. Action selections that would violate the boundary of the visual field are converted into "don't move" actions before the final Action Output is produced. Finally, the "Chauvin Switch" is a simple gating mechanism for the passage of activation and error through the system. In particular, it controls whether the visual and the auditory channels, or just the auditory channel, is carrying the burden of processing. The purpose of this device is to provide a vehicle for bootstrapping structured auditory processing in the solutions to visual processing problems entailed by the tasks of interaction in the world.

$$T(\alpha_\psi) = \frac{1}{\left(\frac{1}{T_0} + \frac{\alpha_\psi}{\tau}\right)}$$

In the simulation reported later, $T_0 = 0.5$ and $\tau = 4000$.

If the agent's Selected-action were to generate a new finger location within the boundary of the visual field, then it is said to be "realised" in the environment because it is materially evidenced in the agent's new finger location. However, if the Selected-action were to generate a new finger location outside the boundary of the visual field, then the action "don't move" is realised and the location of the agent's finger does not change. The action enacted is known as the Realised-action.

Thus, at each time-step in the interaction agents produce (independent) actions, each of which generates either a shift to an adjacent finger location or a pause at the agent's current finger location. We take these actions to be functionally equivalent (in these simple agents) to "shifts in attention" or "pauses in attention", respectively.

Learn

Learning in the simulation is a product of back-propagating the errors produced by applying five different target representations to the activation states of five different layers which have resulted from the forward propagation of activation through the agent. Four of these layers receive their target or teaching signals from representations already produced by the agent or available in the environment. These four layers are all labelled with the term "Reconstruct" in Fig. 3. The fifth layer, labelled Action Preference in Fig. 3, receives a specially constructed target signal called the Action Preference Teacher. We explain the nature of this teacher below.

At the beginning of the simulation, before agents have had any learning experiences, all sounds are nearly identical—the real values that constitute an agent's sounds are all mid-range in value. This is because sounds are a product of the process of propagating activation across sets of constraints or connection weights which, initially, are random values. (See the pathways leading from Visual Input to Sound Output, in Fig. 3.) Each agent gradually learns to produce the same sound as the other agent produces in response to a given object or action. There are no a priori targets. The "correct" behaviour emerges as all the members of the community jointly learn to shape their behaviours to the behaviours of their fellows. Learning tunes the connection weights to minimise output errors. For example, the target of the listener's sound output is the speaker's sound output. Through this process, stable form–meaning mappings gradually develop.

This development takes place because each sound produced by an agent must be capable of reconstructing a coding of the agent's action and (if the action is "don't move") the contents of the visual field where the agent's attention is focused (see layers labelled "Reconstruct object in focus" and "Reconstruct Action Output" in Fig. 3). This requirement to reconstruct these inputs out of the sounds forces the sounds to encode the structure present in the inputs. Furthermore, listener's sound productions are error-corrected in the direction of speakers' productions (see layer labelled "Sound Output" in Fig. 3). As a result, and due to the fact that all agents get experience in both roles, the form–meaning mappings that develop in the course of the simulation become shared by all members of the population.

When agents pause with their foci of attention fixated on some object in the environment, the words they produce take their meanings from association with perception of that *object* (modulated by past experience with this and other objects). On the other hand, during a shift of attention the word produced takes on meaning due to being associated with the *action* itself. This division of the semantic landscape into meanings of two kinds (of objects and of actions) results from a gating mechanism whereby the activations (i.e. real values) that encode the sound produced by the agent must carry the information necessary to reproduce visual perception of the object (internal to the agent) only when the agent pauses its attention. (See the connections that feed forward from Sound Output in Fig. 3.)

Notice that the semantic partition between "action words" and "object words" is a "built-in" feature of the architecture. By architectural design (i.e. principles built into the internal information processing mechanism of the agent) representations at the sound output layer have different constraints placed upon them for the two different action cases of movement and fixation. Only in the later case must the sound representation support reconstruction of the object in focus for the agent. However, this is the complete extent of what is predetermined. What is in fact built-in is simply an internal mechanism which dictates when (according to the agent's Action Output) an already available teaching signal (the object in agent's focus of attention) should be applied as an error-correcting measure. The actual shape of the semantic partition must emerge under the self-organising process of co-ordinating one's actions with words and with one's partner's actions and words.

Agents learn which shifts in attention to make from the results of the actions they and their discourse partners produce in the world. This learning employs an error signal applied to agents' action preferences, called the Action Preference Teacher (see Fig. 3 and Table 1). In this case, learning works to minimise a dynamic error function which integrates the various requirements of the discourse objective. That is, the function seeks to promote co-ordinated convergence of joint attention on the (private)

intentional objects of speakers, but leaves open exactly what the shape of the behaviours serving this outcome will be.

The Action Preference Teacher implements a function based on the following four considerations. First, each agent's actions must provide for convergence on some target location in the visual field. This is accomplished by employing a simple distance metric over the agent's current and target finger locations. However, for each of the two agents of the interaction, this target location may differ: For the speaker the target is the intentional object of discourse, for the listener (who has no direct access to the speaker's internal states) the target is the location of the speaker's finger. In addition

TABLE 1
The Action Preference Teacher

	Teaching Objective	*Information Needed*	*Where info Obtained from?*	*Effect of Generated Signal*
1.	Converge upon target and pause there. (Note: target is intentional object loc for speaker and speaker's finger loc for listener)	(a) What is distance to target following action? (b) Was action "don't move"?	Analysis of new Visual Input Identification of Action Output	Differentially reinforce selected action based on remaining distance to target. Max. reinforce if distance = 0 and action = "don't move"
2.	Agree with other interlocutor	Is there collocation with other agent's finger following action?	Analysis of new Visual Input	Reinforce selected action if YES, else inhibit selected action and reinforce "don't move"
3.	Don't violate the boundary of the visual field	Was selected action realised in the world?	Comparison of Action Output with selected action	Inhibit selected action if NO
4.	Don't revisit the location where located on the previous time-step	Is new location of finger same as location on previous time-step?	Comparison of new Visual Input with memory of location on previous time-step	Inhibit selected action if YES

The teacher function entails the integration of four different objectives. Each objective is shown in terms of what it is, what information is needed to accomplish it, where that information is drawn from, and how the function implements the stated objective in terms of a training signal applied to the agent's Action Preference layer (see Fig. 3).

to convergence of attention upon a target location, pausing of attention *at* the target location also generates positive reinforcement.

A second consideration built into the teacher of agents' action preferences is consensus, defined as agreement with the other interlocutor about where to move on each time-step. This agreement must overcome the asymmetrical access to information: It is the speaker who acts first (with the intention of leading the listener to a privately held target location and object), while the listener employs the speaker's action as a target for itself. Actions that result in co-ordinated shifts in finger location yield positive reinforcement of the action, whereas disco-ordination yields negative reinforcement of this action and positive reinforcement of the "don't move" action.

The third objective of the teaching function is to inhibit revisiting locations just visited. The fourth objective is to inhibit actions which would violate the boundary of the visual field. All of these objectives are accomplished by the teacher generating reinforcement and inhibition of the actions that produce these conditions, when the respective actions are in fact produced.

More formally, the Action Preference Teacher constructs a learning target at each time-step, for each agent, as follows. At most two of the units in the agent's Action Preference layer are constrained by the teaching signal: (1) The "don't move" action choice, and (2) the Selected-action choice. All other units are left unconstrained, effectively implementing a "don't care" condition upon their output activations. Here are the rules for constructing the two target values for the two units in the layer that receive a teaching signal.

For the "don't move" unit the target is given by:

$$t_1 = \begin{cases} \text{Max}[x, 1 - e^{-2x}] & \text{if (Own-finger-loc} \neq \text{Other-finger-loc)} \\ & \text{AND (Realised-action} \neq \text{"don't move")} \\ x & \text{Otherwise} \end{cases}$$

where x is the activation of the "don't move" unit. Notice that the "Otherwise" clause implements a "don't care" condition because the target is equal to the output activation and thus there is no error.

For the Selected-action unit the target is given by:

$$t_2 = \begin{cases} 0.0 & \text{if Selected-action} \neq \text{Realised-action} \\ 1.0 & \text{if (Selected-action = Realised-action} \\ & \quad = \text{"dont' move")} \\ \dfrac{(h - d + \text{Agree Credit} - \text{CycleCost})}{b} & \text{AND (Own-finger-loc = Targ-loc)} \\ & \text{Otherwise} \end{cases}$$

where h is a constant (6.0); d is the shortest distance (number of steps) from Own-finger-loc to Targ-loc; AgreeCredit is a constant (of size 4.0) which is positive if Own-finger-loc = Others-finger-loc and negative otherwise; CycleCost is a large constant (10.0) if Own-finger-loc(t–2) = Own-finger-loc (t) and zero otherwise; and b is simply a normalising factor (10.0).

Roughly speaking, targets constructed according to the above rules teach agents—via weight modifications which reduce output errors—to perform the following kinds of behaviours. (1) If agents disagree over where to move their fingers, then "pausing" is positively reinforced if it is not already a highly activated choice. (2) If the action selected by the agent is "invalid" (violates boundary of visual field) then inhibit it. (3) If the action selected by the agent entails successful "halting" (pausing at agent's own target location) then maximally reinforce it. Otherwise, (4) reinforce the selected action of the agent according to a weighting that aims to (a) encourage actions taken that reduce the distance to one's own target, (b) encourage actions that are in agreement with other agent, and (c) discourage actions that lead to "cycling"—that is, revisiting locations with one's finger that have recently been visited.

What emerges from this learning scheme are conventions for traversing the visual field in the service of reaching all possible target objects from all possible initial conditions (i.e. all possible starting states of the environment). Since word forms come to map one-to-one onto the contexts in which they are uttered, these emergent conventions for traversing the visual field constitute construction principles (i.e. ordering preferences) over the set of possible agent sentences.

Finally, since sentences mediate the practices of the community they must be capable of functioning as stand-ins for the activity itself. This is accomplished by forcing agents to employ the constituents of sentences (and the complete sentence) in the absence of visual input during learning. In effect, agents perform a learning pass without visual input (a kind of "mental simulation"), which maps the sentence onto the functional outcomes obtained with visual input in place, in each context. This is accomplished by implementing a two-pass learning procedure. On each pass, the propagation of activation forward in the architecture and the propagation of error backward through the architecture are gated by the "Chauvin Switch" (Fig. 3).

The Chauvin Switch simply controls what information passes through, and therefore which sets of connection weights (those of the auditory channel or those of both visual and auditory channels) contribute to generating outputs, as well as which weights receive error correction. (See Chauvin, 1988 for a thorough analysis of such cross-modal integration of information as a model of symbol grounding.) Over time, as good words and conventions for concatenating words emerge, agents become able to employ

well-formed sentences to produce the functional outcome of the associated activity—without the aid of any visual input. This is true because, given a meaningful sentence, the auditory channel (with the Chauvin Switch connecting sentence gestalt to motion preference) must do all of the work necessary to produce the outputs associated with the sentence.

Simulation Algorithm

A simulation is composed of an initialisation followed by a sequence of interactions.

Initialisation:

1. Create a population of agents whose architecture is as specified in Figs. 2 and 3. Each agent's learnable weights are randomly selected from the interval $[-0.5,+0.5]$.
2. Create the set of training scenes, a subset (462 random members) of the 512 possible scenes. The remaining 50 scenes are set aside for later testing.

Interaction:

1. Set t = 0 (time-step, or turn counter). Set Exit–Condition = False. Randomly pick two agents from population. Randomly assign roles of Speaker and Listener. Randomly pick a scene from training set. Place each agent's finger (randomly and independently) on some location in the scene. Seed Speaker's "intent" with some (random) location within the visual field. Note: Exit–Condition becomes True if and only if (for Speaker) any of the following hold:
   (a) Selected-action≠Realised-action
   (b) Own-finger-loc≠Others-finger-loc
   (c) Own-finger-loc = Target-loc AND Realised-action = "don't move"
   (d) Own-finger-loc(t) = Own-finger-loc(t–2) OR t = Max.
2. Update–Sensory–Surfaces (both agents). IF (t≠0) THEN Learn (both agents). IF (Exit–Condition = True) THEN Quit.
3. Speaker: Feedforward–Activation (produces sound in environment). Produce–Action (generates new finger location in environment).
4. Listener: Update–Sensory–Surfaces. Feedforward Activation. Produce–Action (generates new finger location in environment).
5. Set t = t + 1. GOTO (2).

In closing this section, it needs to be stressed that agents begin the simulation unorganised with respect to the internal structure necessary to accomplish the tasks required by interactions. At the beginning of the simulation, agent verbal productions are not capable of denoting because

they all take on mid-range values in the representational space of sound. Similarly, agent actions are in co-ordination only by chance, because the parameters responsible for producing shifts in attention are all random values before learning begins. The development of structure to solve the co-ordination problem takes place as a consequence of error correction procedures which take their data from outcomes already available in the world or internally generated by the agent in question, as described.

RESULTS AND ANALYSIS

A small simulation (of population size 2) was run using a random sampling of 462 of the 512 possible scenes. The remaining 50 scenes were set aside for later testing. The outcomes of interactions were plotted across time (Fig. 4). Every interaction terminates under one of four conditions: (1) Speaker "Halts" with focus of attention upon its privately held intentional object (that is, speaker leads listener to the target and selects a "don't move" action); (2) speaker and listener "Disagree" (that is, they fail to co-ordinate their shifts in attention prior to speaker halting); (3) speaker selects an "Invalid" shift in attention, which cannot be realised (that is, speaker's selected shift in attention would violate the boundary of the visual field, resulting in a "don't move" action being realised in the world); and (4) speaker "Cycles" by revisiting a location with its attentional focus that was visited on the previous time-step or some "Max" number (24 in our simulation) of time-steps have taken place in the current interaction. Only those interactions in which the two agents are focused on the same thing at time-step 0 are considered.

As shown in Fig. 4, these four termination conditions provide a good window into the evolution of the system. Each point in the graph (plotted every 2000 interactions) is obtained by averaging the outcomes of the previous, 200 interactions. At each such point in time, the values of the four conditions sum to 1.0 (i.e. they account for all cases of interaction termination).

At the beginning of the simulation "Invalid" moves are the cause of termination for nearly one in five interactions, a value which is close to the expected value for violating the boundary on a random walk over the 3×3 visual field. Over time, the frequency of invalid shifts in speaker attention decreases, and these actions are (nearly) extinguished by the 45,000 interaction mark of the simulation.

The category of interaction termination labelled "Cycles + Max" in Fig. 4 climbs from 0% to almost 20% after 25,000 interactions have taken place. Early in the simulation agents learn to co-ordinate their actions by remaining fixated on the locations where they initiate an interaction. This is a consequence of agents failing to agree on where to shift attention (which,

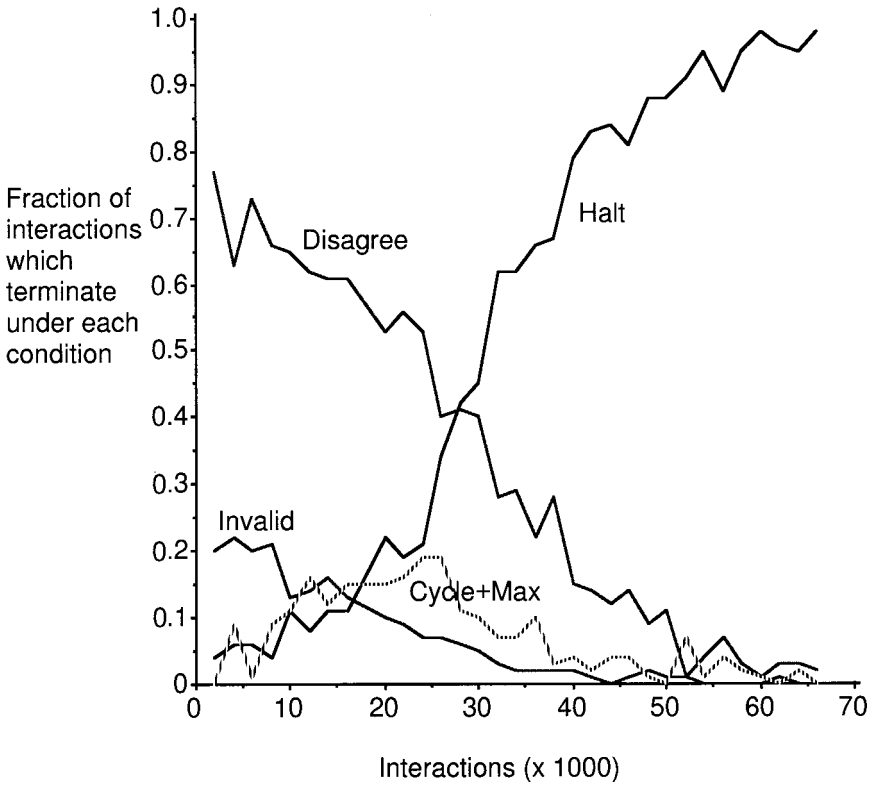

FIG. 4. The progression of one simulation run.

Time is represented in numbers of (thousands of) interactions along the x-axis. At every 2000 interactions four points are plotted representing averages over the previous 200 interactions. These values indicate the proportion of interactions that terminate under each of the (four possible) termination conditions.

in turn is a function of unorganised action selection), encouraging selection of the "don't move" action through positive reinforcement of this choice. A second source of reinforcement for mutual learning of initial "don't move" is derived from learning the "halting" condition. Successful "halting" requires speaker selecting a "don't move" action when current focus of attention equals target focus of attention. These two factors lead to agents over generalising "don't move" and cycling (in this case, a cycle of period 1—remaining fixated on the same location) if the action is executed twice consecutively.

While this tendency to over-generalise "don't move" early on produces some unsuccessful interactions, it is also responsible for agents learning an important building block of their emerging organisation. That is, as a consequence of spending time focused upon objects, agents learn about

individual objects before they learn about moves, and they construct object words before they construct move words. The first words to appear in the emerging lexicon are the words for Object0 and Object1 (Fig. 5). A second consequence of this over-generalisation is seen at the sentence level. Later in the simulation, a robust tendency to pause exactly once at each location visited with agent focus of attention emerges. In learning not to overgeneralize the "don't move" action to all contexts, agents learn to extinguish selection of "don't move" in exactly those cases where it is inhibited—namely, when two pauses in succession lead to a cycle.

At the beginning of the simulation, speakers reach their targets and "Halt" about 5% of the time. By the end of the simulation, speaker-halting appears to be approaching 100% and nearly all of the $9 \times 9 \times 462 = 37{,}422$ possible discourse contexts result in the speaker successfully leading the listener to the intended target location. The number of total speech contexts is arrived at by considering all possible initial locations (nine), all possible target or halting locations (nine), and all possible scenes (462). Of course, since context selection is random in each interaction, there is no guarantee that every context is experienced during a simulation run.

At the beginning of the simulation, approximately 75% of interactions terminate because speaker and listener "Disagree" over a shift in attention. This is primarily due to the fact that agents are not yet organised individuals, and are acting quite randomly. By the end of the simulation, agent disagreement appears to be on the verge of extinction. During the course of the simulation, agents learn to agree on how to shift attention in roles of speaker and listener in specific discourse contexts. As speaker this learning is conditioned by the need to reach a target location, the locus of the agent's privately held intended object of discourse within the visual field. As listener, one's target location is simply the newly evidenced location of speaker's focus of attention at each step in the interaction. Although these roles are asymmetric, they constrain each other in a fashion that provides a mechanism for building internal structure that allows agents to achieve shared understanding about their external world and how to interact with it.

Below we will argue that, through the co-ordination of joint attention, the agents have created a shared system of spatial predication. In this system, language-like sentences which agents produce in interaction are reasonably viewed as instantiations of propositions which predicate spatial arrangements of objects in space within the simulated world.

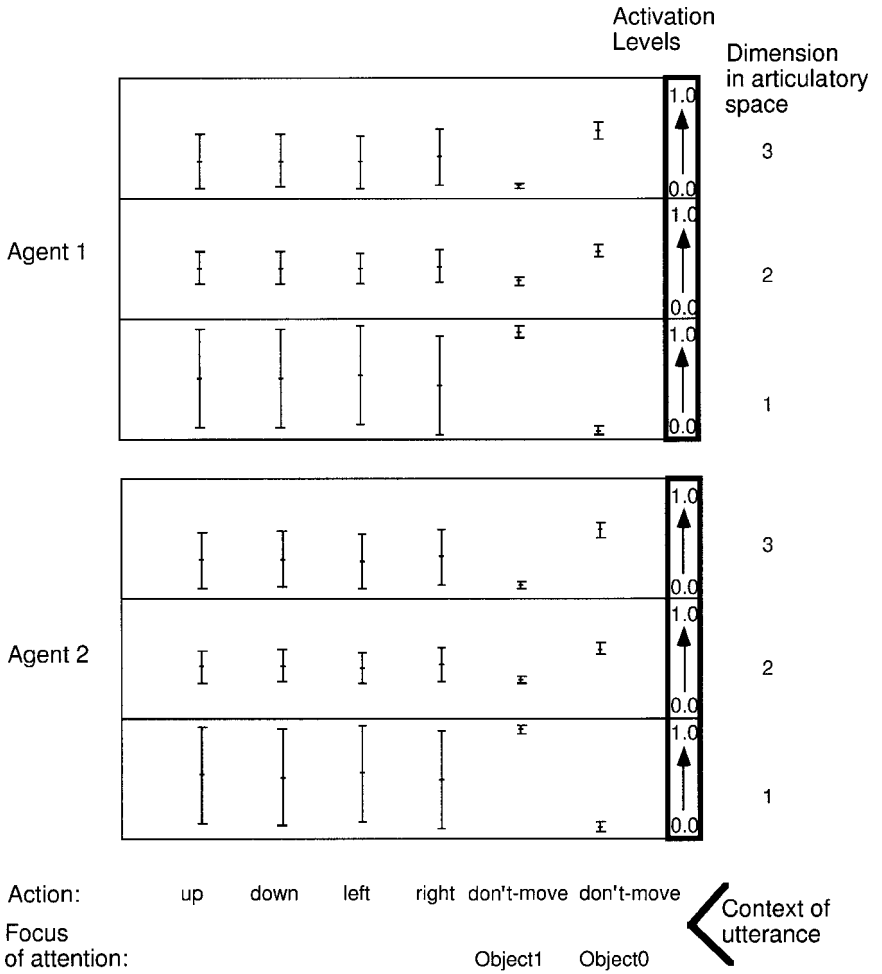

FIG. 5. Lexicon structure at $t = 10$k.

There is one plot for each agent. The data are taken at the 10,000 interaction point in the course of the simulation. The data represent each agent's production of a "test corpus" at this point in time. The data show the relation of agent sounds (encoded as three real values ranging from 0.0 to 1.0) to the context of utterance (what action the agent took and, if the action was "don't move", what the agent was focused upon). The component values of each sound are shown as mean values and standard deviation error bars represent the variation found in agent productions. Notice that agent sounds for actions remain unorganised, while sounds produced in the context of focusing upon objects constitute well-formed lexical items in the sense that: (1) There is low variability in the three components of each utterance in this context, (2) the combination of all three components of each such utterance differentiates it as a unique token in co-ordination with a unique context, and (3) these features are shared among the two agents of the population.

Evaluating the Claims: Does the Structure Developed in the Simulation Entail Propositions?

Propositions, it was claimed, can be viewed as conventional relationships among structures that represent states of the experienced world for interpreters who are members of a community of practice. Do the emergent structures of this system meet the criteria required of this definition, giving us the warrant to claim that the sequences agents construct are instantiations of propositions? The definition decomposes into three criteria, which we will employ to answer this question: (1) The constituent tokens of the sequences constructed must constitute a coherent lexicon; (2) these sequences must exhibit conventional construction or ordering of tokens; and (3) these sequences must function to predicate spatial facts about the world, for the agents in the community.

Later we evaluate the simulation, with respect to each of these three criteria, at specific points in the simulation run. At each of these evaluation points, a corpus of data was collected by having the population of agents perform as speakers in a number of novel (previously unexperienced) contexts. In fact, since agent actions are stochastic, a sample of 10 trials in each context was collected for each agent. The contexts consisted of all possible source–target pairs for novel scenes—there are $9^2 = 81$ such pairs for each scene. Each source–target pair defines starting and ending loci of attention within the visual field. Thus, for each speaker, the total number of speech trials collected was 81 (pairs) $\times$ 10 (sample size) $\times$ 10 (scenes) = 8100. The 10 scenes employed were chosen at random from the set of 50 test scenes that were set aside and never experienced by agents during the normal course of the simulation. At each evaluation point we generated a test corpus consisting of (1) all actions (speakers' shifts in attention), plus (2) all focus locations (speakers' trajectories through the visual fields), plus (3) all verbal productions (speakers' strings) for all 8100 trials and both agents. In order to simplify reference to simulation time in the discussion below, we adopt the notation $t = x$k to specify a point in time x number of thousands of interactions from the beginning of the simulation. During this testing of speakers, only two termination conditions apply: (1) Speakers "Halt" when reaching their target location and select a "don't move" action; otherwise (2) speakers quit when 24 time-steps have been taken.

The Coherence of the Lexicon

At the end of the simulation ($t = 60$k), but not at the beginning ($t = 10$k), verbal tokens produced by agents constitute a well-formed lexicon (see Figs. 5 and 6). That is, while verbal productions are at first ($t = 10$k) incapable of supporting denotation, the emergent lexicon ($t = 60$k) is capable of representing objects in the world (Object0 and Object1) and shifts in

attention (up, down, left, and right). As shown in Fig. 6, at the end of the simulation (t = 60k) the structure of the entire set of tokens agents produce reveals a lexicon in which forms are (a) in one-to-one correspondence with meanings and (b) this form–meaning structure is reliably employed across the population of agents, when each takes the role of speaker.

In the plots of Figs. 5 and 6, each word is shown in terms of the activation levels of the three connectionist units representing speakers' sound productions, in one of six contexts: Moving attention up, down, left, right, or remaining focused on Object0 or Object1, respectively. Notice that each word uniquely represents each meaning. Words are in one-to-one correspondence with the co-occurring contexts of actions and objects. Meanings of the words were determined by actions which were simultaneously produced in the world by the speaker (in the case of the actions "up", "down", "left", and "right"), or by the contents of the speaker's focus of attention (in the case of the action "don't move").

Notice that early in the simulation (t = 10k), words denoting objects have already begun to take on well-defined form, while words denoting actions are still not well-defined. As mentioned before, this is a consequence of agents over-learning the "don't move" action early in the simulation, leading them to learn the words for objects before constructing words for actions.

Error bars in the plots of Figs. 5 and 6 show the distribution of real values observed in all instances of each word produced in the test corpus by the given speaker. The error bars denote one standard deviation (in each direction) from the mean activation level for each unit. The small variance in unit activation (especially where this is necessary in order to distinguish one word from another) suggests that the form–meaning pairs are mapped one-to-one by agents, as required, and that they instantiate a relatively context-free lexicon. Comparing these plots for the two agents reveals that this structure of the lexicon is also shared across the population.

It appears that by the end of the simulation agents have indeed developed a coherent lexicon. The tokens employed by agents reliably represent states of the experienced world for them, and this mapping is shared across the population.

Conventional Sequence Construction

Given a coherent lexicon, a sequence of verbal productions can be said to be conventional if it's ordering is shared—that is, if the ordering represents a small number of agreed-upon sequences of tokens drawn from the lexicon for accomplishing some communicative task. In the simulation, we can assess the degree of conventionality in sequence construction by measuring the distribution of preferences for moving attention through the visual field

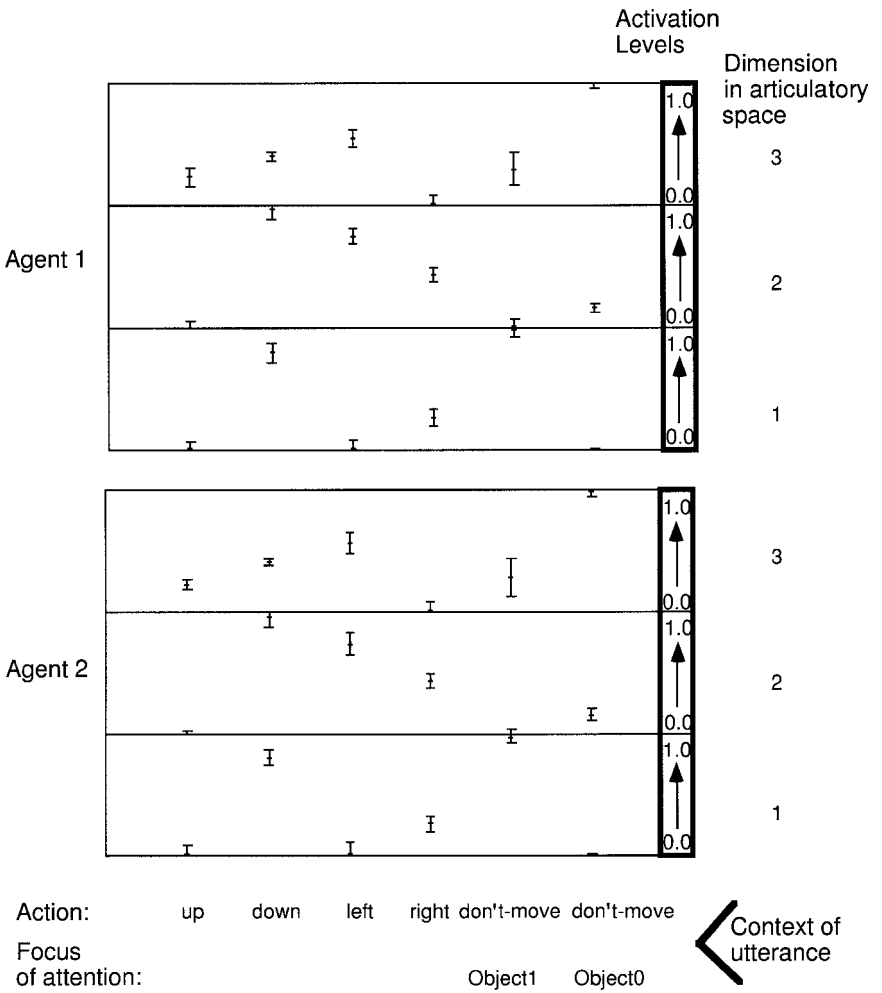

FIG. 6. Lexicon structure at $t = 60$k.

There is one plot for each agent. The data are taken at the 60,000 interaction point in the course of the simulation. The data represent each agent's production of a "test corpus" at this point in time. Notice that all agent sounds have now become organised and the repertoire constitutes a coherent lexicon in that words capture the appropriate distinctions in context and these words are shared among the agents in the population.

in a range of contexts. Out of many possible paths through the visual field, the population settles on a few preferred trajectories.

At any location in the 3×3 lattice of agents' visual fields, there are (on average) about three options for moving attention that stay within the boundary of the visual field. Given a maximum trajectory length of 24 (the maximum number of steps allowed in an interaction), there are on the order of 3^{24} different possible paths starting from a given source location. Approximately one-ninth of these will end on the desired target location. Finally, about one-fourth of these—the approximate fraction of final "don't move" among all legal final moves—will constitute trajectories which "halt" at the desired target location. But in fact, at the end of the simulation, a much smaller number of trajectories are employed by speakers as *preferred trajectories* to reach a desired target from a given source location.

The extent of sharing of preferred trajectories across the population is a measure of *conventionality*. Since we have already demonstrated that words map one-to-one onto experience of objects and agents' transitions of attention between those objects, an analysis of trajectories through the visual field transfers to claims about the ordering of tokens in verbal sequences.

Figure 7 shows the degree of conventionality in speakers' trajectories of attention through the visual field for each test corpus. The "degree of conventionality" is a measure of the randomness in the set of trajectories (pooled across agents) that successfully halt at speakers' intended target locations. Randomness in this set is measured in terms of the distribution of trajectories in the sample. Given that (for each speech context) each agent acted as speaker 10 times (yielding a pool of $10 \times 2 = 20$ trajectories), and that some subset of these successfully "halted" at speaker's intentional object in the visual field, the lack of randomness in that subset should indicate the use of convention for reaching the object in each context.

More formally, the measure of conventionality is derived from the following procedure. For each speech context $S_{i,j,k}$ in a given test corpus (starting or source focus location i and ending or target location j within scene k) there is a sample of 20 trajectories (10 created by each agent in the process of speaking). Let $C_{i,j,k}$ be the number of trajectories (out of the 20) which successfully halted at the speaker's target location (thus, $C_{i,j,k} \leqslant 20$) and let $T_{i,j,k}$ be the number of unique trajectory "types" in this set. Consider now only these $C_{i,j,k}$ trajectories for each speech context. Since there are $9 \times 9 \times 10 = 810$ speech contexts there are 810 samples $C_{i,j,k}$. We measure the degree of conventionality in the following three steps.

First, convert the observed frequencies of each unique trajectory in the samples into probabilities representing the likelihood of each trajectory

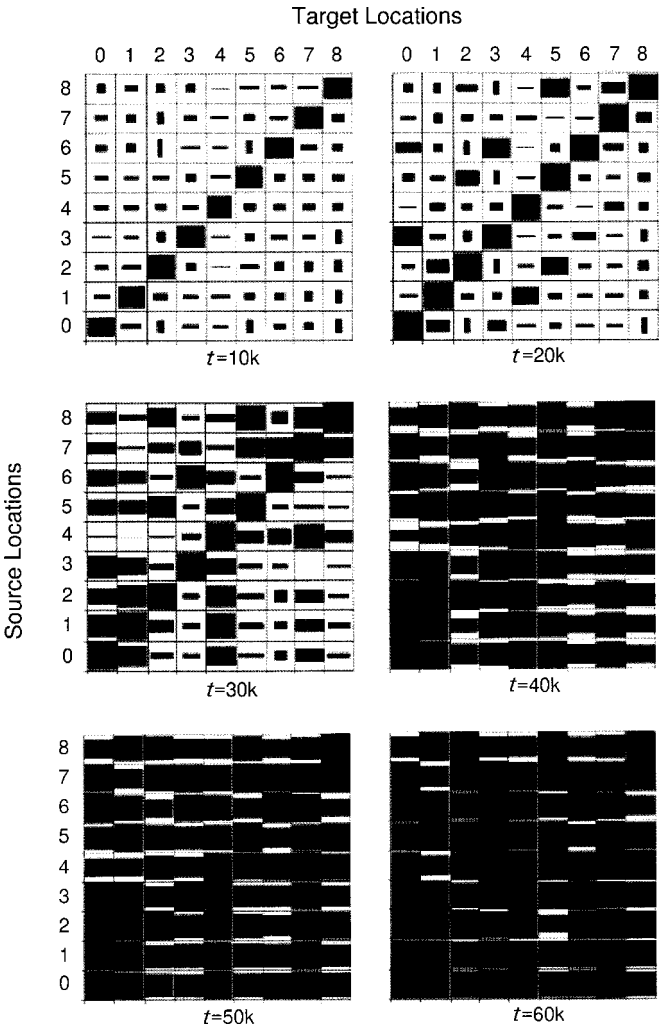

FIG. 7. The emergence of conventionality in sequential behaviour.

Each plot shows the level of successful "halting" (width of black rectangles) and level of consensus or conventionality in trajectory through the visual field (height of black rectangles) in multiple contexts, with results pooled from both agents' performances. (See main text for formulas that produce these values.) The contexts of action are given by each cell of each plot, representing a source and target pair for a given point in the simulation run. Since agent action selection is stochastic, performance in each context was sampled and then averaged across 10 different novel scenes. Since agent verbal productions map one-to-one onto the transitions entailed by an agent's action trajectories through the visual field, these data yield evidence not only about actions but also about the sequential organisation of verbal productions.

occurring in the community's repertoire. These probabilities sum to 1.0 for each sample of successfully halting trajectories. Next, apply the standard information measure to assess the structure in each set of preferences or likelihoods:

$$I_{i,j,k} = \sum_{\alpha=1}^{T_{i,j,k}} -P_\alpha \, \mathrm{Log}(P_\alpha).$$

For sets that include many different trajectories, I will be large (the elements of the set are diverse and the set is unstructured). Conversely, for sets that contain many instances of only one or two different trajectories, I will be small (there are very few different elements in the set and the set is highly structured).

Finally, since there are a number (10) of test scenes to consider, averages are computed across test scenes:

$$\bar{I}_{i,j} = \frac{\sum_{k=1}^{10} I_{i,j,k}}{10} \text{ and } \bar{C}_{i,j} = \frac{\sum_{k=1}^{10} C_{i,j,k}}{10}.$$

Thus, for each source/target location pair (i,j) we have two values measuring the observed average rate of successful halting, $\bar{C}_{i,j}$, and the observed average randomness in the sets of halting trajectories, $\bar{I}_{i,j}$. Figure 7 represents these values in the following way. Each plot is associated with a particular test corpus, collected at a particular point in time of the simulation, as labelled in the plot's title. Each cell of a plot represents a starting or source location (along y axis) and target location (along x axis) which identifies the speech context (i and j, in the earlier notation). Within each cell, a solid black rectangle is drawn according to the following dimensions. Height is scaled by the degree of conventionality in trajectories employed for the given context, averaged across scenes and normalised by the maximum across all plots, $(\bar{I}_{Max} - \bar{I}_{i,j})/\bar{I}_{Max}$. Note that $\bar{I}_{Max}$ is a consequence of agents' random initial states. These states produce (at the beginning of every simulation) trajectories that emulate random walks. Width is scaled by the degree of successful halting in trajectories employed for the given context, averaged across scenes and normalised by 20—the maximum number of halting trajectories possible ($\bar{C}_{i,j}/20$).

For the test corpus collected at t = 10k, we see that agents have already agreed upon a single trajectory in those contexts where they begin speaking at the target location (i.e. the main diagonal of each plot shows the speech context where source equals target location). This phenomenon is derived (as already mentioned) from early learning of the tendency to produce "don't move" actions. As every "halting" trajectory must entail this ability, it

is an important piece of the overall task to master early on. At $t = 20k$, this trend continues. Here we also see an emerging consensus regarding trajectories which require one shift in attention (e.g. from location 0 to 3, 1 to 4, 2 to 5, etc.) away from the source location. This feature is evidenced in the two off-centre diagonals that surround the main diagonal in the plot for the test corpus at $t = 20k$.

At $t = 30k$, things become more complicated. It is clear that speakers are quite successful at halting in nearly all contexts as shown by the width of black rectangles in each cell. However, white space in upper left and lower right regions of plot indicates a lack of consensus for those speech contexts that require the greatest distances to be travelled. This is not surprising since longer trajectories should be more variable (as a group) than shorter trajectories. A second interesting symmetry is the use of location 4, which lies at the centre of the visual field. On one hand, it appears that there is agreement about how to reach the centre (target = 4) from the top two rows (cells 0–5) in the visual field, but little agreement about how to do so from the bottom row (cells 6–8). On the other hand, trajectories that begin at the centre of the visual field (source = 4) show a lot of variability in reaching the top row (cells 0–2), and much less variability entailed in trajectories that halt on the bottom row (cells 6–8) of the visual field. This would seem to imply an overall organisation in which all trajectories pass through the center location in the downward direction, making certain paths that involve this location longer and thus inherently more variable as a group. As described later, this pattern of organised behaviour emerges by the end of the simulation.

By $t = 40k$, speakers have settled into a pattern of agreement in virtually every speech context. This pattern continues to improve slightly over the next 20,000 interactions (at $t = 50k$ and $t = 60k$), but most of this improvement is attributable to the lowering of agent temperature, which reduces the randomness injected into the agents' action selectors. By the end of the simulation ($t = 60k$), agents are in nearly complete agreement about how to employ trajectories through the visual field in order to reach intentional objects of speech with their foci of attention.

Whereas this analysis focuses on trajectories of attention and not the verbal constructions, the ordering principles of sentences co-develop with the trajectories of attention. These two structures mutually constrain each other's development so that the ordering principles of sentences predict, and are predicted by, the trajectories of attention. We are therefore confident that the agents are indeed employing conventional orderings of tokens that have arisen from the co-ordination of joint attention in discourse interactions. In fact, the ordering principles of sentences bear an iconic relationship to the trajectories of attention such that an appropriate verbal token is produced with each shift of attention. While the iconicity of this

mapping simplifies both the simulation model and the current analysis, it is a limitation not observed in natural languages. The development of non-iconic mappings between the structure of sentences and the structure of trajectories of attention is a goal for future simulations. If that were accomplished, it would be necessary to independently establish the conventionality of orderings of tokens.

Predicating Spatial Facts

While action tokens are in one-to-one correspondence with speakers' attention-shifts, and thus have been said to "refer to" those actions, there is another way to interpret the meanings of these words. In the context of verbal sequences containing references to objects in the world, and in the context of discourse which functions to guide systematic navigation between those objects, action tokens also serve to represent the spatial relations between objects, as perceived by agents (i.e. "above", "below", "left-of", and "right-of"). By virtue of representing the relations between objects, action tokens serve as "predicates" over the arguments (object tokens) contained in verbal sequences. Such sequences are propositions about space, because they function to predicate facts about objects in space.

That this "function" is in fact served by the sequences agents produce can be verified in the following way. First, collect a test corpus in the usual fashion (see earlier). Second, employ the agents as "blind listeners" in order to assess their abilities to interpret verbal sequences without access to any visual information. In the terms of agent architecture described in Fig. 3, this is accomplished by opening the agent's Chauvin Switch so that no visual inputs affect the action preference. To the extent that agents in the role of "blind listener" are able to recreate the same trajectory taken by the speaker in producing the sequence, the sequence can be said to function to predicate facts about space and the objects that occupy space.

Figure 8 shows the results of applying this procedure at a number of points in the course of the simulation. Each trace represents a measure of "success" as a function of simulation time. Success is measured in terms of a fraction of the maximum possible success rate (i.e. maximum success = 1.0). For speakers (represented in top two traces of the plot) "success" is simply the standard measure given by fraction of speaker's "halting" trajectories within the test corpus. For speaker/blind listener pairs, "success" is measured in terms of blind listeners' abilities to reconstruct the speakers' (halting) trajectories from verbal sequences only (i.e. without access to visual input).

The success of "blind listeners" in interpreting speaker's strings lags behind the success of speakers in halting at their intentional objects of speech. This order of development is expected because the organisation of sound units into meaningful words and sequences of words must follow the

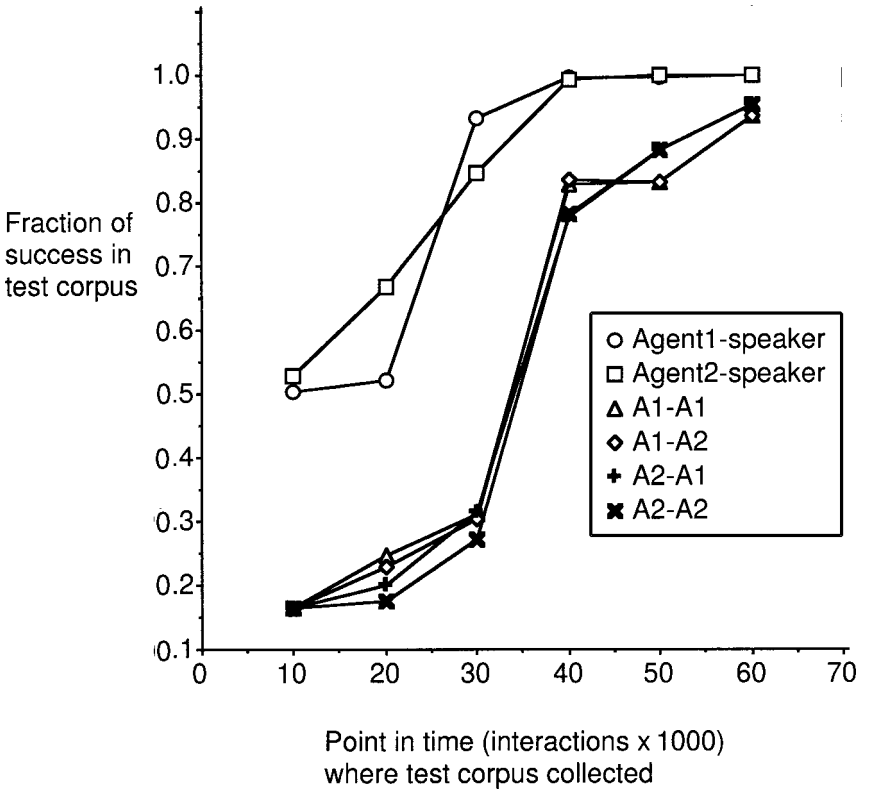

FIG. 8. The emergence of predication in verbal constructions.

This plot shows the evolution of several measures of agent performance during the simulation run. The plot shows data sampled at 10k intervals between $t = 10k$ and $t = 60k$. Each data point is the result of applying the respective performance measure to an entire "test corpus". The top two traces in the plot show degrees of successful "halting" by the two agents acting as speaker alone. The bottom four traces show degrees of successful following by agents acting as "blind listener" to a speaker's productions. (These traces are labelled with two agents (e.g. A1–A2). In each case, the first label denotes the speaker, the second label denotes the blind listener.) In the role of "blind listener" the agent must successfully parse the visual field (i.e. duplicate the speaker's trajectory) with access only to the sequence of sounds. Insofar as the capacity to do this maps onto the contents of the visual field (i.e. entails the denotation of objects there and relations between those objects) we believe that the sentences agents produce instantiate propositions.

organisation of an ability to move attention through the visual field. It is only after the co-ordinated shifts in attention settle into conventional patterns (and the mappings of internal structure to external actions become systematic) that a coherent lexicon can emerge. Furthermore, it is only after the emergence of a coherent lexicon that sequences of words can serve as descriptions for "blind" listeners of the trajectory traversed by the speaker's focus of attention during sequence production.

Over the course of the simulation, (1) a coherent lexicon emerges where there was none before, (2) the sequences composed from words in the lexicon come to have conventional construction, and (3) these sequences function to predicate spatial facts about the world, for agents. At the end of the simulation agents are using propositions about space (and objects therein) that were created in social interactions in the simple world of this simulation.

The Development of Propositions Entails Sharing Emergent Structuring Principles of the Verbal Sequences

For all of the data reviewed in the previous section, action and verbalisation take place in novel environments. Agents had no prior experience with the visual inputs employed in the test corpora reported earlier. Nonetheless, agent verbal productions are well-formed in terms of the three criteria examined. This is evidence that the agents have learned general ordering principles for the set of expressions generated by speakers. Agents share not only knowledge about a set of exemplar productions, but also knowledge of the structure that organises processing of the set.

This claim is supported by Fig. 9, which shows data from the test corpus collected at the end of the simulation (t = 60k). These plots show, for each agent, the frequency of every possible shift in attention without regard for context (i.e. where attention was previously located). The data from all 810 speech contexts (yielding 8100 trajectories per agent) have been collapsed into single state transition diagrams of attention in the visual field. These plots show the general tendency for action produced by each agent, regardless of context. Each cell of a plot depicts a location in the visual field, and represents the frequency of every action taken (for one agent) from that location. Actions are represented by line segments originating at the perimeter of a box in the centre of the cell and radiating in one of four directions (representing the actions up, down, left, and right, respectively) and by the box in the centre of the cell (representing the action "don't move"). Frequency of action is represented by line segment length (for actions up, down, left, and right), and black fill (for the action "don't move"). These magnitudes are scaled with respect to the most frequent action found in the plot.

It is apparent that agents are predominantly employing a single scheme for traversing the visual field with their foci of attention. Agents have a strong tendency to reach the lower edge of the visual field by way of the centre column, and to reach the upper edge of the visual field along either of the two outside columns. In addition, although not self-evident in Fig. 9, agents employ an alternating structure of: pause, move, pause, . . . , for all

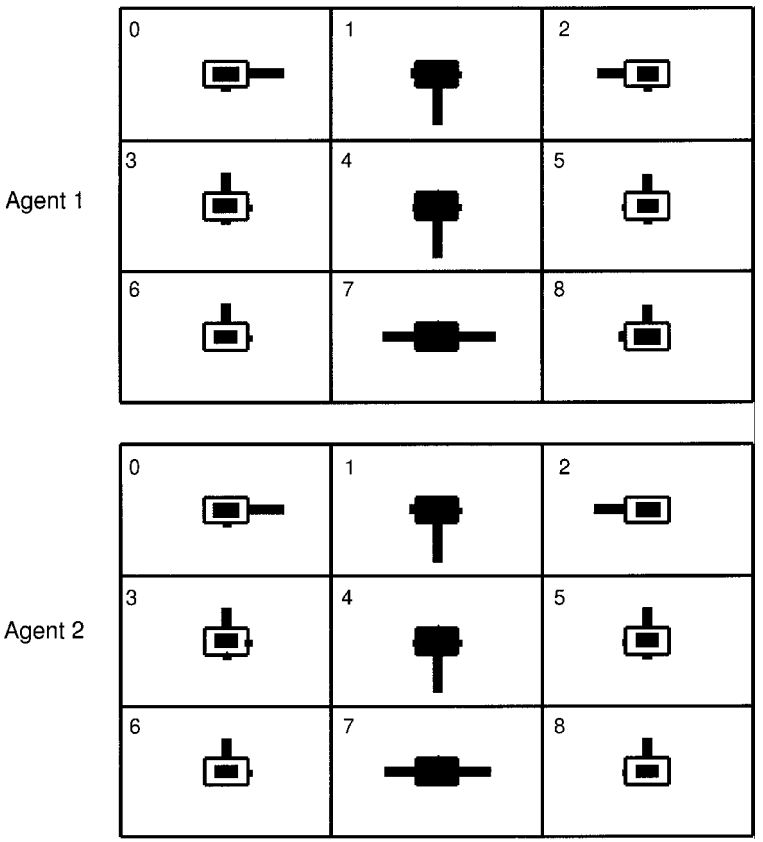

FIG. 9. The shared organisation of behaviour is rule-governed: Preferred transitions of attention through the visual field at t = 60k.

Each plot shows one agent's performance as speaker in all 8100 trials of the agent's "test corpus" after 60,000 interactions in simulation time. Each test corpus constitutes 10 trials in 810 unique and novel contexts for the agent. In each plot, the nine cell locations of the visual field are shown together with information about the agent's behaviour when attention is focused upon (i.e. finger is located in) each cell, summed across all speech contexts. The box in the middle of the cell contains information about the number of times a "don't move" action was produced. The agents preference for "pausing" is represented by the proportion of black fill within this inner box. The agent's preferences for shifting attention in some direction are represented with a length of line segment originating at the perimeter of this inner box and extending in the respective direction. Notice that the two plots, one for each agent's independent behaviours, are quite similar. This indicates sharing of the scheme for behaviour. Notice also the high degree of organisation to this scheme. Agents employ the middle column of the visual field to reach the lower row, and the outer columns to reach the upper row—in all cases. This organisation to behaviour, along with a one-to-one mapping of agent sounds onto speech contexts, allows us to write a grammar for the production of sentences.

trajectories. This latter feature of the structure is a consequence of early over-learning of "don't move" followed by inhibition of consecutive "don't move" actions later on in agent development.

With this systematic scheme, every possible target location can be reached from every possible source location without recourse to a large set of specific transition rules for each speech context. The systematic nature of this scheme demonstrates the emergence of shared general ordering principles over the set of agent expressions. These principles are not explicitly represented anywhere in this system. The principles emerged from the demands of a communication task in which agents must agree with each other over the way to shift attention in order to reach the intended objects of speakers within a problematically shared visual field.

The Shared Structural Principles can be Described as a Grammar

If a grammar is a set of rules for producing a large (even infinite) number of well-formed sentences from a small number of parts, then the agents can be said to have developed a grammar as follows. Let the grammar G include a set of terminal symbols {u,d,l,r,p}—corresponding to the tokens "up", "down", "left", "right", and "don't move", respectively—a set of non-terminal symbols {A,B,C,D,U,D^,D′,U^,U′,L,R}, the start symbol S, and the following rules for producing sentences (strings of terminal symbols) from the start symbol. In the set of production rules for the grammar shown in Fig. 10, parentheses indicate the optional introduction of a symbol into a string, and vertical bars indicate choices between complete strings.

Briefly, all sentences generated by G begin with "p" followed (optionally) by a series of one or more pairs of terminal symbols, where the first member of the pair is not "p" and the second member is "p". The legal orderings of these pairs is defined in the productions of A, B, C, D^, D′, U^, and U′. Rule B forms the central component of the grammar described here. Translating back to action in the simulated world, B defines the behaviour of agents attention shifts which originate in the centre cell of the upper row of the visual field (cell #1). Rule A gives a mechanism for reaching this location, and Rule C for completing trajectories from this location. Rule B itself provides a mechanism for recursive iteration of a loop (around either side, left or right, or a combination of both sides) in the visual field. The grammar accounts for the tendencies in agents behaviours, and should also describe agents' "competence" with verbal constructions.

To verify that agents have developed knowledge adequately described by G, the grammar was employed to generate a corpus of prescribed strings by replacing terminal symbols with corresponding words of the developed lexicon. Each instance of the symbol "p" was replaced by a random selection

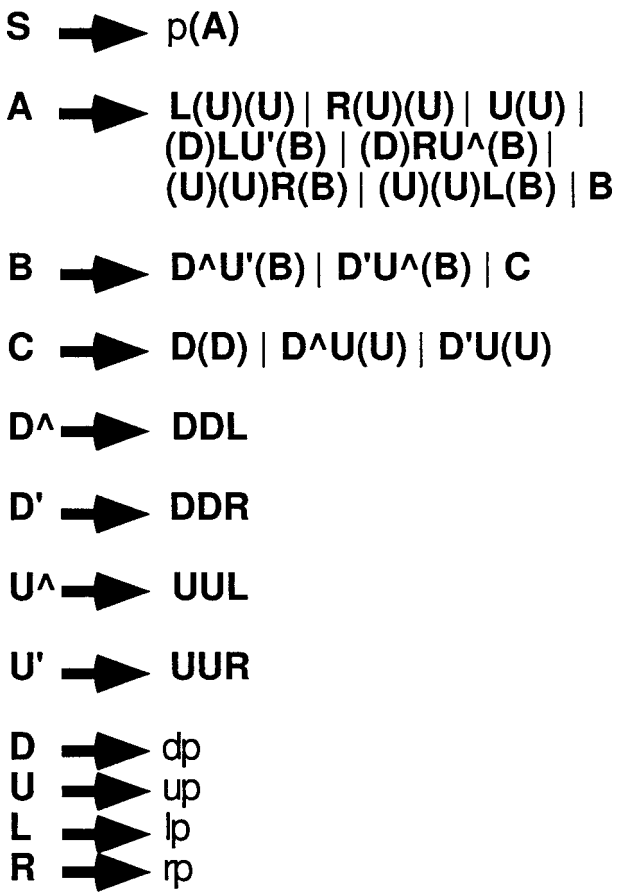

FIG. 10. Production rules for the grammar G.

between the word for Object0 and the word for Object1. All of these replacements were taken from the three-valued word representations which emerged from the simulation (t = 60k, shown in Fig. 6). This corpus was employed to test agents in the "blind listener" condition at t = 60k. Additional corpora were generated by randomly exchanging elements of each string from the original corpus (see later). *Grammaticality* was attributed to those strings which agents were capable of successfully parsing, in the sense that agents (upon "hearing" the string) reproduced the prescribed trajectory. Those strings which agents failed to parse were classified as *ungrammatical*, regardless of whether or not they were prescribed by the rules of the grammar. In all of the gramaticality testing discussed later, action selection was deterministic, meaning that the most

active output in the agent's Action Preference layer was taken to indicate the agent's interpretation of each word.

In the results described later, the corpora in question are defined as follows:

- C A corpus of 1000 strings generated from the grammar G. In generating these strings there was an equal probability of picking each production of a rule, and probabilities ranging from 0.5 to 0.75 for picking each optional symbol within a production.[1]
- M1 For each string in C, randomly exchange one token with some other token from the lexicon.
- M2 For each string in M1, randomly exchange one token with some other token from the lexicon.
- R For each string in C, generate a string of equal length formed from random selection of tokens from the lexicon.
- M1′ For each string in C, maintain the structure of alternating "p" in the string, but randomly exchange one of any other token with another token from the lexicon. The token being introduced cannot be "p" and cannot be of the same type as the token being exchanged out.
- M2′ For each string in M1′, maintain the structure of alternating "p" in the string, but randomly exchange one of any other token with another token from the lexicon. The token being introduced cannot be "p" and cannot be of the same type as the token being exchanged out.
- R' For each string in C, maintain the structure of alternating "p" in the string but, for all other tokens, randomly swap each with some other token from the lexicon.

Figures 11 and 12 show each agent's grammaticality judgments broken down by the distribution of string lengths in each corpus. The height of each bar indicates the percentage of each string-length judged grammatical, for the given corpus.

The first thing to note is that agents' grammaticality judgments are very similar, although not identical. Second, both agents are in complete agreement that the corpus derived from grammar G (corpus C) contains only grammatical strings. Of particular interest are those strings that contain multiple copies of particular sequences. More than one-half of the strings in corpus C greater in length than 12 have this sort of repetition, and all of the strings greater in length than 24 require revisiting some position in the state

[1]The exception here was that the optional symbol A in the starting rule was made non-optional. This meant that there were no strings of length 1 generated.

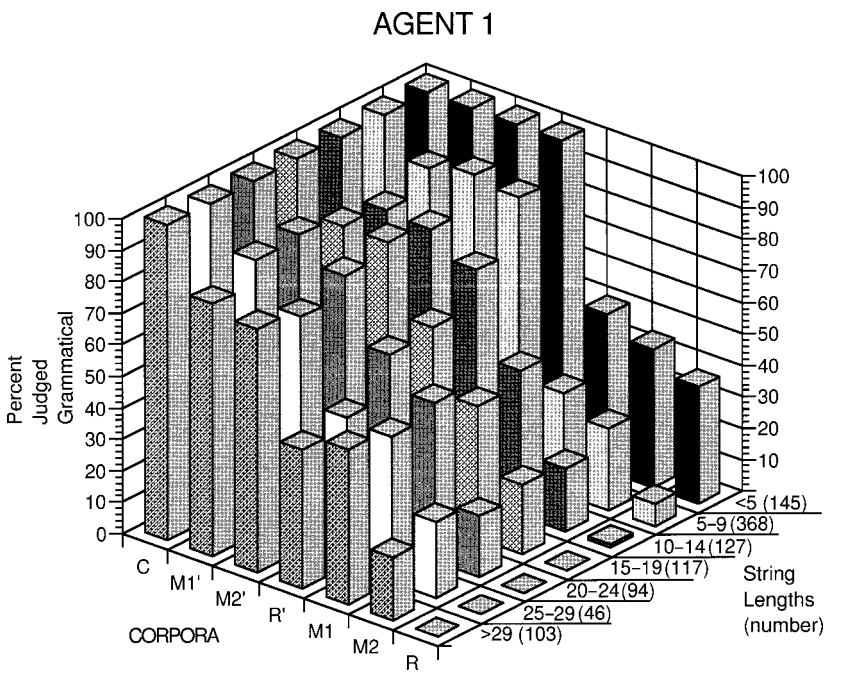

FIG. 11. Grammaticality judgements for Agent 1 on several different corpora.

Corpora were generated from concatenations of tokens drawn from the lexicon shown in Fig. 6. The method of construction differs for each corpus. Corpus C was generated from the grammar G described in the main text. The other corpora were derived from C by various manipulations of the strings in C. Grammaticality is conferred upon sentences for which all agent actions in role of "blind listener" (i.e. with access to sounds only) match those expected from the context-free mapping of token form to meaning, as shown in Fig. 6.

transition diagram of Fig. 9. Apparently, these strings do not present any particular comprehension problem for agents, suggesting that the elements of the sequences are truly being treated according to their syntactic roles in the scheme shown in Fig. 9.

Not all strings composed from tokens drawn from the lexicon are judged by agents to be grammatical. With the introduction of noise into the original corpus, judged grammaticality quickly drops, and approaches 0% in the case of randomly constructed strings of longer length (see Figs. 11 and 12, corpus R).

In between these two extremes we see a gradation in the fraction of strings judged grammatical, which is precisely what we expect when noise is introduced into the prescribed corpus (as in corpora M1′, M2′, R′, M1, M2). The data for M1′, M2′, and R′ demonstrate the significance of "alternating pause" structure in agent language. In each of these three corpora, this

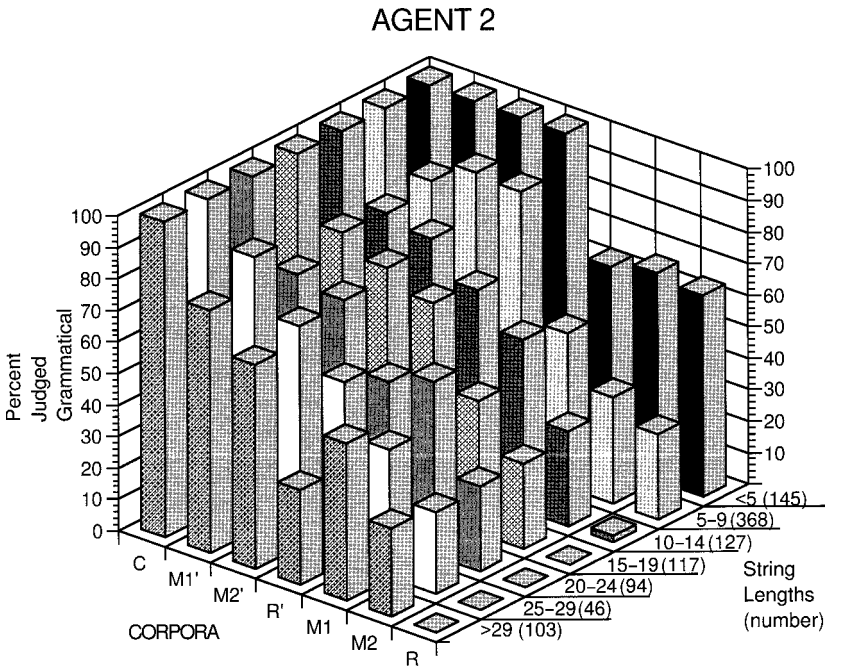

FIG. 12. Grammaticality judgements for Agent 2 on several different corpora. (See Fig. 11 and main text for descriptions of these corpora.)

structure was maintained and changes were only made on the intermediate (action–word) tokens. It should be pointed out that, although built on a random process, this noise-introducing procedure none the less creates many strings contained in G. This is especially true for short strings. This accounts for the high degree of grammaticality conferred upon short strings in these three corpora. Corpus R′ represents the maximum amount of noise that can be added to the alternating pause structure of the language. As can be seen, grammaticality decreases quite reliably with string length.

In sum, agents appear to judge all strings prescribed by G as grammatical, and the amount of syntactic noise introduced is a good predictor of ungrammaticality. This evidence leads us to conclude that G is a good description of the knowledge that agents have developed for constructing and interpreting verbal sequences.

DISCUSSION

The primary purpose of the simulation model presented here is to examine the possibility that a community of simple artificial cognitive agents could construct propositions from interactional processes. Of course, this

simulation is only useful if the elements of the model map onto phenomena in the natural world in a believable, even if simplified, way. In the discussion we first summarise the simulation model. Then we briefly consider an empirical study that addresses related phenomena in the real world and suggest ways in which our model is compatible with that study. Finally, we review some theoretical issues and suggest a modified view of the origins of human symbol processing behaviour.

The agents in the simulation are composed of connectionist network modules that produce functional properties that model simple perceptual, motor, social, and verbal capacities. The agents encounter each other in social interactions requiring co-ordinated action in a problematically shared world of visual perception. The development of inter- and intra-agent co-ordination of shifts in visual attention lead to agents sharing understanding about the spatial arrangements of objects in their world. Propositions emerge from the organisation of agent behaviour around agent-generated verbal structures that describe this co-ordination. Ultimately, these verbal structures can be taken to stand in for the co-ordinated activity independent of the external visual stimuli, motor, and social activity from which they arose.

The words constructed and employed by agents as constituents of verbal sequences differentiate into two types, representing objects and actions, respectively. The object-type tokens develop and come to represent objects in the world. As agents build internal structure that reliably maps these constructed tokens onto visual perception of those objects, object-type tokens are grounded in agent perception of objects located within a momentarily fixated focus of attention within a larger visual field. The action-type tokens come to represent shifts in agent focus of attention as agents build internal structure that reliably maps such tokens onto motor commands for carrying out those actions. Action-type tokens result from speakers generating external structure that becomes co-ordinated with the internal structure developed to control the dynamics of attention within a visual field. Through co-ordination of the use of these public structures, the population comes to share the form–meaning mappings and the ordering principles of sentence construction that affords predication of spatial properties—that is, the community invents propositions.

Building Propositions Through the Co-ordination of Joint Attention

Several lines of empirical research that take a broadly functionalist perspective on language acquisition are compatible with the model presented here (cf. Bates, Benigni, Bretherton, Camaioni, & Volterra 1979; Bruner, 1974, 1976; Ochs, Scheifflen, & Platt 1979; Tomasello, 1988;

Tomasello & Todd, 1983). All of these studies begin with a predisposition for seeing the prelinguistic child as an active participant in a highly structured world of communicative activity. Furthermore, this interaction, as well as developing language skills, are seen to be fundamentally coupled to the dynamics of attention management.

A study by Ochs et al. (1979) concerning the development of propositions in young children is relevant. The authors construct a model from observations of child–child and child–adult discourse which demonstrates the learning of propositional structure through "vertical construction". That is, propositions are first seen to be distributed across speakers (vertically in the written transcript) and only later become mastered by the child as "horizontal" units that are complete within a single discourse turn.

Ochs and colleagues provide an empirical account of an incremental learning process whose end point, but not starting point, is sentential propositions. Their account argues that propositions don't originate in sentential forms; they come in much smaller pieces of attention management, which must be cobbled together in the service of communication. These pieces are put into place by interactive processes that are distributed across participants and across time. Only later, as a consequence of having participated in this co-ordinated activity, is the child able to master propositions as complete sentences. Furthermore, this account applies equally well to child–child interactions as to adult–child interactions. This suggests that the phenomenon is not simply a pedagogical strategy employed by adults, nor the consequence of asymmetry in knowledge between interactants. Rather, it suggests that this phenomenon constitutes a robust mechanism for generating propositional structure from generic human cognitive and social skills. Learning propositions as a novice in a community of speakers who have already mastered them and having a community learn to produce propositions from scratch are clearly different problems. Still, in both cases, the origins of propositional structure appear to be outside the individuals first, and only later inside them.

Must Language be Grounded in Innate Properties of Mind?

The simulation demonstrates that agent–agent and agent–world interactions provide the basis for a solution to the problem of where propositions come from. It is clear that the world alone cannot provide all of the structure necessary for cognitive agents to reach agreement on abstract relational predicates such as those involved in "Above[ball, table]" and "Below[table, ball]". Either of these (and a myriad other) descriptions could be appropriate. Building these propositions requires a decomposition of the structure in the world into parts (ball and table) which have relations (ball

"on top of" table, and table "underneath" ball). If not given by structure in the world, where might predicates (and the propositions constructed on these predicates) come from?

One answer to the question of where such predicates come from is to stipulate that they are innate in the human mind. This explanation has been offered, in various forms and guises, throughout Western history (e.g. Plato, Aristotle, Augustine, Descartes, Fodor). The most recent influential versions of this doctrine—actually, a consequence of a generative theory of language—are attributed to Chomsky (1957, 1965) and Fodor (1975). In its barest outlines the argument these scholars put forward is as follows: (1) Thought and language (psychological processes) are inherently productive, i.e. all humans are capable of grasping an infinite number of thoughts and an infinite number of well-formed sentences; but (2) this is accomplished with finite resources, i.e. with limited experience of structure in the world and access to only finite innate structure in the brain; therefore, (3) these resources must be employed in a generative process that constitutes human language and thought.

From this argument it follows, according to this tradition, that the concern of cognitive science should lie in understanding how mental representations enter into computational processes which generate idealised behaviour. For Fodor, this means positing a "language of thought", which is seen as a self-interpreting formal system constituting an engine for hypothesis confirmation. The elementary representations employed in this system are intrinsically meaningful within it because they are innately given concepts. And they must be innate, for, if they are learned, require interpretation, or are simply stand-ins for something else, then an infinite number of such systems is implied and the work of cognition (i.e. computation over well-formed symbolic constructs) will never get done. Such innate concepts are taken to be "basic encodings", because they are imagined to ultimately ground the recursive "stand-in" relations required by the generative stance on mental representations. Thus, language acquisition is a process of employing basic encodings provided by the language of thought to deduce the nature of language in terms of the truth conditions which classify constituents of the language.

> Learning a language (including, of course, a first language) involves learning what the predicates of the language mean. Learning what the predicates of a language mean involves learning a determination of the extension of these predicates. Learning a determination of the extension of the predicates involves learning that they fall under certain rules (i.e., truth rules).
>
> (Fodor, 1975, 63–64)

> If learning a language is literally a matter of making and confirming hypotheses about truth conditions associated with its predicates, then learning a language

> presupposes the ability to use expressions coextensive with each of the elementary predicates of the language being learned.
>
> (Fodor, 1975, 80)

So construed, the language of thought resembles a body of knowledge (including, importantly, a set of intrinsically meaningful representations or basic encodings) structured so as to enable formal syntactic manipulation (computation), making possible the learning of any language through processes involving hypothesis confirmation of truth conditions. Fodor's answer to the question of where predicates such as "Above" and "Below" come from is that they are innate to the human mind insofar as "coextensive expressions" can be found in the agent's language of thought prior to any communicative act or activity in the world.

The solution proposed in this article is that these predicates are grounded in agents' experience of contextualised actions in the world. Their co-ordinated and conventionalised shifts in visual attention stand as proxies for the abstract relations among objects in the world. Agents come equipped with an innate *capacity* to formulate predicates, but the development of predicates in the simulation is a product of communicative events. There are many possible outcomes to this process, which depend upon a number of mechanisms not localised within any individual. The contrast with Fodor's account is nicely captured in Christina Erneling's (1993, pp. 117–118) description of a Wittgensteinian position on language acquisition.

> While Fodor says that children generate language according to innate mental structures until they reach agreement with the language spoken around them by forming and confirming hypotheses, a Wittgenstein-inspired approach would say that children generate language by moderating innate behaviors until they reach agreement with the language spoken around them. This is accomplished by training. So, just as with Fodor, [the Wittgensteinian] approach takes something innate as the starting point of learning, but unlike Fodor it is not innate language of thought or linguistic competence, but specific behaviors and natural expressions (shared "biological" forms of life). These are not translated into their public linguistic counterparts but are replaced, extended, and refined by the encounter with public language. Something generally new is acquired, not only externalization of something already present, either explicit or implicit, in the mind.

Agents in the simulation do have innate properties. Agents are predisposed to produce verbal structure in co-ordination with action, and are endowed with information processing mechanisms that guarantee effective cross-modal integration of information between vision and action, including acts of verbal production. But these capacities cannot reasonably be construed as an endowment of innate concepts which simply surface in

the emergent system of communication. Different simulations yield different systems of communication. The surface forms of lexical items are contingent upon a particular interactional history and starting point. Sequence ordering principles are constructed conventional arrangements, and the sets of predicates developed in each simulation are not guaranteed to coincide. Simulations have been observed, for instance, where some action (such as "up") never materialises in the community's behavioural repertoire. This, of course, yields a language without the corresponding action word.

We believe that our simulation demonstrates an artificial world in which agent "biology" puts in place information processing mechanisms which enable agents to enter into a process of language construction. This construction process involves the co-ordination of communicative functions in the world and is therefore subject to the problems that accompany the establishment of behavioural conventions. Once such conventions have been formulated, they need not be recreated in each generation; nor is it necessary for such conventions to become incorporated into the innate biological machinery. Rather, through public use, these conventions can serve as templates for learning by novices, making learning easier (and traditions resistant to change) without agents having been programmed by evolution to produce them (cf. Freyd, 1983, 1990; Hutchins & Hazlehurst, 1991, 1995).

While the simulation described here focused on propositions as *products of co-ordinated activity*, propositions also serve simultaneously as *co-ordinators of activity*. This dual nature of human symbol systems is often labelled with the term "culture," and it gives humans cognitive abilities unknown in any other animal (cf. Bruner, 1972; Geertz, 1962; Tomasello, Kruger, & Ratner, 1992).

The Nature of Representations and the Social Constitution of Mind

We hope to have demonstrated the plausibility of an alternative basis for understanding the origins of cognitive representations. In this model, structures internal to agents represent in virtue of their role in a particular kind of process. These processes feed each other in chains of functional dependence radiating inward (from the skin) to mechanisms of physiology and outward (from the skin) to the mechanisms of social living. These processes depend upon properties that reside not only in agent biology but also with socially and historically constituted environments.

It may be that the generative tradition's treatment of representations (what Bickhard, 1987, calls "encodingism") can be rescued by an evolutionary story that puts basic encodings in human brains where they serve encoding and decoding functions for the language of thought. Bickhard (1987; Bickhard & Terveen, 1995) doesn't believe such rescue is

possible, and even if possible leaves encodingism epistemologically bankrupt. First, as mentioned previously, encodingists are left claiming that representations "represent that which they represent". Second, if evolution could deliver basic encodings then there is no reason (in principle) why basic encodings could not be delivered by some other process such as social interaction, for instance. This leaves encodingists with no logically necessary reason to resort to natural selection in the first place.

This point applies not only to the generative tradition, but to the current vogue in cognitive science of resorting to natural selection or biological reductionism to "explain" human behavioural phenomena. We too seek a naturalistic account of mind, but we see no use in rejecting dualism of mind and body (which underlies the generative tradition, for instance) only to embrace an exclusive concern for the body (which underlies many modern rejections of the generative tradition). We see, instead, the need to reconstitute mind in the dynamics of a system of bodies coupled via socio-historical processes.

The revised account does not assume "encodingism" and thus has no need of an evolutionary story to explain the origins of "basic encodings". Stripped of the computational machinery to which the ontology of internal representations have been bound in the generative view of mind, the nature of information processing in the service of cognition looks very different. In the revised view, internal representations do not take on epistemic properties because they provide—through encodings of the world—content for internal symbolic manipulation. Rather, internal representations take on epistemic properties because of the role they play in the processes that organise behaviour in the world. These processes simultaneously extend "outward" to the cultural context and "inward" to the body which must act in that context.

Can This View Account for Intelligence?

It is generally agreed that human intelligence is constituted by a set of properties which is dominated by those known, respectively, as "compositionality", and "systematicity" (cf. Fodor & Pylyshyn, 1988). Briefly, compositionality refers to the ability to generate complex structures (thoughts and sentences) from simpler parts already available. Systematicity refers to the idea that processing some structures (e.g. John loves Mary) necessarily implies the ability to process other related structures: (e.g. Mary loves John).[2] These properties are entailed by the argument for productivity

[2]This "necessity" is imagined to follow from at least two sources. (1) The constituents of sentences (e.g. John, Mary) are members of syntactic classes (e.g. NP) which determine grammaticality and linguistic competence, and (2) form–meaning pairings are context-free or independent, such that swapping constituents of the same syntactic class is computable in virtue of knowing the syntax.

which serves as the epistemological foundation for the generative view of mind. Both compositionality and systematicity derive from the belief that cognitive processes must be productive and must treat representations strictly according to their syntactic roles in a formal, rule-following, computational system.

The simulation allows us to pose the following question. Must compositionality and systematicity be built into the machinery of mental processing, as required by the generative position? Or might the mind construct these properties as products of participation in a cultural process?

The sentences that emerge in the repertoire of the community of agents in the simulation are highly "compositional", in the sense that most sentences serve as building blocks for many other sentences. This is seen in the system of trajectories developed for traversing the visual field. As shown in Fig. 9 a single system of routes through the visual field is used for the entire set of speech contexts. This yields an enormous overlap in the forms employed in verbal constructions.

Furthermore, the repertoire of constructions exhibit "systematicity" in the sense that the form-meaning pairings are context-free and the preferred orderings of constituents are rule-governed. The observed forms are in one-to-one correspondence with meanings regardless of context within a sequence. Constituent ordering is rule-governed in the sense that the same structural principles are employed to construct all sequences from constituents. Regardless of where the trajectory begins or ends in visual space, the same path (including pausing on every other time-step) is employed in all contexts. This makes it possible to write a simple set of rules that adequately describes the agents' abilities to produce and comprehend sentences.

This kind of structure did not develop in simulations consisting of just one individual. When acting alone, speakers are not constrained to shape their behaviours to match those of the partners. Rather than developing a systematic solution that emerges under the co-ordination of action between agents, the individual learns a much more varied, context sensitive, set of trajectories. Long trajectories in this simulation condition tend to be composed out of established shorter trajectories to a much lesser extent than was the case in co-ordinated learning. As a result, when we overlay all of the trajectories that the individual employs, as in Fig. 13(a), we don't see a globally consistent set of state transitions. This is because transitions in the individual learning condition are governed more by context (where one came from, what one is looking at, and where one is trying to get) than by the need to match one's own behaviour to that of another agent.

In the individual learning condition, there is delayed commitment to particular solutions and agents learn the visual field as a true "field". Due to

the stochastic nature of action selection, locations in the visual field are learned as loci on a gradient. The gradient is based on distance to target location (as expected of the teaching function, which employs such a distance metric) without the added constraint of developing conventional trajectories for the shifts in attention. This is seen clearly in Fig. 13(b), which decomposes the data of the top plot into the nine different target locations involved in the corpus. Each major square in the diagram represents a single target location. We see a tendency for actions taken from each position in the visual field to follow the gradient established by distance to target.

This means that there are often several equally probable action choices available to the agent on each time step. This, in turn, means that verbal productions are "noisy" in the sense that they don't reliably predict actions. If we tried to write a set of rules to describe agent trajectories or sentences (as we did earlier for the multi-agent simulation), the number of such rules would approach the total number of speech contexts.

By contrast, in the multi-agent simulation compositionality is a product of demand for co-ordination in the face of only partial sharing of mental states by agents in interaction. During an interaction, the listener employs the speaker's actions in the world as evidence about the speaker's target. The listener doesn't have direct access to the speaker's intention and must treat the speaker's action as its own target for emulation. Furthermore, speakers are constrained by the need to meet listener expectations, and must themselves be listeners half of the time. It appears that a systematic set of trajectories provides a good solution to this co-ordination problem. As the system evolves, speakers learn to produce sequences of action that listeners learn to expect.

This functionality is similar to what happens in the internals of a standard auto-associator network. In an auto-associator, the weights between the input and hidden layer learn to produce patterns of activation at the hidden layer that the weights between the hidden and output layers can learn to decode as the original input. This simultaneous learning of an encoding function and a decoding function that are shaped by each other to fit each other produces the well-known efficient encodings of the auto-associator network. In our simulation, this process is not contained within an individual. Rather it is a property of the interactions among individuals in the community. The simultaneous needs to learn to talk in a way that others can understand, and to learn to understand the way others talk provides the additional constraints that cause the emergence of structure from these interactions.

CONCLUSION

This article describes processes that occur in a community of interacting agents. In these processes, the co-ordination of joint attention leads to the

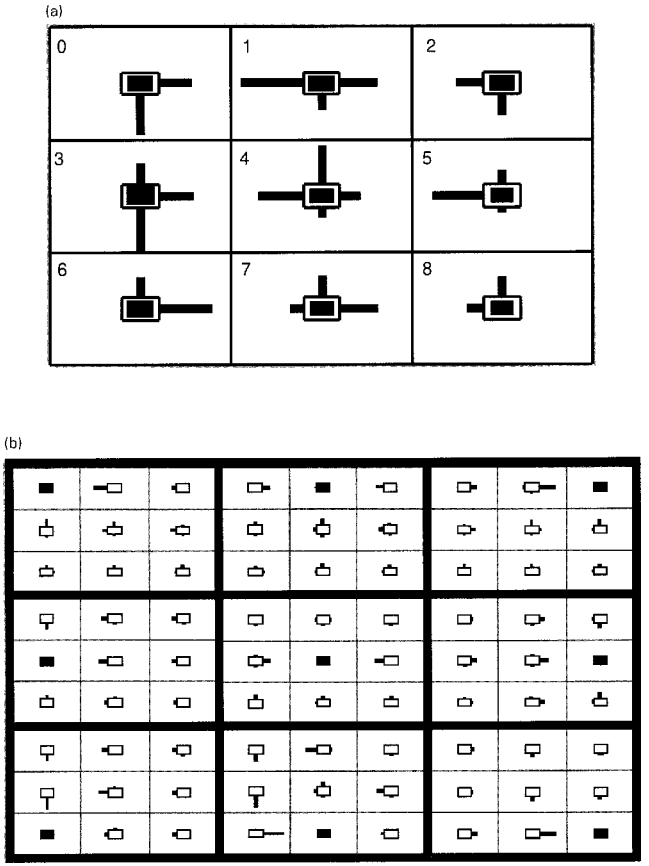

FIG. 13. The organisation of behaviour for individual learning condition: Preferred transitions of attention through the visual field at t = 60k.

In contrast to Fig. 9, agents who learn the task as individuals do not generate rule-governed behaviour. In this "individual learning condition" there is no negotiation of how to act. This means that there is no consensus aspect to the reinforcement of action selections, and agents learn the visual field as a true "field" in which distance to target defines (in each context of action) a gradient that agents come to internalise. The first plot (a) shows the identical plot as Fig. 9 but now for an individual agent that learned in the individual learning condition. Notice that the overlaying of performance in all of the 8100 trials (entailing 810 novel contexts) does not demonstrate a systematic scheme for parsing the visual field with attention. Instead, agents in this learning condition appear to learn the problem by internalising the gradient representing distance to target in all contexts. This is shown quite clearly in (b), which decomposes the overlaid data of (a) into contexts defined by target location. Each major cell now defines the target location involved, and the minor cells represent miniature plots showing transition of attention in each of these contexts. The notion that agent behaviour is organised by the distance gradient can be seen by noticing that distance to target predicts agent preference for movement of attention in nearly every case.

development of structures (both internal and external) that support organised behaviour. We have argued that the simulation model demonstrates the plausibility of propositions arising from such processes. Furthermore, we have argued that similar phenomena are at work in empirical studies addressing language acquisition.

The traditional treatment of the origins of propositions invokes innate concepts that ground the representations of a formal system of computation that is entirely internal to the individual. The ontology of such a system is inferred from certain properties which have been taken to be the most important aspects of cognition: Productivity, compositionality, and systematicity.

Our model presents an alternative account of the origins of propositions. In our simulation agent "biology" includes information processing mechanisms that enable agents to enter into a process of language (or language-like) construction. The language construction process in our model requires the co-ordination of communicative functions in a shared world. The development of such a system puts in place resources for the organisation of behaviour that are normally attributed to the internal mechanisms of individual brains. This process is capable of yielding several of the cognitive properties that are widely accepted to be indicative of intelligence.

In particular, we were able to demonstrate the emergence of a simple rule-described organisation in agent behaviour and language-like structures that mediate agent behaviour. Individuals who were members of a community of practice developed internal organisation reflecting the (apparently) rule-based co-ordination of behaviour with other agents. Individual agents acting alone were unable to generate similar organisation through direct engagement of the world.

It appears, then, under the assumptions of our model, that co-ordination of action in the world can lead to the development of compositionality and systematicity in the structures (both internal and external) that organize agent behaviour. We believe this finding reflects the facts of human existence. Humans live in worlds where the organisation of behaviour through social interaction and use of historically derived public structure, is pervasive. We also believe that the traditional approach to cognitive science has failed to consider how these facts may account for fundamental properties of human cognition. This article represents one attempt to address this shortcoming.

We do not believe that this is a model of the origins of human language. Our simulations are obviously extremely impoverished computational systems compared to the richness of human evolutionary and cultural experience. What we do claim is that any argument which maintains that the origins of symbolic behaviour (or, in our model, conventional sequences of

meaningful tokens) must lie in innate properties of the brain, is wrong. There is another possible source of such structure: Repeated interactions among the members of a community in a shared world of action. This provides an explanation that plausibly fits with what we know about the social nature of our species.

REFERENCES

Bates, E., Benigni, L., Bretherton, I., Camaioni, L., & Volterra, V. (1979). *The emergence of symbols: Cognition and communication in infancy*. New York: Academic Press.

Bickhard, M. (1987). The social nature of the functional nature of language. In M. Hickman (Ed.), *Social and functional approaches to language and thought*. Orlando, FL: Academic Press.

Bickhard, M., & Terveen, L. (1995). *Foundational issues in artificial intelligence and cognitive science: Impasses and solution*. Amsterdam: North Holland.

Bruner, J. (1972). Nature and the uses of immaturity. *American Psychologist*, *27*(8), 1–22.

Bruner, J. (1974). The organization of early skilled action. In M. Richards (Ed.), *The integration of a child into a social world*. Cambridge, UK: Cambridge University Press.

Bruner, J. (1976). Peekaboo and the learning of rule structures. In J. Bruner, A. Jolly, & K. Silva (Eds), *Play: Its role in development and evolution*. Harmondsworth: Penguin.

Chauvin, Y. (1988). Symbol acquisition in humans and neural (PDP) networks. Unpublished doctoral dissertation, University of California, San Diego, CA.

Chomsky, N. (1957). *Syntactic structures*. The Hague, The Netherlands: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Erneling, C. (1993). *Understanding language acquisition*. Albany, NY: SUNY Press.

Fodor, J. (1975). *The language of thought*. Cambridge, MA: Harvard Press.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. In S. Pinker & J. Mehler (Eds), *Connections and symbols*. Cambridge, MA: MIT Press. (A reprinted special issue of the journal *Cognition*).

Freyd, J. (1983). Shareability: The social psychology of epistemology. *Cognitive Science*, *7*, 191–120.

Freyd, J. (1990). Natural selection or shareability? *Behavioral and Brain Sciences*, *13*(4), 732–734. (Comment on "Natural language and natural selection" by S. Pinker & P. Bloom).

Geertz, C. (1962). The growth of culture and the evolution of mind. In J. Scher (Ed.), *Theories of the mind*. New York: Free Press of Glencoe.

Hutchins, E., & Hazlehurst, B. (1991). Learning in the cultural process. In C. Langton, C. Taylor, D. Farmer, & S. Rasmussen (Eds), *Artificial life, Vol. II. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. X*. Redwood City, CA: Addison-Wesley.

Hutchins, E., & Hazlehurst, B. (1995). How to invent a lexicon: The development of shared symbols in interaction. In N. Gilbert & R. Conte (Eds), *Artificial societies: The computer simulation of social life*. London: UCL Press.

Ochs, E., Scheifflen, B., & Platt, M. (1979). Propositions across utterances and speakers. In E. Ochs & B. Scheifflen (Eds), *Developmental pragmatics*. New York: Academic Press.

Prior, A.N. (1976). *The doctrine of propositions and terms*. London: Duckworth & Co.

Tomasello, M. (1988). The role of joint attentional processes in early language development. *Language Sciences*, *10*(1), 69–88.

Tomasello, M., Kruger, A., & Ratner, H. (1993). Cultural learning. *Brain and Behavioral Science*, *16*, 495–511.

Tomasello, M., & Todd, J. (1983). Joint attention and lexical acquisition style. *First Language*, *4,* 197–212.