ELSEVIER

Cognitive Science 28 (2004) 937–962

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Establishing conventional communication systems: Is common knowledge necessary?

Dale J. Barr

Department of Psychology – 075, University of California, Riverside, CA 92521, USA

Received 31 July 2003; received in revised form 14 July 2004; accepted 21 July 2004
Available online 25 September 2004

Abstract

How do communities establish shared communication systems? The Common Knowledge view assumes that symbolic conventions develop through the accumulation of common knowledge regarding communication practices among the members of a community. In contrast with this view, it is proposed that coordinated communication emerges a by-product of local interactions among dyads. A set of multi-agent computer simulations show that a population of "egocentric" agents can establish and maintain symbolic conventions without common knowledge. In the simulations, convergence to a single conventional system was most likely and most efficient when agents updated their behavior on the basis of local rather than global, system-level information. The massive feedback and parallelism present in the simulations gave rise to phenomena that are often assumed to result from complex strategic processing on the part of individual agents. The implications of these findings for the development of theories of language use are discussed.

Keywords: Conventions; Common knowledge; Pragmatics; Communication; Multi-agent simulation

1. Introduction

A critical prerequisite for successful language use is shared knowledge of conventions. Symbolic conventions are arbitrary by their very nature, and human languages are vast collections of such conventions. Even the simplest utterance takes for granted a large amount of knowledge about the phonological, syntactic, semantic, and discourse conventions of a language. It is an impressive feat that despite differences in input, members of a language community acquire semantic representations that are similar enough so that any two arbitrarily chosen individuals can successfully communicate with one another. Where do these shared representations come

E-mail address: dale.barr@ucr.edu.

0364-0213/$ – see front matter © 2004 Cognitive Science Society, Inc. All rights reserved.
doi:10.1016/j.cogsci.2004.07.002

from? What guarantees that the members of a language community will acquire the same conventions and use them in similar ways? Given the absence of some benevolent designer, what is needed is a theory that relates macro-level effects—the establishment and maintenance of symbolic conventions in the language community—to micro-level causes—language users and their behavior in conversational exchanges.

One such theory is the theory of signaling that David Lewis (1969) developed as part of a more general theory of social conventions. In Lewis' view, conventions evolve out of the recurring coordination problems communities face. For example, the lexicon of a language grows out of the recurring problem of communicating about certain entities or states of affairs. By characterizing communication as a coordination problem, Lewis situated his analysis within the framework of classical game theory (Schelling, 1960; von Neumann & Morgenstern, 1944). Coordination problems are problems of interdependent decision making in which individual agents select an action from a set of possible actions, seeking to align their selections with those of other agents. In itself, the selection of any particular action is arbitrary, and no agent has any prior preference for choosing any one action over another. Participants in a coordination game reason strategically about the behavior of other agents in order to reach equilibrium, a state in which no single agent would prefer to change its decision given the decisions of its co-participants. Because of the interdependent nature of decision making in coordination problems, Lewis assumed that agents should base their decisions upon mutual expectations: that is, they should consider what other agents will do, what other agents will expect them to do, what other agents will expect them to expect others to do, and so on, ad infinitum.

In Lewis' proposal, when coordination problems recur among the members of a community, the community develops regular patterns of solving these problems. These regular patterns become conventional through the accumulation of common knowledge. Common knowledge is important because it provides agents with a basis for generating the mutual expectations that are necessary for solving coordination problems. Assume that individuals from some community C believe that a coordination problem, P, and its solution, S, are common knowledge in the community. When two individuals A and B who are members of C meet and face a specific instance of P, they can mutually expect adherence to S rather than to some other solution, S'. This is by virtue of their mutual belief in their joint membership in C, where the practice is common knowledge.

How do individuals develop common knowledge of conventions? Although there can be various bases for common knowledge of social conventions, in the case of language, the critical basis is provided by communicative interaction. Each member's interactions with a sample of individuals from the community enable them to generalize to the entire community, even if the sample is relatively small. "If one has often encountered cases in which coordination was achieved in a certain problem by conforming to a certain regularity and rarely or never encountered cases in which it was not, he is entitled to expect his neighbors to have had the same experience" (Lewis, 1969, p. 40).

In sum, Lewis' theory, which I refer to as the Common Knowledge theory of conventions, assumes that the use of conventions relies upon strategic reasoning and global representations; that is, representations about the community. Lewis' theory and its attendant assumptions have wielded an important influence on theories of language use, exemplified in the work of Clark (1992, 1996). This influence can seen in the idea that language use can be characterized

as a special kind of coordination problem, in which interlocutors strive to achieve shared understanding by developing mutual beliefs. Thus, Clark and Marshall (1981) argued that determining the referent of a definite expression such as *the movie playing at the Roxy* requires interlocutors to coordinate their beliefs. Specifically, they suggested that definite references are only guaranteed to succeed if speakers and listeners produce and interpret them against the background of a special kind of shared knowledge, known as "common ground," which consists of the set of mutually held beliefs, assumptions, and suppositions. A critical part of common ground is common knowledge of conventions.

Clark and Carlson (1981) proposed that the context against which people process language can be reduced to common ground. However, recent experimental studies of referential communication indicate that people violate common ground in their processing of referential expressions (Horton & Keysar, 1996; Keysar, 1994; Keysar, Barr, Balin, & Brauner, 2000; Keysar, Barr, Balin, & Paek, 1998; Keysar, Lin, & Barr, 2003), including in their processing of conventions (Barr, 1999; Barr & Keysar, 2002). Keysar et al. (2003) found that when listeners interpreted a speaker's referring expressions, they considered objects not known to the speaker in spite of clear evidence of the speaker's ignorance of these objects. Surprisingly, they did so even when they thought that the speaker had a false belief about the identities of these hidden objects. Although other studies find that the task-relevant common ground may at times partially constrain production and comprehension (Arnold, Trueswell, & Lawentmann, 1999; Hanna, Trueswell, Tanenhaus, & Novick, 1997; Nadig & Sedivy, 2002), it is clear that even when a language user has the proper beliefs about common ground there is no guarantee that they will use it at all during processing (Keysar et al., 2000, 2003).

Against the background of the common knowledge theory, the non-normative, egocentric behavior of language users uncovered in the laboratory presents something of a paradox. Common ground and common knowledge are both forms of mutual knowledge: knowledge that is shared and known to be shared. If successful use of conventions relies upon representations of mutual knowledge, and language users do not reliably consult such representations, then how is possible for communication to be routinely successful? In other words, if we remove mutual knowledge's guarantee of successful communication, why don't people constantly misunderstand one another?

Clearly, the egocentric use of language will only succeed to the degree to which speakers and listeners can typically count on overlapping perspectives. To the extent to which they typically diverge, interlocutors will have to rely on their common ground. However, currently very little is known about what causes the semantic representations of language users to converge or diverge. This leads us back to the original question of how communities of language users establish and maintain symbolic conventions. The critical question is, can such conventions emerge without the benefit of common knowledge?

In what follows, I argue that common knowledge is not necessary for the emergence of symbolic conventions, proposing instead that semantic representations are coordinated through use; that is, as a by-product of individual attempts at coordination among speakers and listeners which are distributed over time and across the language community. I begin by reviewing previous work on conventions which suggests that high-level common knowledge and strategic reasoning may not be necessary for community-level convergence. In particular, past research on language evolution provides some compelling arguments against the common knowledge

theory. However, because these studies specifically target questions regarding language evolution rather than common knowledge, they do not test many important assumptions about the extent of global knowledge necessary for convergence. The current work seeks to advance these models by systematically exploring the minimal agent microtheory necessary for the emergence of conventions. Using multi-agent simulation, I show that convergence is a possible by-product of interaction even when agents have extremely limited abilities and sample narrowly from the community. The finding that coordinated semantic representations can occur without global representations of common knowledge casts doubt upon the claim that language users must routinely call upon such representations when producing and interpreting utterances.

1.1. Prior work

Previous research has challenged the view that the emergence of social conventions requires rational agents who strategically reason about common knowledge. Young (1998) argued that "social feedback mechanisms can substitute for high levels of knowledge and deductive powers on the part of individuals" (p. 662). Using analytical techniques, Young (1993) showed that social conventions can emerge even in populations in which agents lack complete knowledge or sometimes make irrational choices. In a similar vein, Shoham and Tennenholtz (1997) used computer simulations to show that a community of agents can adopt social conventions without common knowledge. However, neither of these studies addressed the problem of the emergence of symbolic conventions. Symbolic conventions might present a different case, since as some have forcefully argued, the inherent ambiguity of language would seem to make routine assessments of mutual knowledge necessary (Clark & Carlson, 1981; Clark & Marshall, 1981; Gerrig, 1986).

Studies of human discourse by Garrod and colleagues have shown that people can adapt existing symbolic conventions to suit their needs (Garrod & Anderson, 1987; Garrod & Doherty, 1994). Garrod and Anderson (1987) found that dyads who worked together to move a marker around in a maze converged upon a set of temporary conventions for producing and interpreting utterances pertaining to locations in the maze. Garrod and Doherty (1994) formed virtual communities in the laboratory wherein each member interacted dyadically with other members of the community. As a by-product of these individual dyadic episodes, certain referring conventions appeared and eventually generalized to nearly the entire community, even though each individual participant was uninformed that the series of partners he or she experienced were drawn from a larger community of players. Thus, it seems possible that the conventions in these studies emerged without common knowledge, although it is possible that participants inferred their membership in a community based on the commonality of behavior they observed across partners.

Few other laboratory studies of the emergence of symbolic conventions exist, in part because of the clear logistical challenges posed by observing the behavior of a large population of language users over a relatively long period of time. Furthermore, attempting to study populations of individual agents poses analytical problems, because the complexity of large scale social aggregates often makes it extremely difficult, and sometimes impossible, to predict the behavior of the system over time (Holland, 1998). For these reasons, a technique commonly

used to study social dynamics is multi-agent computer simulation. In multi-agent computer simulation, the researcher attempts to simulate community-level patterns of behavior as emergent products of the actions of individual community members, or agents. The researcher can independently manipulate various parameters, including population size, community structure, and agent implementation, and then observe how the system behaves.

A common use of multi-agent simulation is in the study of the evolution of language. Much of this research is directed primarily at questions that are specifically linguistic in nature (e.g., explaining universals of language or the emergence of structural regularities) rather than examining the necessity of common knowledge for conventions. Nonetheless, this work can offer some insight into the conditions under which conventions emerge more generally.

1.1.1. Multi-generational theories of conventions: linguistic conventions are invented and re-invented by each generation of language learners

The multigenerational approach to language evolution suggests that certain non-genetic selectional pressures cause languages to evolve to fit human minds. These pressures can arise from non-linguistic information processing limitations in the cognitive system (Christiansen & Ellefson, 2002). More commonly, however, work in this tradition takes up the thesis put forth by Lightfoot and others (Lightfoot, 1991, cited in Niyogi & Berwick, 1997), that language acquisition is a motor for language change. A few models combine cultural with genetic evolution (e.g., Cangelosi, 2001; Cangelosi & Parisi, 1998), although other work has found that cultural learning mechanisms alone are sufficient to produce community-level coordination. A common assumption in these models is that the forces that shape language have their impact over multiple generations of language users.

In these multi-generational models, the language of a population of agents evolves over the course of many generations through observational learning. In a typical model, at each generation new agents enter the population and are trained on a sample of form-meaning pairs produced by adults. After this learning phase, the now-adult agents remain in the population to transmit their knowledge to the next generation. Agents are often implemented in the form of neural networks that operate under error-driven learning. Oliphant (1999) has shown that given an appropriate learning algorithm, the iterated process of cultural transmission can yield a stable set of form-meaning mappings in the population. Using similar techniques, Livingstone and Fyfe (1999; see also Livingstone, 2002) found that dialects emerged in a spatially organized populations of agents.[1] In Kirby and Hurford's Iterated Learning Model (Kirby & Hurford, 2001) language acquisition imposes a bottleneck on transmission that regularizes language. Because each learner could not possibly be exposed to all possible form-meaning mappings in the language of the adult population, those mappings that can be captured by generalizations have a better chance of propagating from one generation to the next (Hurford, 2000). Such iterated learning can give rise to structural hallmarks of human language such as recursive compositionality and regular versus irregular morphology (Kirby, 2000, 2001).

Multigenerational models demonstrate that certain kinds of conventions can arise in a community without the benefit of common knowledge. But because these models operate on a generational time scale, they are perhaps best suited to explain processes of grammaticalization, which tend to be slow in nature, sometimes requiring centuries before innovations become fully general (Aitchinson, 2001). However, these models are less well-suited to explain lexical

innovations, where generalization can take place considerably more rapidly. Within a single generation, new coinages can take root and spread throughout a relatively stable community of language users (e.g., the emergence and generalization of Internet-related terms such as *website* in the last decade). Therefore, mechanisms beyond cross-generational learning must be invoked to fully explain lexical conventions.

1.1.2. Emergence-through-use models: linguistic conventions are established and sustained through use

The "emergence-through-use" approach poses a more direct challenge to Common Knowledge theory, suggesting that the conventions of language can emerge as a by-product of the individual communicative actions of agents. According to this view, the origin of symbolic conventions lies in the act of communication itself—in the efforts of individual agents to understand, and to be understood by, their interlocutors. Unlike the common knowledge theory, however, agents have no explicit global representation of community behavior. Instead, each agent has a lexicon which they update on the basis of individual successes or failures to communicate. Learning continues throughout the agent's lifetime, in contrast to multi-generational models where learning only occurs when an agent enters the community.

In a typical model, agents play a communication game that is similar in logic to the referential communication tasks used by psychologists (e.g., Krauss & Weinheimer, 1964, 1966). In each round of the game, a speaker agent encodes an intended meaning for one or more listener agents, who in turn attempt to decode the intended meaning. The lexicon is updated based on the success or failure of the exchange. Studies have shown that these systems can spontaneously give rise to phonological (de Boer, 2002) and syntactic or semantic conventions (Batali, 1998; Hazlehurst & Hutchins, 1998; Hutchins & Hazlehurst, 1995; Steels, 1996, 1998, 2002b), merely as a by-product of the interactions themselves.

The behavior of systems of interacting agents is often clarified by analogy to the processes of self-organization that are found in insect or non-human animal societies. For instance, Reynolds (1987) demonstrated that the flocking behavior of birds can be simulated by assuming that individual birds make local adjustments based on the velocity and bearing of neighboring birds. Thus, despite appearances, such complex flocking behavior need not be generated by individual birds following a leader bird. Likewise, individual language users adapt their language use locally to match that of their individual conversational partners, instead of to the global standards in their community. These local coordination processes serve to maintain cohesion in semantic space for the language community, just as local processes maintain cohesion in physical space for a flock of birds.

Emergence-through-use models have a powerful appeal because they require no specialized mechanisms beyond simple learning mechanisms and the feedback mechanisms involved in communicative interactions. Furthermore, the basic assumptions of the models are consistent with a growing body of psycholinguistic data which finds that coordination processes in the dyad serve to align the conceptual and linguistic representations of language users (Barr, 1999; Barr & Keysar, 2002; Branigan, Pickering, & Cleland, 2000; Brennan & Clark, 1996; Garrod & Anderson, 1987; Garrod & Doherty, 1994; Markman & Makin, 1998; Pickering & Garrod, in press). The idea that processes in the language dyad could be responsible for coordination at the community-level is appealing because of its parsimony and strong empirical motivation.

However, because current models were designed to investigate aspects of language evolution, they do not directly address a central problem addressed by the common knowledge theory of conventions: The generalization of a regularity on the basis of limited knowledge. In Lewis' theory, it is common knowledge that allows agents to generalize a behavior beyond their own experiences, and individuals can perform this generalization even when their experience with other agents is limited to a small sample. Thus, common knowledge serves as a surrogate for direct interaction. But in previous models, agents were provided with ample opportunities for direct interaction. Agents sampled indiscriminately from the entire population, unlike in human populations, where the range of sampling is strongly constrained by individual preferences and the structure of the community. Furthermore, many of the simulations used small population sizes (e.g., 5–15 agents in Hutchins & Hazlehurst, 1995 and 30 agents in Batali, 1998). Given the long run times in these simulations, every agent would have had multiple opportunities to interact directly with every other agent. Thus, they do not really test whether common knowledge is necessary for convergence to occur in situations where repeated direct interaction is not possible. Similarly, in Steels' simulations of the self-organization of the lexicon (Steels, 1996, 1998, 2002b) each agent had an unlimited memory that tracked the performance of a word–meaning association over that agent's entire interactional history. Because each agent sampled uniformly from the population, these unlimited memories might effectively track frequency of use in the population, a kind of global information. In summary, while findings such as these provide important insights into aspects of language evolution, these models cannot speak to whether common knowledge would be necessary for generalization to occur in cases where agents' abilities are more limited and their knowledge is more local in nature. Thus, the question that the current study explores is, What is the minimal amount of knowledge necessary for coordinated communication to emerge in a community of game-playing agents?

1.2. The present study

To address this question, I conducted a set of multi-agent simulations in which I manipulated agents' opportunities to gain global knowledge about conventional practices. Specifically, in this study the following variables are manipulated: (1) agent learning regime; (2) agent memory size; (3) population size; and (4) community structure (a variable pertaining to how agents sample partners from the community). Two sets of simulations were conducted to demonstrate the robustness of the emergence-through-use phenomenon over broad ranges of these variables. In the first set, population size and agent learning regime were varied, with agents randomly sampling partners from the entire population. Convergence was observed with short run times even when agents' memories were extremely limited. In fact, it was observed that larger memories produced inferior convergence, challenging the idea that more global knowledge leads to more efficient convergence. In the second set of simulations, the population was spatially organized on a two-dimensional plane, with each agent sampling partners according to a Gaussian function that related the probability of interaction inversely to the distance between the two agents. In these simulations, each agent's "neighborhood size" was manipulated by changing the breadth of this function. While convergence to a single symbolic scheme was a less likely outcome than in the previous simulations, the populations developed spatially organized 'dialects.' Counterintuitively, convergence was sometimes better

when neighborhood sizes were small; i.e., when agents played the game repeatedly with the same players instead of with different players every time.

2. The signaling game

This section describes the signaling game used in the simulations. The formalism adopted here is based on Lewis' analysis of signaling games (cf. Lewis, 1969, chapter IV). Agents played the game in pairs, with one agent, the "speaker," encoding a message for the second agent, the "addressee," to decode.

The game consists of a population of agents, P; a set of four meanings, M: $\{m_1, m_2, m_3, m_4\}$; and a set of four forms, F: $\{f_1, f_2, f_3, f_4\}$. Every individual agent A_K from P has a "lexicon," or mapping function, G_k, that maps meanings onto forms. Any act in which A_K uses G_k to produce a form corresponds to an act of "production." The mapping function is bi-directional or Saussurean (cf. Hurford, 1989); that is, the function's inverse can be used to produce a meaning given a form, corresponding to the process of "comprehension."

We say that a population has *converged* when all agents use the same mapping function, G_t. This target function is not pre-ordained in any way, but is defined circularly as the function that everyone ends up using. In game-theoretic approaches, agents strive toward this equilibrium by adapting their behavior to mutual expectations. In other words, they choose the function that they expect others to choose based on common knowledge (Lewis, 1969). However, in the emergence-through-use approach, this goal itself is *not* programmed into the agents; all they are programmed to do is to update their mappings based on observed feedback from specific partners.

It is assumed in these simulations that agents have a rudimentary ability to evaluate the success or failure of their communication. This evaluation is reflected in a reinforcement signal b indicating success or failure (activation of the same or different meanings) that is provided to the agents. Such a reinforcement signal is the minimum necessary for learning. However, unlike simulations where observational learning is used, agents are not directly trained on the specific meaning activated by the other party.

At each epoch, dyads are formed through a stochastic process. Each agent A_K has a selection function, D_K, that determines the probability that it will speak to any other particular agent in the community. For instance, $D_K(A_G)$ is the probability that agent A_K will attempt to speak to with agent A_G. In the simulations below, we explore two different selection functions. To simplify matters, we will use symmetric functions such that $D_K(A_G) = D_G(A_K)$.

Training was organized into a series of epochs. The algorithm for a single epoch is:

1. Choose a speaker, A_S, at random from the population.
2. Use that agent's selection function, D_S, to select a listener, A_L. Agents are sampled without replacement such that once an agent is selected as either speaker or listener, it is removed from the pool of eligible agents for that epoch.
3. Randomly permute set M to create a sequence of meanings to be communicated, M'.
4. For each m_i in M':
   (a) A_S passes meaning m_i through its mapping function G_S to obtain form f_j.

(b) A_L passes the observed form f_j through its mapping function G_L^{-1} to obtain meaning m_p.
(c) Compare m_i to m_p to obtain success or failure b.
(d) Apply update functions based on outcome b to obtain G'_S and G'_L.

5. Repeat steps 1–4 until all (or nearly all; see Simulation Set 2) agents have been assigned to dyads.

Clearly, the dynamics of a given simulation will depend upon the nature of the agents' update functions. In the current simulations, the effects of two different update functions on system dynamics are explored. These functions are called Reinforcement Learning and the Stay–Switch algorithm.

2.1. The Reinforcement Learning model

In multi-agent simulations of language evolution it is common to instantiate agent learning in the form of connectionist networks (e.g., Batali, 1998; Hutchins & Hazlehurst, 1995; Oliphant, 1999). Such a practice is adopted here to provide comparability with previous work. Under the Reinforcement Learning regime used in these simulations, the mapping function for each agent is implemented by a one-layer neural network of weights, W. The network consists of two input, or meaning nodes, $C = \{c_1, c_2\}$, and four output, or form nodes $F = \{f_1, f_2, f_3, f_4\}$. Meanings are represented as distributed patterns of activation along the input nodes.[2] Thus:

$$M = \{\{0, -1\}, \{-1, 0\}, \{1, 0\}, \{0, 1\}\}.$$

Forms have a localist representation on the output nodes. Each network begins life with its weights set to small, random real values. The weights are normalized before any learning and after every update.

During "production," a given pattern of activation m^p from M is clamped onto the speaker agent's input nodes. Activation is passed through the weights to the form nodes, according to the function:

$$f_i = \sum\nolimits_j w_{ij} m_j^p$$

Then, the most strongly activated output node, f_w, is selected as the "winner." During "comprehension," the listener agent observes f_w, activation is clamped to the corresponding output node, and is passed to the input nodes, the result of which is represented by vector C:

$$c_i = f_w W_{wc}$$

A vector, m^w, is selected from set M as the "winner," where m^w is defined as the vector that is closest in Euclidean space to C.

The networks update their weights according to a competitive reinforcement learning algorithm. The algorithm depends upon the observed success of the communication. There are two learning parameters, δ and η, which correspond, respectively, to the learning rate for a correct and an incorrect response. The parameter η also is used to implement the principle of *mutual exclusivity* (Markman & Wachtel, 1988). This principle, which is assumed to govern word learning, asserts that learners strive to maintain one-to-one mappings between symbols

and meanings. In observance of this principle, when an agent successfully uses a particular form to convey a meaning, it will not only strengthen the link between that meaning and form but will also weaken the link between the same meaning and other forms. Likewise, if an agent fails to communicate a meaning with a particular form, it will weaken the corresponding link and strengthen the links between that meaning and other forms.

The Reinforcement Learning update rule is as follows:

If $m^w = m^p$: (correct response)
  where $i = f_w$:
  $\forall_j\ w_{ij} = w_{ij} + \delta\ (m^p_j - w_{ij})$; (move weights closer to the meaning vector)
  where $i \neq f_w$:
  $\forall_j\ w_{ij} = w_{ij} - \eta(m^p_j - w_{ij})$. (move weights away from meaning vector)
If $m^w \neq m^p$: (incorrect response)
  where $i = f_w$:
  $\forall_j\ w_{ij} = w_{ij} - \eta(m^p_j - w_{ij})$; (move weights away from meaning vector)
  where $i \neq f_w$:
  $\forall_j\ w_{ij} = w_{ij} + \eta(m^p_j - w_{ij})$. (move weights closer to the meaning vector)

2.2. The Stay–Switch model

Given the goals of the current study, two objections could be made to the use of neural networks. First, agent lexicons as implemented in the networks can occupy certain "ill-formed" states, in which more than one form is mapped to the same meaning. This is different from human lexicons, where perfect synonymy is avoided and mappings tend to follow a "one form, one meaning" principle of contrast (Clark, 1987). A second possible objection to using neural networks is that the learning in such networks is cumulative. It might be argued that such cumulative learning is akin to maintaining community-level statistics, and therefore that these agents implement a rudimentary kind of common knowledge. In any case, this kind of learning makes it difficult to directly quantify the amount of "social information" agents use in encoding and decoding. The Stay–Switch model was developed to counter these objections, as well as to show the generality of the self-organization phenomenon.

The Stay–Switch model, which is a form of "best-response" model (Young, 1998), implements discrete, memory-dependent learning. The model is somewhat similar to that used by Steels (1996), in that each agent has a lexicon with discrete mappings and a basic memory that tracks the performance of each of these mappings. However, the mapping update procedure differs, and the size of the agent memory was varied. The memory size parameter allows for the exploration of the issue of whether coordination is best achieved when agents adjust their behavior based on local versus global information.

The update algorithm implemented by these agents is as follows. When an agent successfully uses a form to communicate a meaning, it simply tallies its success and retains its current set of mappings (*stay*). On the other hand, when an agent miscommunicates it will, with some probability, seek to exchange the unsuccessful mapping that produced the error with the least reliable mapping in its lexicon (*switch*). The probability of a switch is given by the proportion of previous uses of the mapping within the agent's stored history that led to failure.[3] Thus,

agents are conservative in the sense that they are reluctant to switch a mapping that has worked well in the past even though it may have failed in the current interaction. In implementing a switch, the agent consults its memory and ranks the forms from best to worst according to their success, and switches the meaning of the offending form with that of the worst performing form from that set. If there is no clear worst mapping, the agent will select randomly among the mappings with the lowest rank. Upon switching, the memory for each of the swapped forms is cleared.

3. Simulation set 1: conventions in unstructured populations

The first set of simulations investigates the following three questions: (1) Can a population converge to a common set of symbolic conventions even when the population size is large enough to preclude repeated direct interaction between all members? (2) Does this convergence depend upon learning regime? (3) How much "social information" must each agent have access to for this convergence to take place? In this set of simulations, unstructured communities were used in which each agent had an equal probability of interacting with every other agent in the population.

First, results for the Reinforcement Learning model are considered. Before running the main simulations, a parameter search was conducted to determine optimal parameter settings for the update algorithm as well as to explore the robustness of convergence. Ten runs of 100 agents each were conducted at each .02 interval for each parameter (including only cases where δ was greater than η). If a given run failed to converge after 10,000 epochs, the run was terminated. In the parameter search, convergence was observed at least once per 10 runs in 28% of all of the settings explored, over a large range of δ and η. However, it was observed that convergence was most robust at settings in which the ratio of δ to η was between 3 and 10. The values of δ and η that yielded the best performance were .76 and .16, respectively. The results presented in this section are based on these parameters.

At each population size setting, 100 runs were conducted. Fig. 1 presents the mean efficiency of convergence for Reinforcement Learning populations by population size.[4] Without exception, all populations eventually converged to a single signaling system before the 10,000 epoch criterion. The efficiency of convergence had a non-linear relationship with population size. At the largest tested population size of 10,000, a single signaling system was reached after only 178 interactions, only 86 epochs later than for a population of 100. At 178 epochs in a population of 10,000 agents, each agent could only have interacted with a maximum of 2% of possible partners. This shows that convergence can take place under conditions of partial knowledge where direct interaction is not possible.

The dynamics of Reinforcement Learning populations were explored by examining the proportion of agents in each epoch that spoke the target language that eventually won out. This variable is plotted in Fig. 2. The sigmoidal shape of the function is similar to a class of critical mass or "dying seminar" models identified by Schelling (1978) wherein agents have a preference to conform to some course of action contingent upon the number of other agents conforming to the same action. This characteristic pattern of behavior has been assumed to arise through the game-theoretic preferences and strategic reasoning of agents who base their

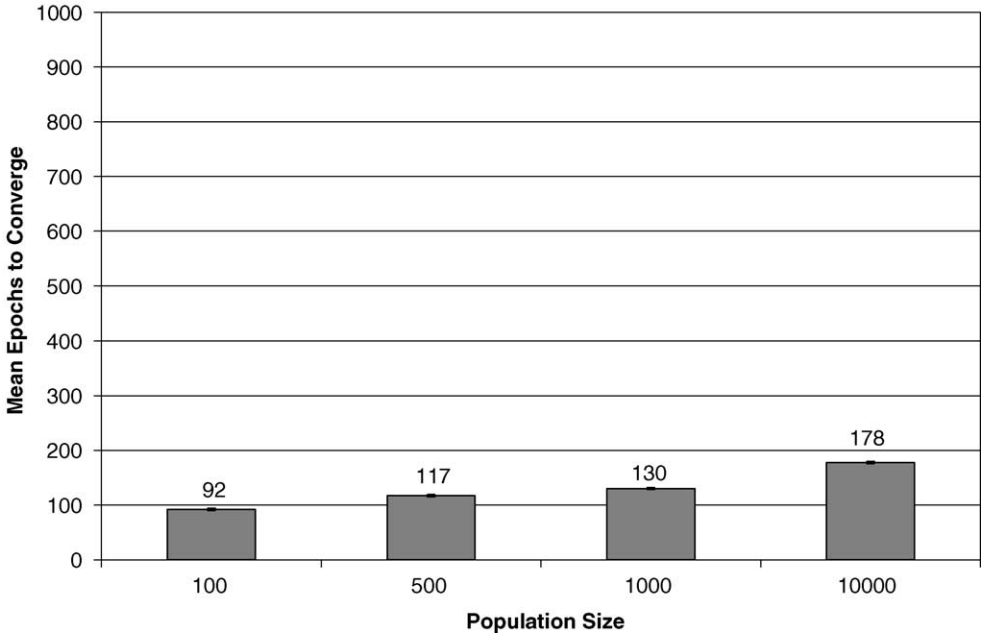

Fig. 1. Mean epochs to converge, Reinforcement Learning model.

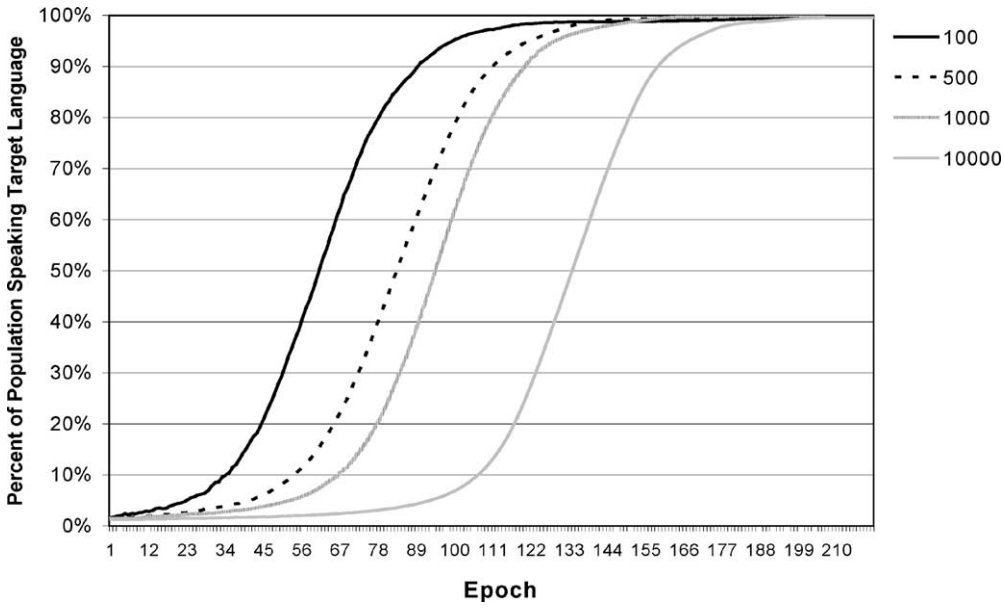

Fig. 2. Percentage of agents speaking target language by epoch and population size, RL model.

decisions on the expected global behavior of the community. However, the simulations show that this pattern of behavior can emerge as a product of local interaction.

Because the reinforcement learning agents engage in a form of cumulative learning, it is not clear how to quantify the breadth of social information encoded by the agent. In contrast, the Stay–Switch agents allow the direct quantification of this variable through the memory size parameter. This makes it possible to determine what extent of local versus global knowledge leads to most efficient convergence.

One-hundred runs were conducted at each of the four population sizes and each of the following memory settings: 1–10, 500, and 1000. For small populations of 100 agents, convergence was most efficient at small values of the memory parameter (Fig. 3). The optimal setting of this parameter was two. In other words, populations of 100 agents arrived at a set of semantic conventions most quickly when each agent updated its weights considering only the current and the previous two interactions. However, when the memory size was larger than 10, full convergence never occurred before the criterion—and did not do so even when the memory was large enough to accommodate an agent's entire history of interaction. Thus, not only does this set of simulations show that convergence can occur without global knowledge, but it actually suggests that there are circumstances in which attempts to use global knowledge actually would *impair* community-level coordination.

However, in contrast to the Reinforcement Learning simulations, convergence was rarely observed in Stay–Switch populations larger than 100 agents. For populations of 500 agents, convergence was only observed at a memory setting of two, with a mean convergence epoch

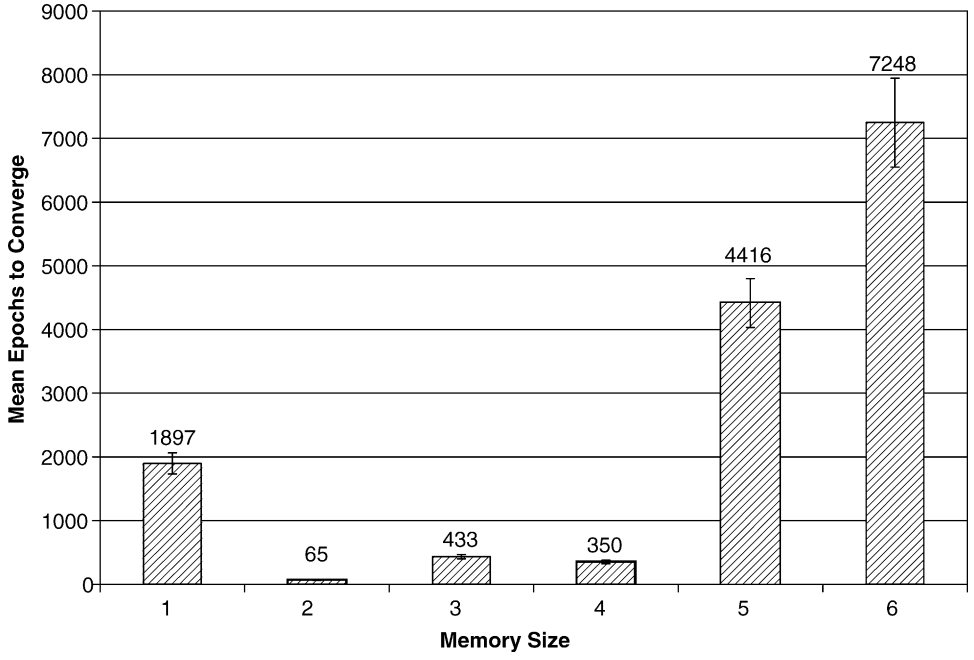

Fig. 3. Efficiency of convergence in the Stay–Switch model by memory size, with a population of 100 agents. Error bars represent the standard error of the mean.

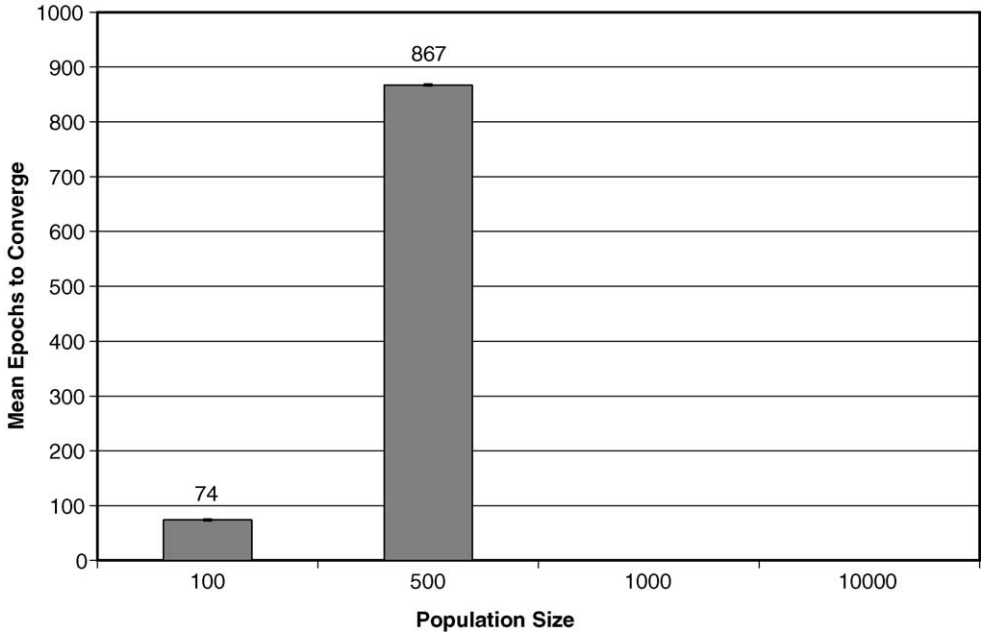

Fig. 4. Mean epochs to converge, Stay–Switch model.

of 867 (*SD* = 73). Finally, convergence was never observed at any of the memory settings in the two larger populations (1000 and 10,000) (Fig. 4).

The source of these differences in performance between the two learning regimes is not entirely clear. One possibility is that the superior performance of Reinforcement Learning is due to the continuous nature of the semantic representation. While Stay–Switch agents are always in 1 of 24 possible states, corresponding to a well-formed set of form-meaning mappings, Reinforcement Learning agents can be in any infinite number of intermediate states. Thus, each agent can be a more sensitive barometer of changes going on in the system.

To summarize, it has been shown that under some circumstances, a population of "egocentric" agents can develop and sustain coordinated signaling systems as a by-product of the parallel dyadic coordination efforts that are distributed across the community. This convergence can occur even when the population size is large enough so that repeated direct interaction among all members is not possible. The simulations also show that it is not always the case that the more social information each agent has access to, the better the convergence. In smaller populations, performance was best when each agent based its decisions on strictly local information. Thus, limitations on memory can be adaptive for community-level coordination (for a similar result, see Shoham & Tennenholtz, 1997). A reason for this "less is more" phenomenon is that populations of agents are shooting at a moving target. Thus, for convergence to obtain in memory-dependent learning, there needs to be some congruence between the size of the agent memory and the rate of change of the population. When the rate of change is rapid, previous communicative episodes quickly lose their predictive value, because the system may have already moved into a different region of the state space.

In the current simulations, agents were equally likely to interact with any other member of the population. This means that each agent obtained a random sample of partners from the population, and such random sampling could be optimal for the generalization process. However, in human societies the selection of interactional partners is biased by geographical and social factors such as the existence of social networks. This biased sampling provides yet another constraint on an individual's ability to gain knowledge about the practices of the community. Intuitively, it would seem that communities in which agents samples narrowly (the same few interactional partners again and again) would be less likely to converge than communities in which agents sample indiscriminately (e.g., new partners every time).

Alternatively, it is possible that this additional constraint might actually improve convergence. Note that in the simulations above, Stay–Switch populations of 1000 agents or more never converged. This performance indicates that under this discrete-learning regime some amount of repeated interaction among individuals might be necessary for a target convention to gain a foothold in the community. In a very large population, such repeated interactions would be extremely unlikely, and thus a critical mass of agents using the same conventional system could never develop. It is possible that limiting each agent's interactions to a small surrounding "neighborhood" of agents might enable a conventional system to become established even in very large populations, and once established, to eventually spread to other agents.

4. Simulation set 2: conventions in structured communities

In this set of simulations, agents inhabited a two-dimensional plane and interacted with other agents depending upon their proximity. Each agent had a selection function that determined the probability that it would interact with another agent.[5] The selection function was Gaussian, such that the probability that an agent S would interact with agent L dropped off exponentially as a function of distance:

$$\frac{\exp(-||S-L||/2\sigma^2)}{\sum_k \exp(-||S-K||/2\sigma^2)}$$

The term $||S-L||$ represents the Euclidean distance between speaker and listener. The parameter σ represents the "spread" or "coverage" of the function, determining its relative steepness or flatness. Effectively, this parameter sets the *neighborhood size* of each agent, with smaller values of σ yielding smaller, more peaked neighborhoods, and larger values, larger and flatter ones. The denominator of the equation normalizes the values such that the probability of interacting with other agents in the population will sum to 1. Note that the spatial organization of agents need not be construed literally as organization in physical space, but could also be considered as representing agents' preferences for social affiliation.

Agents were placed at random locations in a 1024×768 grid (with each agent occupying an 18×18 pixel square). Performance of the models was examined under six settings of σ: 10, 15, 20, 25, 30, and 35 pixels. Neighborhood size was defined as the average number of distinct agents that a typical agent will interact with over 100 epochs. At each of these settings, the neighborhood sizes were effectively: 3, 5, 9, 14, 20, and 27. For simplicity, the population size was fixed at 1000, and the update function parameters from Set 1 were used.

As in the previous simulations, the results presented here are averaged over 100 runs for each learning model at each population size. If a population had not yet converged by 1000 epochs, training was terminated. Fig. 5 presents the efficiency of convergence by population size, including only those populations that converged by 1000 epochs. Recall that in Simulation Set 1, where populations were unstructured, Stay–Switch populations including more than 500 agents never converged to a single system. In this case, Stay–Switch showed improved performance, even though the population size was 1000. This suggests that Stay–Switch is most effective under circumstances where agents limit their interactions to a small portion of the entire population. Of the settings tested, Stay–Switch was optimal at a neighborhood size of nine agents. At this setting, 85% of the runs converged to a single signaling system, at an average of 280 epochs. In contrast, at larger neighborhood sizes, convergence was slower and less likely. For instance, at a neighborhood size of 27 convergence was observed in only 44% of the runs, on average at 526 epochs.

The opposite was the case for Reinforcement Learning—convergence was worst at smaller and best at larger neighborhood sizes. At the smallest setting of 3, populations never achieved perfectly coordinated signaling. At the largest setting of 27, only 40% of the populations converged before 1000 epochs, with an average efficiency of 302 epochs. In contrast to Stay–Switch, performance degrades when interaction is limited to small neighborhoods.

In sum, discrete, memory-dependent learning worked best in small, unstructured populations and in larger populations that were socially structured. What these two circumstances have in common is that they provide agents the greatest opportunity to interact repeatedly with a set of other agents. As suggested above, this repeated interaction may be necessary for the population to establish a critical mass of agents speaking a single target language. Once this critical mass has been established, a positive feedback loop between success and use will drive the population toward convergence.

Even when communities did not converge to a single system, as was often the case, they still exhibited some degree of self-organization. Fig. 6 shows the number of signaling systems used by more than 1% of the population (10 agents) by learning model, neighborhood size, and epoch.[6] Both models showed increasingly greater organization by epoch regardless of neighborhood size, though there were vast differences in efficiency both within and between the learning models. Stay–Switch exhibited most efficient performance at settings of 5 and 9, while larger settings yielded ever-worse performance overall. At a setting of 27, Stay–Switch populations cut down the number of signaling systems to 12 by epoch 1000, and the slope of the graph suggests that performance would continue improving with further training.

An interesting phenomenon observed in the simulations was that even when populations failed to converge to a single target language, they inevitably exhibited spatially organized "dialects." Two examples are provided in Fig. 7. As one might expect, the number and size of dialects was inversely proportional to neighborhood size. Agents that inhabited regions near the corners of the plane tended to form more stable dialects than did agents located toward the center, because the former agents were somewhat insulated from the influences of other subcommunities. Subcommunities that formed in the middle of the plane tended to eventually become swallowed up by these subcommunities at the borders. Furthermore, the dialect of a subcommunity typically exhibited similarities to the dialects spoken by other, nearby

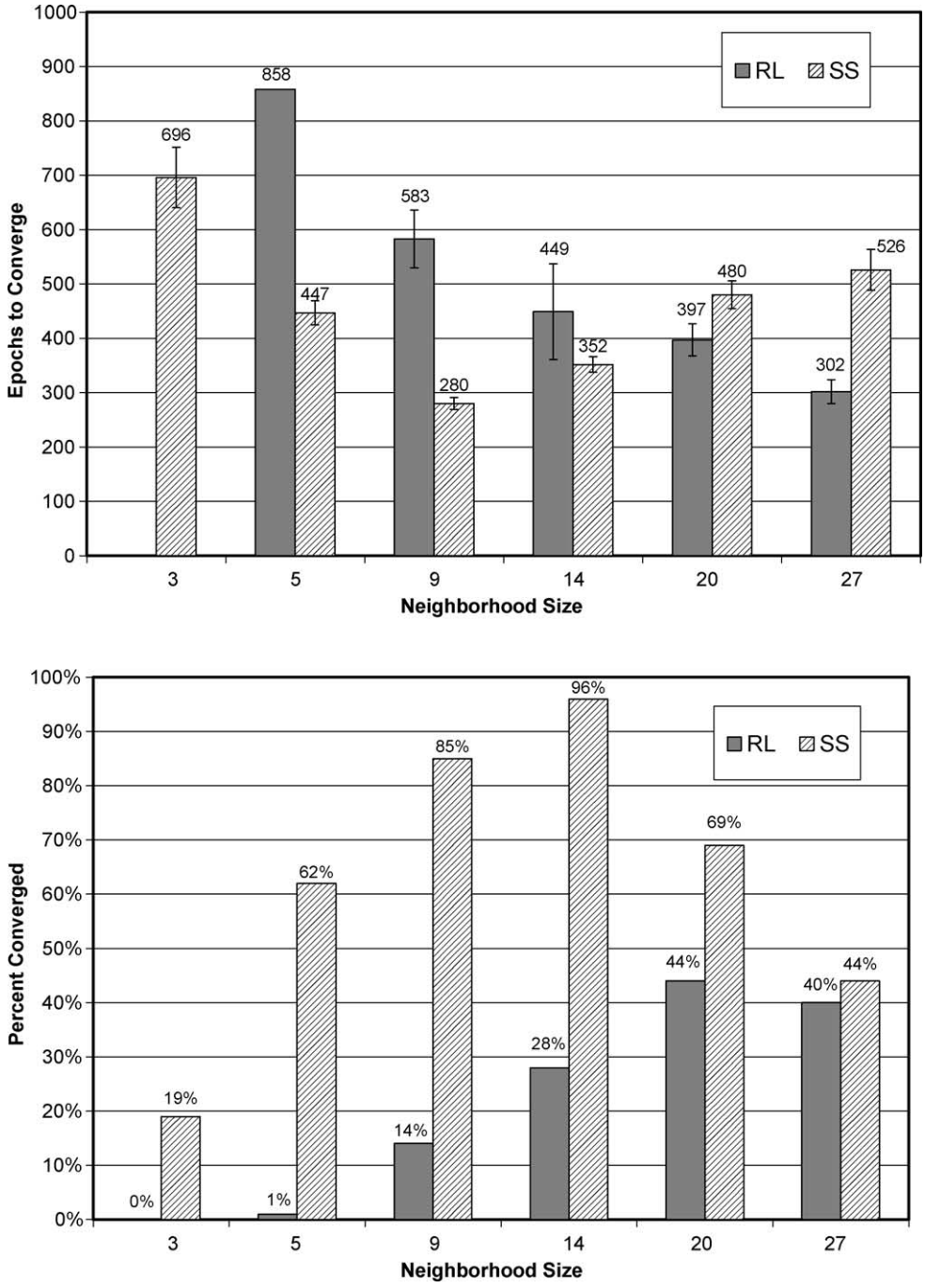

Fig. 5. Efficiency of convergence (top panel) and percent convergence to a single system (bottom panel) by neighborhood size and learning model.

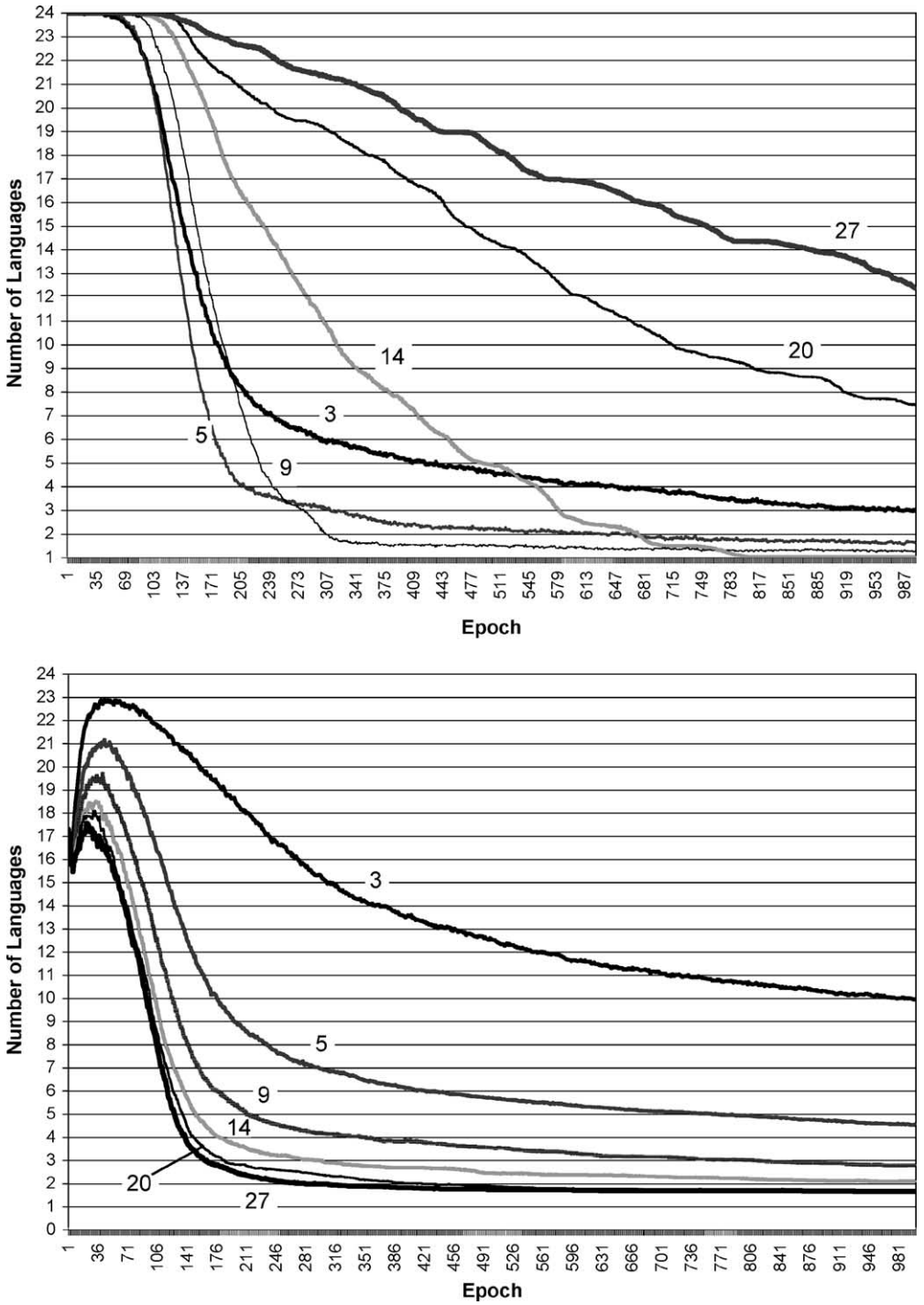

Fig. 6. Average number of languages spoken by over 1% of the population by neighborhood size, for the Stay–Switch (top) and Reinforcement Learning (bottom) models.

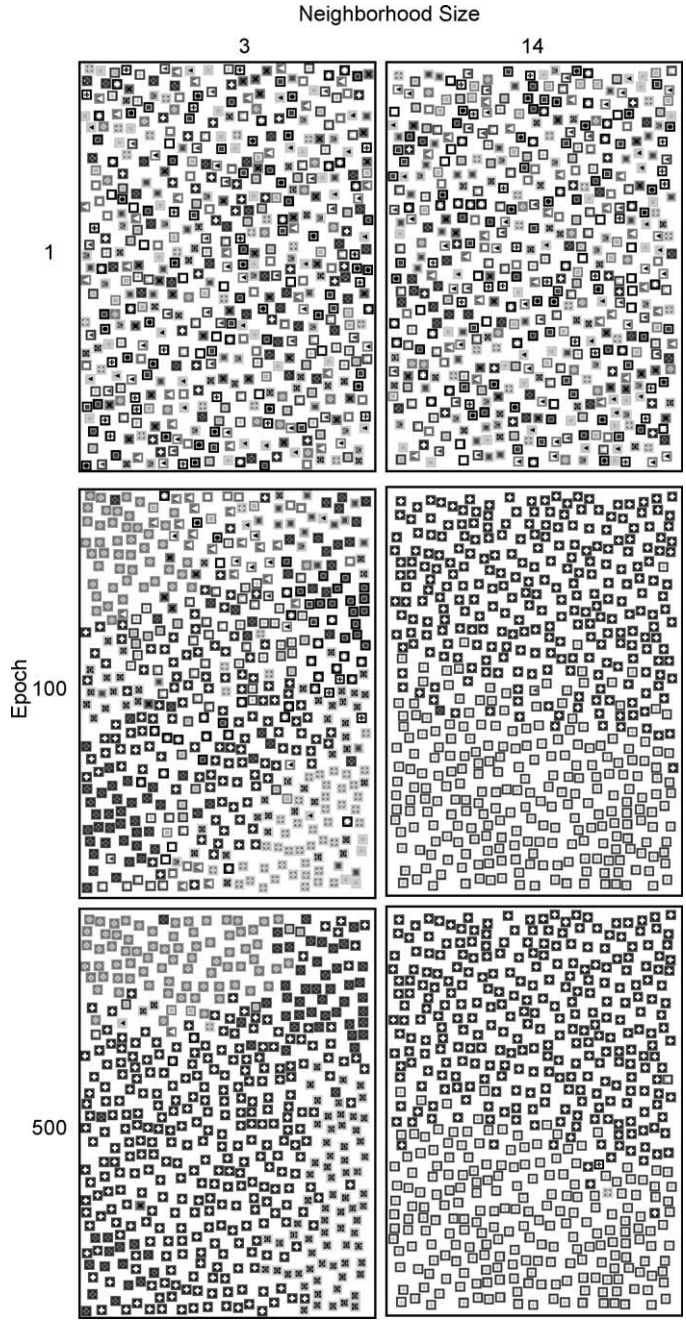

Fig. 7. Emergence of dialects.

subcommunities. Because of these similarities, even when there was no single communication system, the overall coordination of the population was quite high.

The current study is not the first to show that dialects can arise through processes of cultural learning (see Livingstone, 2002 for a review). However, previous studies showing dialectical organization of semantic conventions found that they emerged over multiple generations (e.g., Kirby, 1998; Livingstone & Fyfe, 1999). In contrast, the semantic conventions in the current study emerged in a closed population of agents, as a by-product of interaction. Livingstone (2002) reports similar findings for phonological conventions. In his study, which used a modeling framework established by de Boer (2002), a closed population of 20 spatially organized agents learned vowel sounds from one another through imitation. A "dialect continuum" emerged, wherein the vowel systems of neighboring agents were more similar than those of more distant agents.

To summarize, the second set of simulations shows that convergence is likely even when agents' sampling of partners from the community is not representative but biased to a small number of partners in the agent's neighborhood. Although this biased sampling imposed another limitation on each agent's ability to acquire global knowledge regarding community practices, convergence took place even in large populations. The performance of populations of agents using memory-dependent learning improved markedly from the last set of simulations, even though their opportunities for social interaction were more strongly constrained. Reinforcement Learning populations showed a different trend: as social connectivity increased, convergence became more likely. Taken together, these results demonstrate that more knowledge is not always better. Under certain circumstances, limitations on agents' social knowledge may actually be adaptive for community-level coordination. Lastly, even when convergence to a single system was not achieved, self-organization of communication was still observed in the form of dialects.

5. General discussion

Supporting the emergence-through-use view, it was found that large populations of agents can establish and sustain conventional signaling systems without common knowledge. The agents in the simulations were "egocentric" in how they produced and interpreted signals, in the sense that they were indifferent to the identity of their conversational partners. In fact, the agents even lacked a concept of community. Nonetheless, in most cases convergence to a single language or to a set of spatially organized dialects was an extremely likely outcome. Unlike previous studies, in which repeated, direct interaction among individuals could have rendered common knowledge unnecessary, in the current simulations convergence was observed when agents had limited opportunities to gather global information, either due to repeated interactions with the same individuals or to memory limitations.

The efficiency and likelihood of convergence seemed to depend upon learning regime. A continuous, neural-network style of learning produced optimal convergence in broadly interconnected populations, but failed to reliably converge when agents were organized into very small neighborhoods. In contrast, Stay–Switch populations performed best when agents updated their lexicons based only on highly local information. Specifically, performance was

optimal when agents considered only their current and last two interactions. As agents' abilities to track global properties of the system were improved through larger memories or broader social connectivity, convergence was impaired. Since common knowledge is a kind of global knowledge, this suggests that at least under some circumstances, agents' attempts to acquire and use common knowledge might actually work *against* the establishment of conventions.

In the current study, agents have things in common such as the same learning rate, a fixed number of forms and meanings, and so on. Can they be said to have common knowledge? Although agents have things in common, they do not have common knowledge because they do not represent what other agents believe, or what is generally accepted in the community of agents. This distinction between knowledge and meta-knowledge (knowledge about knowledge) is crucial. In Lewis' theory, agents will encode or interpret a signal in a certain way *because they know that it is shared*. For example, imagine that you believe that it is common knowledge in your community that people solve a problem by conforming to some regularity *R*, but you happen to have just interacted repeatedly with an individual who did not conform to *R* but to *S* instead. In spite of this experience, you would have little reason to expect the next person in your community to conform to *S* over *R*, even though this option has just been made salient. In other words, common knowledge of a convention will override the salience of another alternative (Lewis, 1969). However, the agents in this simulation would go on conforming to *S* with the next person because they just do what is salient. According to game theory, that kind of behavior is non-optimal.

The theory of games was first developed as a theory of single interactions among dyads or groups and was only later generalized by Lewis (1969) to repeated interactions at the community level. The goal here is not to question the appropriateness of mutual expectations theory for the situation in which a player encounters a series of isolated partners who do not form a community or share information with each other. In that case, coordination on the basis of mutual expectations seems optimal. Yet what is optimal or rational for agents to do at the level of the individual dyad and at the level of the community need not coincide. For example, Lewis asserted that individuals base their judgments on what is conventional because "the salience of an equilibrium is not a very strong indication that everyone will choose it" (Lewis, 1969, p. 57). However, the simulations demonstrate that salience can be a valid heuristic for individuals organized into a community because interactional processes can cause representations in the community to converge. In many cases, the similarity of these representations would obviate the need for strategic reasoning. Against this background of common representation, the routinely egocentric behavior observed in actual language users (e.g., Keysar et al., 2000) makes sense.

Although meta-level knowledge about community practices is not necessary for a shared communication system to emerge, this does *not* mean that meta-knowledge of conventions does not exist, nor that this kind of knowledge does not play *any* role in language use. As Gilbert (1995) observed, one possibly unique characteristic of dynamical systems that are composed of human agents is that humans can learn about emergent properties of the systems they inhabit and strategically exploit this knowledge. In support of this view, research finds that people are in fact quite good at estimating what other people in their community know, though these estimates are systematically biased toward the estimator's own knowledge (Fussell & Krauss, 1991; Krauss & Fussell, 1991). For people to make such assessments, they must have some concept of belonging to a community. However, note that the perception of belonging

to a community might be a *consequence* of the establishment and recognition of common practices, rather than a *pre-condition* for their emergence.

The simulations presented here provide a demonstration that certain communication systems analyzed by Lewis as requiring common knowledge (Lewis, 1969, cf. chapter IV) can be explained by principles of self-organization. Still, the communication games presented here are vastly simpler than the kinds of coordination problems humans face in using language. In contrast to human languages, which are open-ended, the agents live in a fixed linguistic universe where there are only four forms and four meanings. Steels and colleagues (see Steels, 2002a for a review) have attempted to tackle the open-ended nature of lexical forms by using more sophisticated language games, and endowing agents with a secondary, non-verbal channel of communication as well as primitive sensorimotor abilities through which they can ground symbolic meanings. In these so-called "Talking Heads" experiments, pairs of agents view a visual scene through a camera and must communicate about a target object. The agents have no prior conceptual repertoire, and can invent new forms when necessary to communicate a distinction. In these models, conceptual structure and the lexicon co-evolve as a by-product of interaction among agents. However, in these simulations, everyone talks to everyone else with equal probability, and each agent tallies the success of each form every time that is used. The results of the current study suggest that limiting agents' access to social information through smaller memories and spatial organization might actually improve the overall performance of these models.

Important advances in the study of the evolution of language could be achieved by greater integration of these efforts with others both within the same area and from other fields. First, modelers have studied the emergence of syntax and the lexicon through multi-generational approaches as well as via emergence-through-use. These two approaches are clearly not mutually exclusive, and both are probably at work in promoting stability and change in human languages. For example, it is possible that multi-generational iterated learning shapes structural aspects of signaling systems, while processes at the interactional level promote conformity in lexical representations. An important avenue for future research is to understand how these approaches might be combined to account for phenomena of linguistic change.

Second, it would be useful for computational work to be more closely integrated with the growing body of psycholinguistic work investigating negotiation and alignment in dyads (Barr & Keysar, 2002; Brennan & Clark, 1996; Clark & Brennan, 1991; Garrod & Anderson, 1987; Markman & Makin, 1998; Pickering & Garrod, in press; Schober & Brennan, 2003). As more sophisticated models of coordination in dialogue arise, these models can be directly embedded in agent systems. This would help empirically ground many of the assumptions about learning and communication implemented in multi-agent models. In turn, multi-agent modeling will provide researchers in the area of language use with better insight into the broader informational environment in which their models must operate, and can serve as a testbed for these theories.

To date, theories of language use have developed under an acute preoccupation with the difficulties of achieving shared understanding in the face of the ambiguity of language. This preoccupation has led theorists to posit specialized cognitive mechanisms for dealing with this ambiguity. However, the development of theories of language use should also be guided by an appreciation for how distributed acts of coordination can promote a commonality of

representation in the language community. Armed with such knowledge, we will be in a better position to assess the degree of ambiguity that language users actually face and the strategies that enable them to overcome it.

Notes

1. Although agents were trained on one another's communicative behavior both across generations and within generations, within-generation training was found to be unnecessary for convergence (Livingstone, 2002).
2. For diagnostic purposes, two instead of four input units were used, enabling the weights to be visualized as projections onto a 2D space. A set of runs using four meaning nodes was conducted for Set 1, and yielded qualitatively similar results, with a faster time to converge.
3. For the first epoch, in which the agents' memories are empty, the probability of a switch was set to .75.
4. The standard error for each of these means is approximately 2 epochs.
5. One side effect of this function is that a few agents are left without a partner in a given epoch because all potential nearby partners are already taken.
6. The plots in the Reinforcement Learning chart (bottom of Fig. 5) do not begin at 24 at epoch one because the random initialization of agents' weights causes some agents to inhabit intermediate states that do not correspond to any of the 24 conventional signaling systems.

Acknowledgement

This research was conducted while the author was a postdoctoral fellow at the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign. The author thanks the Beckman Institute for their support during the development of this project, and Gary Dell, Boaz Keysar, and Terry Regier for their valuable feedback.

References

Aitchinson, J. (2001). *Language change: Progress or decay?* Cambridge: Cambridge University Press.

Arnold, J. E., Trueswell, J. C., & Lawentmann, S. M. (1999). Using common ground to resolve referential ambiguity. Paper presented at the 40th Annual Meeting of the Psychonomic Society, Los Angeles, CA.

Barr, D. J. (1999). *A theory of dynamic coordination for conversational interaction*. Unpublished Ph.D. thesis, The University of Chicago.

Barr, D. J., & Keysar, B. (2002). Anchoring comprehension in linguistic precedents. *Journal of Memory and Language, 46*, 391–418.

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

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*, B13–B25.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *22*, 1482–1493.

Cangelosi, A. (2001). Evolution of communication and language using signals, symbols and words. *IEEE Transactions in Evolutionary Computation*, *5*, 93–101.

Cangelosi, A., & Parisi, D. (1998). The emergence of a 'language' in an evolving population of neural networks. *Connection Science*, *10*, 83–97.

Christiansen, M. H., & Ellefson, M. R. (2002). Linguistic adaptation without linguistic constraints: The role of sequential learning in language evolution. In A. Wray (Ed.), *The transition to language*. Oxford: Oxford University Press.

Clark, E. (1987). The principle of contrast: A constraint on acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 1–33). Hillsdale, NJ: Erlbaum.

Clark, H. H. (1992). *Arenas of language use*. Chicago, IL, USA: University of Chicago Press.

Clark, H. H. (1996). *Using language*. Cambridge, England, UK: Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC, USA: American Psychological Association.

Clark, H. H., & Carlson, T. B. (1981). Context for comprehension. In J. Long & A. Baddeley (Eds.), *Attention and performance IX* (pp. 313–330). Hillsdale, NJ, USA: Lawrence Erlbaum Associates.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshe, B. L. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge: Cambridge University Press.

de Boer, B. (2002). Evolving sound systems. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 79–97). London: Springer.

Fussell, S. R., & Krauss, R. M. (1991). Accuracy and bias in estimates of others' knowledge. *European Journal of Social Psychology*, *21*, 445–454.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*, 181–218.

Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, *53*, 181–215.

Gerrig, R. J. (1986). Process models and pragmatics. In N. E. Sharkey (Ed.), *Advances in cognitive science: Vol. 1* (pp. 23–42). New York: Wiley.

Gilbert, N. (1995). Emergence in social simulation. In N. Gilbert (Ed.), *Artificial societies: The computer simulation of social life* (pp. 144–155). London: UCL Press.

Hanna, J. E., Trueswell, J. C., Tanenhaus, M. K., & Novick, J. M. (1997). Consulting common ground during referential interpretation. Paper presented at the 38th Annual Meeting of the Psychonomic Society, Philadelphia, PA.

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*, 373–424.

Holland, J. H. (1998). *Emergence: From chaos to order*. Reading, MA: Perseus Books.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*, 91–117.

Hurford, J. R. (1989). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua*, *77*, 187–222.

Hurford, J. R. (2000). Social transmission favours linguistic generalization. In C. Knight, M. Studdert-Kennedy, & J. R. Hurford (Eds.), *The evolutionary emergence of language: Social function and the origins of linguistic form*. Cambridge: Cambridge University Press.

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* (pp. 157–189). London: UCL Press.

Keysar, B. (1994). The illusory transparency of intention: Linguistic perspective taking in text. *Cognitive Psychology*, *26*, 165–208.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*, 32–38.

Keysar, B., Barr, D. J., Balin, J. A., & Paek, T. S. (1998). Definite reference and mutual knowledge: Process models of common ground in comprehension. *Journal of Memory and Language*, *39*, 1–20.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*, 25–41.

Kirby, S. (1998). Fitness and the selective adaptation of language. In J. R. Hurford, M. Studdert-Kennedy, & C. Knight (Eds.), *Approaches to the evolution of language* (pp. 359–383). Cambridge: Cambridge University Press.

Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight, J. R. Hurford, & M. Studdert-Kennedy (Eds.), *The evolutionary emergence of language: Social function and the origins of linguistic form*. Cambridge: Cambridge University Press.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions in Evolutionary Computation*, *5*, 102–110.

Kirby, S., & Hurford, J. R. (2001). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language*. London: Springer.

Krauss, R. M., & Fussell, S. R. (1991). Perspective-taking in communication: Representations of others' knowledge in reference. *Social Cognition*, *9*, 2–24.

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, *1*, 113–114.

Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality & Social Psychology*, *4*, 343–346.

Lewis, D. (1969). *Convention: A philosophical study*. Cambridge, MA: Harvard University Press.

Lightfoot, D. (1991). *How to set parameters*. Cambridge, MA: MIT Press.

Livingstone, D. (2002). The evolution of dialect diversity. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 99–117). London: Springer.

Livingstone, D., & Fyfe, C. (1999). Modeling the evolution of linguistic diversity. In D. Floreano, J. D. Nicoud, & F. Mondada (Eds.), *Proceedings of the Fifth European Converence on Artifical Life (Lecture Notes on Artificial Intelligence, Vol. 1674)* (pp. 704–708). Berlin: Springer-Verlag.

Markman, A. B., & Makin, V. S. (1998). Referential communication and category acquisition. *Journal of Experimental Psychology: General*, *127*, 331–354.

Markman, E., & Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121–157.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective taking constraints in children's on-line reference resolution. *Psychological Science*, *13*, 329–336.

Niyogi, P., & Berwick, R. C. (1997). A dynamical systems model for language change. *Complex Systems*, *11*, 161–204.

Oliphant, M. (1999). The learning barrier: Moving from innate to learned systems of communication. *Adaptive Behavior*, *7*, 371–384.

Pickering, M. J., & Garrod, S. (in press). Toward a mechanistic psychology of dialogue. *Behavioral & Brain Sciences*.

Reynolds, C. W. (1987). Flocks, herds, and schools: A distributed behavioral model. *Computer Graphics*, *21*, 25–34.

Schelling, T. C. (1960). *The strategy of conflict*. Cambridge, MA: Harvard University Press.

Schelling, T. C. (1978). *Micromotives and macrobehavior*. New York: Norton & Co.

Schober, M. F., & Brennan, S. E. (2003). Processes of spoken interactive discourse: The role of the partner. In A. C. Graesser & M. A. Gernsbacher (Eds.), *Handbook of discourse processes* (pp. 123–164). Mahwah, NJ, USA: Erlbaum.

Shoham, Y., & Tennenholtz, M. (1997). On the emergence of social conventions: Modeling, analysis, and simulations. *Artificial Intelligence*, *94*, 139–166.

Steels, L. (1996). Self-organizaing vocabularies. In C. G. Langton & T. Shimohara (Eds.), *Artificial life V: Proceedings of the Fifth International Workshop*. Cambridge, MA: MIT Press.

Steels, L. (1998). Synthesizing the origins of language and meaning using coevolution, self-organization and level formation. In J. R. Hurford, M. Studdert-Kennedy, & C. Knight (Eds.), *Approaches to the evolution of language: Social and cognitive bases* (pp. 384–404). Cambridge: Cambridge University Press.

Steels, L. (2002a). Crucial factors in the origins of word-meaning. In A. Wray (Ed.), *The transition to language*. New York: Oxford University Press.

Steels, L. (2002b). Grounding symbols through evolutionary language games. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 211–226). London: Springer.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.

Young, H. P. (1993). The evolution of conventions. *Econometrica*, *61*, 57–84.

Young, H. P. (1998). Individual learning and social rationality. *European Economic Review*, *42*, 651–663.