Chapter 10

Auto-Organisation and Emergence of Shared Language Structure

Edwin Hutchins and Brian Hazlehurst

The principal goal of attempts to construct computational models of the emergence of language is to shed light on the kinds of processes that may have led to the development of such phenomena as shared lexicons and grammars in the history of the human species. Researchers who attempt to model the emergence of lexicons make a set of shared assumptions about the nature of the problem to be solved. First, there are constraints on what counts as a shared lexicon. A *lexicon* is a systematic set of associations (a mapping) between *forms* and *meanings*. Forms are patterns. *Tokens* of a form are physical structures that bear the pattern of a particular form. For example, words are forms in this sense. Each instance of a particular word is a token of that word because it bears the pattern (sequence of sounds or letters) of that word. Forms must be discriminable from one another. *Meanings* are generally taken to be mental structures which, on the one hand, shape agents' interactions with a world of objects and, on the other hand, also shape agents' interactions with forms.

A lexicon is said to be *shared* when the members of a community adopt similar forms, meanings, and the mapping between these two elements. This is a requirement for the communication of meanings via forms. A shared lexicon is thus a systematic set of form-meaning mappings in which the forms are discriminable, the mappings are (roughly) one-to-one, and the set of associations between forms and meanings is shared by members of a community. The mappings are roughly one-to-one because synonyms (two or more forms for a single meaning) and homonyms (two or more meanings for a single form) are possible but do not dominate the mappings. The lexicons of natural languages can be described by these properties (among others).

The emergence of a shared grammar presents a more complex problem. Grammar refers to properties of language involving sequences of lexical forms.

These sequences are called expressions or sentences. A grammar implies constraints on the internal organizations of expressions. Grammatical expressions of a given language constitute a subset of the universe of possible sequences of forms drawn from the lexicon. This property of grammar is called systematicity and is often characterized by reference to the structure of expressions. For example, in English, structure is evidenced in a class of words called "noun" which includes the member forms "John" and "Mary." Language is systematic because the members of the set of allowable sequences share patterns (e.g., NVN → John Loves Mary, Mary Loves John, Marry Hates John, etc.). Looked at from the standpoint of language production, a grammar enables complex expressions to be easily built from simpler parts. The meaning of a grammatical expression depends on both the meanings of the lexical forms from which the expression is composed, and on the relations among the forms in the expression. For example, the sentences "John loves Mary." and "Mary loves John." contain the same lexical forms, but have different meanings because the forms bear different relations to one another in the two sentences. This property of grammar is called compositionality. Finally, it should be possible to create novel meanings by composing new expressions from the available set of forms. This property of grammar is called generativity. A grammar is said to be shared when the members of a community adopt similar systems for composing expressions, including similar mappings between expressions and meanings.

Using definitions of shared lexicon and shared grammar such as these, researchers ask, "What sort of process could lead to the development of a shared language?" Clearly, some historical process led human ancestors from the condition in which there was no shared language to the condition in which a shared language exists. It is assumed that language, and many other aspects of culture, develop without any central control. That is, there could not have been a "teacher" who knew the language first and then taught it to others. Rather, a shared language should be expected to emerge somehow from the interactions among the members of a community who must communicate. Since we have no direct access to the historical events that led to the development of language, a common strategy for addressing this question is to construct a computational simulation model. Such a model begins in a state in which a shared language clearly does not exist. The model is then run and eventually reaches a state in which a shared language does exist.

In our discussion of models addressing the emergence of language, we will attempt to clarify the different stances that modelers take regarding the elements of language and the relations among those elements. Figure 1 depicts the elements of language (in boxes), and their relations (connections among boxes), as exemplified in the simulation models that address the emergence of lexicon. The question marks (?) indicate relations that are treated differently by different simulation models. In addition to these components, each model also specifies processes that bring agents into interaction with one another. In a successful simulation, language-like structures emerge from these interactions.

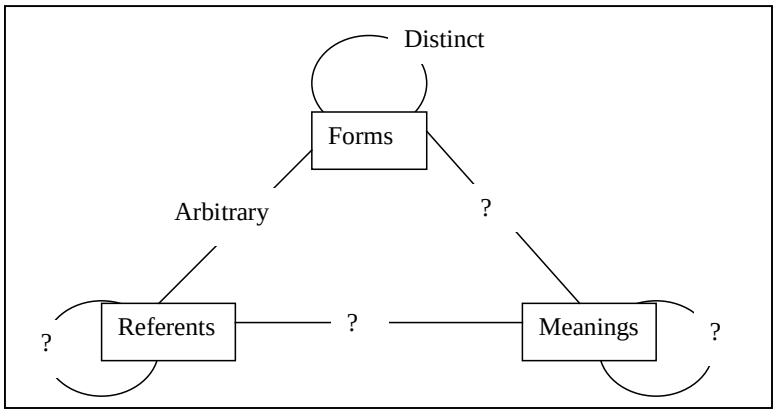

Figure 1 - Elements of language (in boxes) and their relations (connections among boxes) in the simulation models on the emergence of lexicon.

In a computational simulation of the emergence of a shared lexicon, a community of virtual agents is created. Each agent is capable of implementing a form-meaning mapping. In the initial condition of the simulation, no systematic form-meaning pattern exists. The simulation will include a model of the interactions among agents, such that agents can be changed by the experience of interaction. This interaction protocol determines the organization of the interactions among agents and between agents and their simulated world. As the simulation runs, and as virtual agents are changed by repeated interactions with meanings and forms, and possibly with objects in the world and other agents, a shared lexicon *emerges*. The lexicon is not explicitly specified in advance and there is no teacher in the system telling the agents how to construct the mapping that will become shared. The term *auto-organization* refers to this process of emergence in which the community of agents organizes itself over time.

Auto-organization is not magical. New patterns result from the organized interaction of patterns that are present in the initial conditions of the model or are present in the history of interaction itself. All of the models discussed in this chapter operate on a principle of modulated positive feedback. Modulated positive feedback is positive feedback with a resonant filter that favors some signals in the loop and causes others to dissipate. This principle underlies many kinds of auto-organizing processes including those that produce bio-convection and many other animal built structures (Turner, 2000). In the models discussed here some kinds of structure are simply given or assumed by fiat, some random process interacts with the assumed structure to produce small initial differences and similarities, and then a modulated positive feedback loop operates to amplify the initial differences and similarities. When we consider the significance of the models with respect to the goal of contributing to our understanding of the processes that might have led to language-like phenomena, it is essential to understand two kinds of question. The structure question concerns the kinds of structure the emergence of language might plausibly have been built upon. In our review of the models we ask, What has been given by fiat? What produces the initial patterns that are later amplified? The

process question asks, What sort of process implements the modulated positive feedback that leads to the emergence of new structure? Is this process a plausible candidate given our expectations about the conditions under which language-like behaviors arose?

In all of the models considered in this chapter, the initial condition of the model includes structure in the architecture of the agents, and structure in the interaction protocol. Many of the models begin with structure in a set of candidate meanings. Some of the models also begin with structure in a pre-determined set of possible forms. Other models begin without structure in the forms, but include structure in the world with which agents interact. Auto-organization is the transformation and propagation of these structures into new, sometimes surprising, structures inside or between the individual agents. The most compelling models are those in which the emergent structures cannot be anticipated from an inspection of the initial conditions.

In all models of the emergence of lexicon, each agent takes the behaviors of other agents as the target for its own behavior. The idea that humans are, and that their ancestors were, prodigious imitators is an important theme of contemporary studies of primate behavior (Tomasello, 1996). This mutual and reciprocal targeting of behavior creates a positive feedback loop. Once a behavior enters the repertoire of one agent, for whatever reason, it is likely to enter the repertoires of others, which makes it even more likely to enter the repertoires of others, and so on. In order to produce a shared lexicon, the positive feedback loop created by mutual and reciprocal behavioral targeting must be modulated or filtered in some way. The solutions to the modulation problem vary depending on the assumptions on which the model is built and on the choices made regarding the representation of the various elements of the model. We turn now to the details of the models.

Three Frameworks for modeling the emergence of shared lexicons

Three major frameworks have been employed to simulate the emergence of shared lexicons. Each framework makes a different set of explicit and implicit theoretical assumptions. The implicit assumptions are often revealed by choices concerning the representation of meanings, forms, and referents, the mappings among these, and choices of algorithms that constitute the interaction protocols.

The three frameworks are:

1. Expression/Induction (E/I)
2. Form-tuning
3. Embodied guessing game.

Expression/Induction (E/I) models (Hurford, 1999; Oliphant, 1997)

In these models, forms are provided in a large pre-defined closed set. The forms have no relevant internal structure and there are no relevant relations among the forms. Typically, the set of forms are constructed by the researcher through random selection of characters from the English alphabet.

There is also a relatively small closed, pre-defined set of structured meanings. The meanings are unrelated and atomic. Meanings are assumed to reside inside individual agents. The agents are completely disembodied, and the emergence of lexicon is taken to be a formal problem. Typically, meanings are represented in the model as a set of strings that invoke a simple "world" for the reader of the research. This world has no independent representation and plays no role in the simulation.

The constraints on forms and meanings require that the form-meaning mappings must be learned as an unstructured list. The agents cannot learn or exploit any higher-level regularity in the set of forms or in the set of meanings, or in the set of possible form-meaning mappings. This is done in part to ensure arbitrary relations between forms and meanings. Because the distinctiveness of forms is built into the architecture of the model, this approach cannot address the processes by which distinctiveness of forms might arise. Similarly, because meanings are pre-defined and arbitrary the model cannot address any role that relationships among meanings might have upon the emergence of a lexicon. . (See Kaplan 1999, for a model which explicitly addresses this question. See also the discussion below of Hutchins & Hazlehurst, 1995.)

A simulation begins with the creation of a community of agents. Agents then participate in pair-wise interactions that provide a mechanism for the transmission of form-meaning associations. A speaker, a listener, and a meaning are chosen for an interaction. From the outset, each agent is capable of representing all of the meanings internally. To express a meaning, the speaker chooses a form from the set of forms it has experienced other agents use for that meaning. If it has not yet experienced other agents expressing that meaning, it chooses a form at random (from a very large set). The listener knows which meaning the speaker is trying to express (the model forces this to happen) and simply adds the form-meaning pair expressed by the speaker to its list of observations. Note that since the meaning to be represented is given to both agents, the interaction is not a model of the communication of meanings via forms.

A system with only these parts will produce a lexicon in which each meaning is associated with many forms (as many forms as there were interaction events in which the speaker had not previously experienced a form for that meaning). To produce a lexicon in which each meaning has a single form, a production bottleneck is introduced. A *production bottleneck* exists if the method the speaker uses to choose a form to express a meaning results in some forms never being used by that speaker for that meaning, even though these form-meaning pairs may have been acquired by the agent. For example, a production bottleneck can be implemented by biasing agents to utilize those form-meaning mappings that they have experienced most frequently in the past. When a production bottleneck is in

operation, form-meaning pairs will gradually drop out of the system, and the community will eventually converge on a single form for each meaning.

There is a positive feedback loop here because the contracting set of form-meaning pairs in use by speakers limits the range of experience of listeners, which further limits the range of production when those listeners become speakers, and so on. The positive feedback amplifies small differences in the frequencies of form-meaning associations in the "aboriginal" lexicon. The positive feedback is modulated or filtered by the rule that implements the production bottleneck.

As long as the production selection rules lead agents to fail to produce some observed form-meaning pairs, the production bottleneck will act as a weak filter, favoring some forms over others. Sharing among agents is produced by the elimination of some forms and the inability of an individual to eliminate a form that is frequently used by others (even if dropped by an agent, it will force itself back into the agent's repertoire). E/I models assume the distinctiveness of forms and the a priori sharing of distinct meanings. They use modulated positive feedback to produce the emergence of shared mappings by amplifying differences in the frequencies of form-meaning mappings. They adjust the distribution of intact form-meaning mappings in a population of agents. As Oliphant points out, "Innately specifying the set of signals and meanings, however, negates what is perhaps the primary benefit of a learned system – extendibility." (Oliphant, 1997:117)

Form-Tuning model (Hutchins and Hazlehurst, 1995)

In this model, there is no pre-determined set of forms. In some sense there are no internal meanings either. Instead, there is a set of structurally related "visual experiences" imagined to result from stimulation of a visual sensory surface caused by objects within the simulation world. Visual experiences can stimulate production of agent behaviors without any explicitly represented meanings. Every agent produces a behavior (taken to be a verbal production) in response to each visual experience. Early in the simulation, the behaviors are undifferentiated. On each cycle of the model, a pair of individuals encounters a particular visual experience and each agent produces a behavior in response to that experience. Each agent then tries to shape its own behavior in two ways. First, it tries to shape its behavior so that it matches the behavior produced by the other agent. Second, its internal organization (implemented as a connectionist autoassociator network) leads it to change the structure of the behavior evoked by an experience so that it is different from the behaviors evoked by the other experiences. In this way, the behaviors are gradually tuned so that they (a) match the behaviors produced by other agents in the presence of the various visual experiences, and (b) discriminate among the visual experiences. The model will produce as many distinct behaviors as there are distinguishable visual experiences.

Each individual is an autoassociator network consisting of 36 visual input units, 4 hidden units, 4 verbal input-output units and 36 visual output units, as shown in figure 2. The simulation proceeds via interactions – one interaction is one time step in the simulation. An interaction consists of the presentation of a scene (drawn

from the set of 12 scenes, see figure 3) to two chosen individuals, a *speaker* and a *listener* (drawn from the set of 4 individuals). One of the individuals chosen, (say A) responds to the scene by producing a pattern of activation on its verbal output layer (A speaks). The other individual, (say B) also generates a representation of what it would say in this context. As listener, B uses what A said as a target to correct its own verbal representation. This moves B's behavior toward a match with the behavior produced by A. The listener, B, is also engaged in a standard autoassociator learning trial on the current scene, which means its own verbal representation – in addition to being a token for comparison with A's verbal representation – is *also* being used to produce a visual output by feeding activation forward to the visual output layer. The pattern of activation on the visual output is compared to the input signal to establish an error pattern, which is back-prapagated through the network. This changes the structure of the behavior evoked by an experience so that it is different from the behaviors evoked by the other experiences. These two sorts of learning together produce the matching of discriminable verbal behaviors. Over time, by choosing interactants and scenes randomly, every individual has the opportunity to interact with all the others in both speaking and listening roles in all visual contexts.

At the outset, the behaviors that are evoked by the visual experiences are not really forms that can play a role in a form-meaning mapping. All visual experiences give rise to approximately the same verbal behavior in all agents as shown in figure 4. This means that responses to visual stimuli initially carry no information. As the simulation progresses, however, the behaviors come to have the properties of forms and play the role of forms in form-referent mappings. These properties emerge because the networks develop internal structure that solves the dual problem of producing forms that distinguish visual experiences and producing shared form-referent maps among members of the community.

Contrary to what is generally expected of the relationships between forms and meanings, the mappings produced by this model are not completely arbitrary. Because the architecture of the agent requires the forms to produce an efficient or condensed encoding of the structure of the set of visual experiences, parts of the form (individual unit activations) encode features of the visual scene.

The patterns that arise on each agent's verbal medium come to discriminate among the patterns on that agent's visual medium and come to agree with the patterns that arise on the verbal medium of the other agents in the presence of each visual pattern as shown in figure 5. It is important to note that the individual agents become functionally equivalent, but not structurally identical. Agent internal organization provides functional equivalence yet is variable among members of the community.

In this model mutual reciprocal targeting of behavior supports a feedback loop as described for the E/I models. Instead of simply choosing a discrete form that matches the discrete form chosen by the other agent (as was done in the E/I models), agents in this model tune the structure of the forms they produce. What is shared in the "tuning" model is a skill to produce particular behaviors, rather than choices for pre-formed tokens. As a shared solution begins to emerge among a few agents, the experience of other agents is changed so that they are drawn toward that solution, which then affects more agents, and so on.

In fact, this model involves positive feedback organized by two kinds of tuning simultaneously. Consider the constraints on forms as forces applied to the weight matrices of each agent in response to each interaction. The constraint that *forms be shared* applies a force to shift the weights of the agent. This shifting has the effect that the next time the same experience is encountered the weights will produce a form more like the form that was produced by the other agent. The constraint that *forms be discriminable* one from the other applies a second force. The effect of this force is to shift the weights so that the form produced for each experience is maximally different from the forms produced for other experiences. Imagine both forces pushing the value of each weight. At the outset, there is no reason for these forces to be aligned. In fact, they often act in opposite directions. If two agents do not agree on how to represent an experience, the sharing force may act opposite the discrimination force on every weight. However, if two agents agree, even partially, then the sharing force may work in concert with and add to the discrimination force.

The ease of finding a solution that satisfies both constraints depends on the initial conditions of the model, especially the initial distribution of weight values. One positive feedback loop, call it the distinctiveness loop, amplifies initial random differences in the forms produced within agents across visual experiences. Another positive feedback loop, the sharing loop, amplifies initial random similarities in the forms produced among agents for each particular visual experience. As learning proceeds, the two loops may become mutually reinforcing. It is crucial to note that the structure of the forms is a product of the interaction of structure that is present in the visual experiences and emergent structure inside the agents. The two positive feedback loops amplify differences and similarities in that interaction between external structure and developing internal structure. Form-tuning models assume a structure of visual experiences and small random differences among the internal structures of individual agents. The distinctiveness of forms and the sharing of form-meaning mappings emerges from the operation of two, interacting, positive feedback loops.

It is easy to see that E/I models have given forms and given meanings and no referents at all. It is more difficult to say what the elements of the form-tuning model represent. One might take the visual experiences to be meanings analogous to the meanings of the E/I models. After all, it is only the modeler's assertion that these things are external to the agents that makes them so. In that case, meanings are given a priori. The process could then be said to produce distinctive forms and shared form-meaning mappings. Alternatively, the visual experiences could be taken to be referents outside the agents. In that case, the behaviors that are produced in response to the referents might be either meanings or forms. The interpretation of the behaviors as meanings is supported by the fact that they entail feature decompositions of the visual experiences. Feature theories of meaning have a long and venerable history. However, if the behaviors are meanings, then where are the forms and what is meant by the direct sharing of meanings? The interpretation under which the visual experiences are referents and the behaviors are forms seems at first glance somewhat anomalous because then the relation of form to referent appears not to be mediated by meaning. A third alternative is that the behaviors are forms and the meanings are represented implicitly in the structure

of the weight vectors of the agent, rather than explicitly as an autonomous interpretable representation. Under this interpretation, the meaning of a visual experience is the process that is required to produce its representation as a form. This last interpretation is important because under it, the models can be said to model the emergence of referent-meaning-form mappings with emergent forms and meanings, but without making a commitment to symbolic representations of meanings.

Embodied guessing game (Steels, 1996; Steels & Kaplan, 1999)

As was the case in the E/I models, forms in these models are arbitrary strings of syllables drawn from a fixed and given alphabet.

These models have the most complex representation of meanings among the three frameworks considered in this section. Meanings are symbolic representations of context-sensitive distinctions over a set of complex reference objects. As was the case in the form-tuning model, objects have properties that impinge upon agents' visual sensory surfaces. This process produces agent perceptions of features that are represented by continuous ranges along the properties sensed by the sensory surface. The task facing an agent is to discriminate topic (foreground) objects in the context of ground (background) objects within a scene using sets of features. In order to solve this problem agents employ an error signal generated by a built-in need for producing unique feature sets among all topic/ground possibilities.

In these models, there is an explicit distinction between objects in the world (referents) and symbolic representations of properties of those objects (meanings). This is the only framework of the three that includes an explicit representation of form, meaning, referent, and the relations among the three terms. In interaction, speaker and listener share awareness of the reference object. The speaker creates an utterance to encode a feature of the object. The listener tries to anticipate which feature the speaker will encode. When the listener guesses correctly, the association of the form (utterance) to the meaning (feature description) is strengthened for both speaker and listener. When the listener guesses incorrectly, the form meaning association is weakened. Importantly, the feature sets held by different agents may not be identical. However, as the universe of objects about which agreement is required becomes large, convergence toward shared perceptual distinctions develop, and shared form-meaning-referent mappings emerge.

The entire set of all possible form-meaning pairs is implicitly present in the model. Interactions strengthen some pairs and weaken others. A modulated positive feedback loop drives the process because stronger associations are more likely to lead to successful communication, which will make them even stronger. Meanings are grounded in the sensed properties of referents. The structure that is amplified by the positive feedback is produced by fortuitous agreements in the application of the meaning-making process. Forms have arbitrary relations to meanings and to referents.

Modeling the emergence of grammar

One goal of models that simulate the emergence of lexicon is to demonstrate possible origins for the "denotation" function of language. Denotation is made possible through development of shared and coherent relationships among forms, meanings, and referents. In general, forms within these models are atomic units. Models of simple lexical denotation make no effort to account for combinations of forms. Instead, the single relation of "distinctiveness" holds among the forms of the lexicon.

Simulations of the emergence of grammar attempt to explain the origins of a more complex but related phenomenon, namely the systematic nature of language. Every language organizes words of the lexicon into sequences that express complex meanings. The grammar of a language describes the internal structure of these sequences and accounts for the mapping from this structure to the complex meanings that are expressed. The fact that sequences of forms have internal structure implies new classes of relations among forms. The goal of models simulating the emergence of grammar is to describe the origins of complex form-form relations and the manner in which these constructs map onto complex meanings.

The computational requirements of a language with syntax entail:

1. **Compositionality**
   The capacity to construct composite forms (representing complex meanings) from simpler parts. Because human language involves a serial production device, complex forms (e.g., sentences) are composed through sequential concatenation of atomic forms (e.g., words).
2. **Systematicity**
   The capacity for complex constructs to entail "roles" for elements in the construct (e.g., noun) such that these elements retain their individual meanings (e.g., as a word) yet also serve the meaning defined by the roles they fill within the construct (e.g., as the subject of sentence).

Combining the properties of compositionality and systematicity accounts for the open-ended yet structured nature of language. The notion that language is inherently "generative" follows directly from these properties. With such a system, novel meanings can be expressed with novel forms and yet these sentences are easily understood by virtue of conforming to the grammar, the form-meaning mapping of the language.

A great challenge in the study of language is capturing the structural properties of such an arrangement, modeled as an abstract formal system, while addressing what we know about human evolution and history as well as what we know that language accomplishes in the world as a vehicle for situated communication. Attempts to model the emergence of grammar highlight certain relations among the elements of language while disregarding others. We now turn to an examination of a set of models addressing the emergence of grammar. For each model we try to illuminate the consequences of choices made in the representation of language elements and relations among the elements, as well as the processes which employ these elements and relations.

The capacity to learn systematic form-meaning mappings (Batali 1998, Kirby 1999)

In this group of models, which all employ the E/I framework (Hurford, 1999), each simulation includes a finite set of discrete tokens and a process that can assemble the tokens into complex forms or sequences. The entire set of such sequences is open-ended or potentially infinite. The simulation world also includes a set of structured symbolic meanings which agents share. The structure of meanings is given in some type of propositional language (e.g., a simplified predicate logic). There are no referents included in the simulation world.

These simulations demonstrate the development, through learning, of shared mappings between the structure of meanings and sequential patterns of forms. At the start of the simulation, there is no patterning within or among the forms (i.e., the relationships among forms is unspecified). A primary goal of these simulations is to demonstrate that learning devices employed in the service of inter-agent communication can produce the specific type of mappings characteristic of language. If such mappings can emerge from sharing propositional meanings through verbal encounters, then the nature of these encounters may provide a non-genetic generator of language structure.

As with models simulating the emergence of lexicon, agents in these models come together in interaction in the roles of speaker and listener. The speaker produces a sequence from a meaning, and the listener is challenged to produce the inverse mapping from form back to meaning. Importantly, the meanings in the encounter are unproblematically shared by the interactants. In all models of this group, the error signal that promotes consensus is made available from the a priori sharing of structured meanings together with the experience of specific forms produced by the speaker. These models assume that the capacity to construct complex propositional meanings arose prior to and independent of the capacity to express meaning, and that the sharing of meaning is accomplished through some non-linguistic means.

In Kirby'99, the agents' abilities to map between forms and meanings are provided by symbolic rules. The set of such rules used by an agent constitutes that agent's grammar. Rules are induced and maintained through an algorithm that strives to accommodate all experienced sentence examples in the simplest form possible. Given a complex form and the meaning that this form expresses, the induction algorithm of the learning agent first checks to see if the form conforms to the agent's grammar. If that check succeeds, then nothing else is required. Otherwise, a new rule, which produces this sentence, is entered into the agent's grammar. The new rule uses the given meaning to create the given form. This new rule produces the sentence as a singleton. At this point a generalization algorithm attempts to decompose the new rule in a fashion which exploits parts that are already available within other rules of the grammar. This is possible because of the well-formed symbol structures representing meanings. Over time, the set of rules becomes compact and coherent, making the grammar efficient while accommodating all experienced sequences.

At first, agent productions are simple one-to-one mappings from complex meanings to random sequences of tokens. This state of affairs results from the fact

that agents have minimal linguistic knowledge and thus speakers engage in many instances of "invention", modeled as the random selection of forms for gaps in production knowledge. Over time, after many instances of language production, rule creation, and generalization, agents converge upon systematic and shared sets of rules constituting a single grammar with properties of compositionality and recursion. Convergence in this process results from the properties of learning, which strives to generalize linguistic knowledge. Each new form-meaning mapping acquired by a learner is "chunked" so as to fit within the existing rule structure of the grammar, whenever possible. As a consequence, early in the simulation experienced forms are one-to-one with rules while later on the number of form instances dominates the number of rules. The more comprehensive the rule set becomes, the more likely it is that the rule set provides the mechanism for production (thus, "invention" is not required). This process creates positive feedback, which builds the structure of the rule set. The rule set, in turn, acts as a filter on the kinds of forms that can be produced.

In the general multi-agent case, the distribution of rule sets among agents produces the observed distribution of forms[1]. At the same time, the observed distribution of forms sets the targets for the agents' rule sets. If fortuitous independent invention increases the representation of a particular form for a particular meaning, that form-meaning pair will be a more frequent target for the rule sets of other agents. If it is the sort of mapping that the induction algorithm can learn, it will become more likely that the rule sets of other agents will come to produce that form for that meaning, which will further increase the representation of that form-meaning pair. The positive feedback loop amplifies the effects of fortuitous coincidental form-meaning pairs, and is modulated by the learning processes that govern the modifications of rule sets. The induction algorithm is a sort of resonant filter on the positive feedback loop, reinforcing some signals and causing others to dissipate.

In Batali'98, a very different representational mechanism is employed. However, the methodology is similar. The objective of the model is a shared mapping between structured meanings and patterned sequences. In this model, an agent's ability to map between forms and meanings is implemented by a recurrent neural network. The weights of the network propagate structure from input units to output units. Network weights develop in such a way as to support a systematic mapping from sequences of basic tokens to complex meaning structures. This mapping is taken to be a grammar, and it emerges from the interactions among agents.

As Elman (1991) showed, sentence processing by an agent with this type of architecture can be understood in terms of trajectories through the network's internal state space. The representational space of the network can simultaneously encode information about individual tokens (inputs to the network) and the positional or context-sensitive information given by the sequence of tokens constituting a sentence. The learning algorithm applied to a recurrent neural network partitions this space so as to accommodate all of the examples

[1] In Kirby (1999) the population is composed of a single speaker/listener pair.

experienced, while simultaneously enabling systematic generalization, capturing information about the structure inherent in the set of examples.

In Batali's model, as with all of the others examined in this paper, agents interact via the roles of speaker and listener. More accurately, the model explicitly treats the problem as that of a "learner" sampling the productions of "teachers" in each iteration of the simulation. Each teacher produces sequences through a protocol that selects the sequence of tokens which "best" invokes the given meaning. This meaning and the produced sequence then provide the learning instance for the learner. Over time, preferred mappings emerge because learners are also teachers. These preferences result from (1) the distribution of initial conditions inherent in the starting weights of the networks, (2) biases which may be associated with the ordering of interactions, and (3) the constraints of emerging agent representational spaces which may promote some mappings and inhibit others.

Despite very different choices of representational frameworks, the Kirby model and the Batali model share high-level assumptions. Objects in the world are assumed to be irrelevant to the problem of language systematicity. Abstract structures are declared to be meanings by stipulation alone.

At a finer level of description, the shared assumptions made by the two models entail four important features: (1) the structure of meanings (i.e., the meaning-meaning relation) is given rather than emergent, (2) communication has no role because meanings are always shared perfectly before interactions take place. Interaction is present only as a context for the production and comparison of form-meaning mappings, (3) the distribution of understandings held by agents plays no role in the models because all agents understand the world in exactly the same way and, (4) the framework assumes that the origin of grammar is explained by the propagation of a pre-existing internal language of thought to an external language in the world.

We will return to this set of shared features in the discussion section. First, we review a very different simulation that nonetheless has the shared goal of elucidating the nature and possible origins of language systematicity.

The emergence of propositional descriptions about the world (Hazlehurst & Hutchins, 1998)

In H&H'95 and Steels'96, the emergence of "lexicon" is modeled as the development of *shared descriptions of perceptual distinctions*. Agents in these models inhabit shared worlds containing a variety of objects. Meanings are emergent perceptual structures that mediate relations between forms and aspects of objects in the simulated world. Forms are descriptions of the perceptual structures. The simulations model the development of mappings between emergent forms, emergent meanings, and given environmental structure. These meanings are, thus, not structural entities created in advance by the researchers but rather are themselves developed in the course of the simulated interaction processes. The agents engage the objects as part of an explicit communication task.

In modeling the emergence of grammar, H&H'98 add the social and cognitive problems of *sharing attention* to their earlier model. Sharing of attention introduces a time-based sequential coordination constraint on interactions, providing a temporal scaffold for the simultaneous emergence of patterned forms and systematic form-meaning mappings.

H&H assume that language is a set of resources that have been shaped in response to the problem of coordinating action. Agents have some built-in cognitive properties. For example, agents have the ability to perceive the world via a visual modality. But even basic cognitive properties may be shaped by their use. For example, perceptual structures are tuned by constraints that are imposed by the need to represent that which is perceived. Acts of communication are both shaped by the organization of cognitive processes and put constraints on the organization of cognitive processes. Meanings and forms arise together. The requirement to produce forms that coordinate with meanings shapes the nature of meaning.

In H&H '95, agents that engage each other as discourse participants share the scene about which they are speaking. The sharing of a scene in interaction is given as a property of the simulation and both participants attend to the shared scene in its entirety. In H&H '98 discourse participants also share a scene, but rather than a scene being a single object, a scene is composed of multiple objects located on a spatial grid which is mapped onto the visual sensory surfaces of each agent in interaction. Now, the agents must negotiate a shared focus of attention in order to communicate successfully. The problem of communication now includes the creation of shared understanding about which referent is being discussed.

The speaker in each interaction has some specific object in mind that is present within the current scene. We refer to this as the "intentional object" of an interaction. For the purposes of the interaction, this chosen (but privately held) object can be thought of as foreground against a background comprised of a scene of spatially arranged objects. The listener sees the entire scene but at the start of the interaction only attends to the foregrounded object by chance. When taking the role of speaker, agents produce non-linguistic structure that may lead the listener (in concert with produced linguistic structure) to attend to the referent held in the speaker's mind. When taking the role of listener, agents may make use of this structure to determine what the speaker has in mind. Communication is successful when the listener identifies what is on the speaker's mind by achieving shared focus of attention upon the ~~foregrounded~~ intentional object (see Figure 6~~2~~).

The agents' abilities to accomplish this communication task depend on the speaker's ability to guide the listener's attention to the object held in the speaker's private mind. It is also dependent on the listener's ability to follow the speaker's lead. Agent focus of attention is materially available to interactants as direction of gaze, which identifies a location within the grid of objects that make up the visual scene. In other words, coordination is made possible through (non-identical) visual and verbal inputs to each agent. Coordination is only accomplished through success at the respective (and asymmetric) tasks of the two participants, and failure to coordinate within an interaction terminates the interaction.

Agents are composed of complex, modular, connectionist networks which map from aural and visual sensory surfaces (inputs) to motor controls effecting gaze and verbal actions (outputs). In interaction, motor productions by the speaker create

sounds (forms) that impinge upon the aural sensory surfaces (the "ears") of both speaker and listener. In addition, each agent perceives the location of gaze (the "focus of attention") of both agents on each time step of the interaction. Actions are produced on each time step of the interaction, first by the speaker and then by the listener. Actions are realized in the world as shifts in the location of gaze of each agent and a public sound (form) produced by the speaker.

As with H&H'95, ~~early in the simulation,~~ the forms that might do communicative work in this process are of no use early in a simulation run. Tokens are represented by continuous values along dimensions in agent verbal "articulatory space," so the set of possible tokens is infinite. At the beginning of the simulation, there is no useful structure in agent verbal productions (see Figure 7a~~3~~). Stringing such tokens together in sequential constructs also carries no information at this point.

Early in a simulation run the agents often terminate interactions quickly, because they are unable to sustain the coordination required for the "follow the leader" communication task, and therefore the verbal sequences they produce are short. In fact, early in a simulation run, the communication task is accomplished by chance alone and verbal forms are meaningless entities – just noise. At this point, there is no artifactual structure (forms or behaviors) capable of mediating agent accomplishment of the task. Later in the simulation run, structure in support of the task (both internal to and among agents) does develop and success at the communication task rises.

One way to measure this success is by accounting for how interactions terminate. There are five possible conditions for interaction termination in a simulation run:

1. *Invalid Shift.* Speaker attempts to shift gaze outside of the visual field.
2. *Disagree.* Speaker and listener gaze become uncoordinated.
3. *Halt.* Speaker and listener successfully conclude the interaction by halting at the intentional object held in the speakers mind and located within the shared visual field.
4. *Cycle.* Speaker revisits a location with gaze that was attended to on the previous time step.
5. *Max.* Some maximum number of time steps have occurred within the interaction.

Figure 8~~5~~ shows the evolution of a simulation run in terms of the frequency of each type of termination condition as the run proceeds. As shown, the simulation evolves to a point where the communication task is reliably accomplished. This is seen by the rise in occurrence of the *Halt* condition and the extinction of all other conditions for terminating interactions. ~~Early in a simulation run the agents often terminate interactions quickly, and so the verbal sequences they produce are short. Typically, they fail immediately at the "follow the leader" communication task.~~

~~However, over~~Over time within a single simulation run, several kinds of structure begin to emerge that mediate the organization of behavior which accomplishes the communication task and which provide the foundations for a simple language. First, agents produce a class of forms and meanings, and the mapping between them, which denote objects in the world of the simulation (see Figure 7~~3~~a). This occurs because despite the problem of coordinating attention,

agents nonetheless attend (by chance) to the same object on the first turn of some interactions. When this happens, the agents can tune their verbal productions for the object to match that produced by the other. This emergence of lexicon is a replication of the result obtained in H&H'95. Structure which builds internally in service of perceptual distinctions among referents propagates to the verbal form-producing and form-interpreting mechanisms of agents. These properties enable simple denotation of the current focus of attention on the first turns of interactions, precisely as happened in H&H'95.

Second, agents develop internal structure that permits them to use the gaze of the other as a guide to attention. This produces the joint management of attention. This attention management function is structurally similar for speakers and listeners although it entails distinct tasks for each of them. As discussed above, the speaker learns to produce ~~cues~~ shifts in gaze that can direct the listener's gaze to the object held in mind, while the listener learns to ~~use the cues produced by the speaker to~~ follow the speaker's gaze. In the process of producing coordinated attention management, agents develop another (second) class of verbal forms that refer not to objects in the visual scene but rather to *the actions required to maintain coordination in visual attention*. In other words, a new class of internal structure, involving the motor control of attention, propagates to the verbal form-producing and form-interpreting mechanisms of agents (see Figure 7~~3~~b). These forms (when fed back to agents via an auditory sensory surface) produce internal meanings which, when allowed to activate motor control portions of agent architecture, produce actions such as "shift gaze leftward". Agents develop this second class of forms (and the sharing of these forms) through the process of agreement in motor control structure associated with controlling gaze.

This development can be seen as analogous to the externalization of perceptual structure that serves as the basis for development of the lexicon. In each case there is the requirement for coordination between something internal (motor or perceptual structure), something external (an object or an action), and something verbalized (a language form). As the attention management task is sequential, so is the production of language forms that come to reliably represent the coordination task. In fact, after a period of time agents can use the language constructs alone to simulate the appropriate motor events, without actually manifesting any of the entailed actions in the world (see Figure 9~~5~~).

This development provides the agents with a system for predicating and communicating the relative arrangement of objects in visual space. Clearly, this system has emerged from the interactions of the agents and has not been specified by any central designer. In this complex system, multiple interacting modulated positive feedback loops, operating simultaneously on several aspects of agent/environment coupling, amplify emergent patterns that link perceptual processes to words and features of objects, motor processes to words and shifts in attention, and attention management processes to gestures and direction of gaze.

At this point a system of language with some very interesting properties has been bootstrapped. H&H show that this language exhibits both systematicity and compositionality, and that agent facility in producing valid strings of the language (and rejecting all others) demonstrates agent grammatical competence in the language. In particular, there are an infinite number of valid strings that are

possible, yet they all conform to a compact structural description that demonstrates word classes (objects and actions), context-free composition from the lexicon (e.g., the words representing objects take the same form regardless of position in a complex construction) and recursion (i.e., complex constructions involve the systematic use of component sequences).

Discussion

There are some important similarities and differences between the assumptions made by the two classes of models reviewed in this section. Both classes of models deal with the emergence of grammar, taken to be a mapping between forms and meanings. Both classes of models address how this mapping takes on specific properties entailing the emergence of certain complex constructions or form-form relations. Of particular interest to all of the models are constructions that demonstrate compositionality and systematicity in language. Finally, in none of the models is the mapping built into the "genetic material" of the agents. Neither is there a centralized designer that defines that grammar. Rather, the mapping emerges in the course of exchanges that produce a feedback effect encouraging convergence upon a shared grammar.

The classes of model differ, however, in the processes and component structures of the simulations that put the mapping (the grammar) in place. Although both classes of models employ a framework of "communication" as the mechanism by which shared grammar emerges, the models take very different stances on what this actually means. In the E/I models of Kirby and Batali, agents come together in interaction *already* sharing the meaning of what is being talked about *without* requiring a language to accomplish this. The goal of these simulations is to demonstrate that the structure inherent in complex proposition-like meanings that are already available and shared can propagate into form-meaning mappings (the grammar) simply through the process of agents learning from each other's productions. Thus, interacting agents are imitating each other's speech behavior but not really communicating, per se. Hurford (1999) explains that the emergence of systematic form-meaning mappings in these models results from the imposition of "bottlenecks" which prune the choices (all implicitly present in the meaning and form repertoire of the agents) down to a compact, shared, and explicit set. He speculates that these models would work equally well by eliminating the multi-agent component and simply imposing constraints upon an agent's production or semantic repertoire.

By contrast, in the model of H&H the task presented to agents is built upon the premise that agents engage each other in order to resolve the communication problem that arises when one agent (the speaker) has something in mind, an attentional relationship to a perceived object, that is not directly accessible to the other agent (the listener). In this case, the structure of meanings and forms must co-emerge. There is no "language of thought" available to agents prior to engaging the world and negotiating its meaning. Rather, agents learn about the world through perceiving it, and coordinating action within it, and talking about it. Thus, the constraints of language production are a resource for developing cognitive

structure involved in perception and motor tasks. Simultaneously, as structure associated with solving these tasks begins to emerge, that structure can become a resource for language production. H&H claim that the origins of language systematicity and structured meanings reside in the processes that produce this coordination among structures.

In contrast to Hurford's (apparently correct) observations about the E/I framework employed by Batali and Kirby, isolated individual agents in H&H'98 *cannot* produce a systematic language. This is so, even though an individual agent is fully equipped in principle to produce a private language. In the case of an individual learner, the cognitive task of tracking a target location in visual space (without requiring coordination with a listener) does not produce a systematic language. For individual learners, this motor control task is solved by treating the visual field as a gradient encoding the distance to target. In this case, any path that minimizes distance to target is a good path, and there is no need to adopt a convention for parsing the visual field. Such a representation of the visual problem (unencumbered by the constraints of communication) does not yield the cognitive structure necessary to produce a systematic language. It is only in the course of coordinating action among agents in interaction that a systematic language emerges from this model. When agents must coordinate their actions with others, but not when they act alone, the preferred pathways through visual space build upon conventions that yield the compositional and systematic construction of sequences.

General discussion of the models

Framing the issue in terms of the modulation of positive feedback permits us to apply the process and structure questions to the modeling attempts. What is the process in which modulated positive feedback could operate to produce language-like structure? What is the relation of the process by which structure develops to the processes by which structure is thought to change in historical time? What is the substrate on which the growth of structure takes place? The answers to these questions clarify the contribution of the modeling efforts to the high-level goals of understanding the sorts of processes that may have led to the development of language.

We do not expect language to emerge from a process that does not include the challenge of communication. We find it implausible to look for the origins of language in interactions where fully composed meanings are injected into the mind of the listener before a public expression of that meaning is encountered. E/I models may be interesting engineering solutions, but this aspect of the process violates our understanding of the conditions under which language-like behavior might have arisen.

The structure question is more difficult to handle. The field of cognitive science is nowhere near a consensus on the nature of linguistic representations for contemporary humans. There seems little hope that we could resolve this question for proto-humans at some unknown point in the evolution of intelligence. If the argument cannot proceed from evidence, what else is there?

Expression/Induction models have nothing to say about the development of the language of thought or the representation of meanings because they simply assume that complexly structured meanings exist prior to the language phenomena that emerge. The problem they address is the development of shared mappings between pre-existing meanings and pre-existing forms (in the case of lexicon) or pre-existing complex meanings and emergent complex forms (in the case of grammar). These models fit a view of language in which the language of thought somehow arises prior to and independent of the development of public language. As the assumptions of the E/I models make clear, if confronted with the children's riddle, "Which came first, the chicken or the egg?", E/I modelers would blurt out "Egg!". The language chicken simply emerges from the well-formed meaning egg. This view is grounded in the Physical Symbol System Hypothesis (Newell & Simon, 1990, Newell, et al., 1989), which assumes that all intelligent processes are characterized by the manipulation of symbolic structures.

The positive feedback loop in E/I models of the emergence of lexicon amplifies small differences in the frequencies of form-meaning associations in the "aboriginal" lexicon and is modulated by the rule that implements the production bottleneck. E/I models have no effect on the structure of forms, the structure of meanings, or the structure of form-meaning mappings because these are all given. Rather, the models simply adjust the initially random distribution of intact abstract form-meaning mappings in a population of agents.

H&H, and we suspect Steels as well, have a different image of the historical processes that led to the development of language and the cognitive abilities that go with it. Their response to the chicken-and-egg riddle is "neither". Imagine a history in which structural regularities arise in the interactions among individuals. Imagine further that some of these regularities are true emergent properties in the sense that they cannot be fully constrained by the behaviors of either interactant alone. In such a situation, individual agents might begin to develop internal structure that permits them to maintain coordination with the emergent regularities. Eventually, they might even develop the ability to produce the formerly emergent structure alone. In such a world, symbols could arise between agents first, and become internalized later. Meaning structures could be embodied in forms that are assembled in interaction long before those meaning structures are internally represented by any agent. Language of thought need not precede public language. Structure and process always go hand in hand. If the process through which language emerged involved complex acts of embodied communication, then public propositional expressions could be the source rather than the result of the development of propositional mental representations. The news here is not just that chicken and egg develop together. The bigger point is that this viewpoint introduces another place to look for the origins of cognitive complexity. It has long been argued that the ontogeny of high-level cognition depends on social interaction. (Vygotsky, 1978, Wertsch, 1985) We propose that the same may be true of the phylogeny of high-level cognition.

The form tuning models of Hutchins and Hazlehurst (1991, 1995, 1998) attempt to demonstrate the dialectical development of internal and external structure. Form-tuning models assume a structure of visual experiences and small random differences among the internal structures of individual agents. The

distinctiveness of forms and the sharing of form-meaning mappings emerges from the operation of multiple, mutually reinforcing, positive feedback loops that amplify differences and similarities in the interaction among stable external structure (referents), emergent external structure (forms), and emergent internal structure (meanings). These models have weaknesses. For example, the lexicon model lacks arbitrariness in the form-meaning relation, and the grammar model derives its sequentiality from a course of shared action (which cannot account for much of the sequential structure of real languages). Nevertheless, the models do suggest, and are informed by, a wider view of the nature of the issues that should be addressed by models of the emergence of language. The coordination of emergence of internal and external structure suggests that we should look for the origins of cognitive complexity in embodied interactions among agents.

The structure of language must come from somewhere. All of the models we have looked at try to show how it emerges in a process of auto-organization. All of these models work by constructing positive feedback loops that amplify certain, nearly invisible, initial differences. But what are the differences and what are the processes that amplify them? We have tried to use the positive-feedback-loop framework to identify the signals and the processes that modulate the propagation of those signals in the emergence of novel structure. In E/I models of the emergence of lexicon, the original signal is simply irrelevant noise introduced by random correspondences of arbitrary patterns. It may indeed be possible to get a community of agents to discover regularities in those patterns and to conventionalize those regularities. One must ask, however, could real language come from a process like that? Could language be a structure that has been extracted entirely from random fluctuations in arbitrary patterns? We think not. We expect language to be a structure that highlights and focuses patterns of embodied experience. E/I models of the emergence of grammar assume a complex propositional representation of meaning. Is the problem of the emergence of the grammar of public language simply a matter of propagating fully formed, complex, internal representations into public representations? Inserting meaning structure by fiat in these models means it must be accounted for some other way. Surely some kinds of meaning precede the advent of language. The physical symbol system hypothesis assumes that propositional representations of meaning precede the advent of language. We offer an alternative; a cultural symbol systems hypothesis, according to which symbols arise in interactions among agents concurrently with, or even before, the internal structures with which they are coordinated. Under this hypothesis, the ability to give meanings a propositional representation is a consequence rather than a cause of the ability to create grammatical external forms.

References

Batali J (1998) Computational simulations of the emergence of grammar. In: Hurford J, Studdert-Kennedy M, Knight C. (eds) *Approaches to the evolution of language: Social and cognitive bases.* Cambridge University Press, Cambridge UK pp 405-426

Elman J (1991) Finding structure in time. *Cognitive Science, 14*, 179-211.

Hazlehurst B, Hutchins E (1998) The emergence of propositions from the co-ordination of talk and action in a shared world. *Language and Cognitive Processes*, 13(2/3): 373-424.

Hurford J (1999) Expression/Induction models of language evolution: Dimensions and issues. In: Briscoe T (ed) *Linguistic evolution through language acquisition*. Cambridge University Press, Cambridge UK

Hutchins E, Hazlehurst B (1991) Learning in the cultural process. In: Langton C, Taylor C, Farmer D, Rasmussen S (eds) *Artificial Life (Vol. II. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. X)*. Addison-Wesley, Redwood City CA, pp 689-706

Hutchins E, Hazlehurst B (1995) How to invent a lexicon: the development of shared symbols in interaction. In: Gilbert N, Conte R (eds) *Artificial Societies: the computer simulation of social life*. UCL Press, London, pp 157-189

Kaplan F (1999) A new approach to class formation in multi-agent simulations of language evolution. In: *Proceedings of the Third International Conference on Multi Agent Systems ICMAS'98*. IEEE Computer Society Press

Kirby S (1999) Syntax out of learning: The cultural evolution of structured communication in a population of induction algorithms. In: Floreano D, Nicoud J-D, Mondada F (eds), *Advances in Artificial Life, Proc. of the 5th European Conference, ECAL'99*. Springer, Berlin, pp 694-703

Newell A, Simon H (1990) Computer science as empirical enquiry: Symbols and search. In: Garfield JL (ed) *Foundations of cognitive science: the essential readings.* Paragon House

Newell A, Rosenbloom PS, Laird JE (1989) Symbolic architectures for cognition. In: Posner M (ed) *Foundations of cognitive science.* MIT Press.

Oliphant M (1997) *Formal approaches to innate and learned communication: Laying the foundation of language.* Unpublished doctoral dissertation, University of California, San Diego

Steels L (1996) Self-organizing vocabularies. In Langton C, Shimohara K (eds), *Proceedings of Alife V*. MIT Press, Cambridge MA, pp. 177-184

Steels L, Kaplan F. (1999) Collective learning and semiotic dynamics. In: Floreano D, Nicoud J-D, Mondada F (eds), *Advances in Artificial Life, Proc. of the 5th European Conference, ECAL'99*. Springer, Berlin, pp 679-688)

Tomasello M (1996) Do apes ape? In: Heyes C, Galef B (eds), *Social learning in animals: the roots of culture.* Academic Press, San Diego, pp. 319-436

Turner JS (2000) *The extended organism: the physiology of animal-built structures*. Harvard University Press

Vygotsky LS (1978) *Mind in society: The development of higher psychological processes.* Harvard University Press

Wertsch J (1985) *Vygotsky and the social formation of mind*. Harvard University Press