Language Evolution, Coalescent Processes and the Consensus problem in a Social Network

Hariharan Narayanan * **and Partha Niyogi** †

*Laboratory for Information and Decision Systems
Massachusetts Institute of Technology, and †Department of Computer Science
The University of Chicago

Submitted to Proceedings of the National Academy of Sciences of the United States of America

In recent times, there has been an increased interest in theories of language evolution that have an applicability to the study of dialect formation, linguistic change, creolization, the origin of language, and animal and robot communication systems in general. One particular question that has attracted some interest has the following general form: how might a group of linguistic agents arrive at a shared communication system purely through local patterns of interaction and without any global agency enforcing uniformity? In this paper, we consider a natural model of language evolution, prove several theoretical properties, and establish connections to related phenomena in biology, social sciences, and physics.

language evolution | social network | group consensus

In recent times, there has been an increased interest in theories of language evolution that have an applicability to the study of dialect formation, linguistic change, creolization, the origin of language, and animal and robot communication systems in general. (see [14, 22, 11] and references therein). One particular question that has attracted some interest has the following general form: *how might a group of linguistic agents arrive at a shared communication system purely through local patterns of interaction and without any global agency enforcing uniformity?* The linguistic agents in question might be humans, animals, or machines in a multi-agent society. For an example of interesting simulations that suggest how a shared vocabulary might emerge in a population , see Liberman (2005) (other simulations are also provided by [1, 2, 27, 8, 28] among others). In this paper, we consider a generalization of Liberman's model, prove several theoretical properties, and establish connections to related phenomena in biology, social sciences, and physics.

Our model is as follows. For simplicity, we consider how a common word for a particular concept might emerge through local interactions even though the agents had different initial beliefs about the word for this concept. For example agents might use the phonological forms "dog","kukur", "farama" etc. to describe the concept of a canine animal. Thus we imagine a situation where every time an event in the world occurs that requires the agents to use a word to describe this event, they may start out by using different words based on their initial belief about the word for this event or object. By observing the linguistic behavior of their neighbors agents might update their beliefs. The question is - will they eventually arrive at a common word and if so how fast.

Model.

1. Let W be a set of words (phonological forms, codes, signals, etc.) that may be used to denote a certain concept (meaning or message).
2. Let each agent hold a belief that is a probability measure on W. At time t, we denote the belief of agent i to be $\mathrm{b}_i^{(t)}$.
3. Agents are on a communication network which we model as a directed weighted directed graph where vertices correspond to agents. We further assume that the weight of each directed edge is positive and that there exists a directed path from any node to any other. An agent (say i) can only observe the linguistic actions of its out-neighbors, i. e.nodes to which a directed edge points from i. We denote weight of the edge from i to j by A_{ij}.
4. The update protocol for the $\mathrm{b}_i^{(t)}$ as a function of time is as follows:
   (a) At each time t, each agent i chooses a word $w = w_i^{(t)} \in$ W (randomly from to its current belief $\mathrm{b}_i^{(t)}$) and produces it. Let $X_i^{(t)}$, denote the probability measure concentrated at $w_i^{(t)}$. Since $w_i^{(t)}$ is a random word, $X_i^{(t)}$ is correspondingly a random measure.
   (b) At every point in time, each agent can observe the words that their neighbors produce but they have no access to the private beliefs of these same neighbors.
   (c) Let P be the matrix whose ij^{th} entry satisfies

$$P_{ij} = \frac{A_{ij}}{\sum_{k=1}^{n} A_{ij}}.$$

   At every time step, every agent updates its belief by a weighted combination of its current belief and the words it has just heard, i.e.,

$$\mathrm{b}_i^{(t+1)} = (1-\alpha)\mathrm{b}_i^{(t)} + \alpha \sum_{j=1}^{n} P_{ij} X_j^{(t)},$$

   where α is a fixed real number in the interval $(0, 1)$.

At a time t, let the beliefs of the agents be represented by a vector

$$\mathrm{b}^{(t)} := (\mathrm{b}_1^{(t)}, \ldots, \mathrm{b}_n^{(t)})^T.$$

Similarly, let the point measures on words $X_i^{(t)}$ be organized into a vector

$$X^{(t)} := (X_1^{(t)}, \ldots, X_n^{(t)})^T.$$

Then the reassignment of beliefs can be expressed succinctly in matrix form where the entries in the vectors involved are measures rather than numbers as

$$\mathrm{b}^{(t+1)} = (1-\alpha)\mathrm{b}^{(t)} + \alpha P X^{(t)}. \qquad [1]$$

Reserved for Publication Footnotes

Remarks:.

1. If beliefs were directly observable and agents updated based on a weighted combination of their beliefs and that of their neighbors,

$$\mathrm{b}^{(t+1)} = (1-\alpha)\mathrm{b}^{(t)} + \alpha P \mathrm{b}^{(t)}, \qquad [2]$$

   the system has a simple linear dynamics, where all beliefs converge to a weighted average of the initial beliefs. Thus eventually, everyone has the same belief (see [4] for pioneering work and [12] for a recent elaboration in an economic context.)
2. Our focus in this paper is on the situation where the beliefs are *not observable* but only the linguistic actions $X_i^{(t)}$ are (and only to the immediate neighbors). Therefore, the corresponding dynamics follows a Markov chain. The state space of this chain (defined by Equation 1) is the set of all n-tuples of belief vectors. Since this is continuous, the standard mixing results with finite state spaces do not apply directly.
3. Note that in our setting we have assumed that the communication matrix A_{ij} does not change with time. If this matrix changes with time the evolution is not Markovian in the usual sense but the arguments in this paper when combined with results in [5] would lead to a proof of convergence under suitable conditions. We omit this analysis for ease of exposition.

Results:. Our main results are summarized below.

1. With probability 1 (w.p.1), as time tends to infinity, the belief of each agent converges in variation distance to one supported on a single word, common to all agents.
2. w.p.1, there is a finite time T such that for all times $t > T$, all agents produce the same fixed word.
3. The rate at which beliefs converge depends upon the mixing properties of the Markov chain whose transition matrix is P.
4. The rate of convergence is *independent* of the size of W. One might think that a population where every agent has one of two words for the concept would arrive at a shared word faster than one in which every agent had a different word for the concept. This intuition turns out to be incorrect.
5. The proof of these results exposes a natural connection with coalescent processes and has a parallel in population genetics.
6. Our analysis brings out two different interpretations of the behavior of a linguistic agent. In the most direct interpretation, the agent's linguistic knowledge of the word is internally encoded in terms of a belief vector. This belief vector is updated with experience. In a second interpretation an agent's representation of its linguistic knowledge is in terms of a memory stack in which it literally stores every single word it has heard weighted by how long ago it heard it and the importance of the person it heard it from. Such an interpretation is consistent with exemplar theory (see [9]). An external observer looking at this agent's linguistic actions will not be able to distinguish between these two different internal representations that the agent may have.

Connections to other fields. The general theme of predicting the macroscopic behavior of a system from the local behavior of its microscopic components arises in many different areas of physics, biology, and the social sciences. It is also a fundamental issue in the analysis of distributed systems in computer science.

In Spin systems, which originated as models for Ferromagnets, atoms are pictured to be in a 2-Dimensional square array, each possessing a spin "up" or "down." The effect that an atom has on the spin of a neighbor is a function of temperature. Typically, coherence is observed at low temperatures, while at high temperatures atoms tend not to align, which is in agreement with the demagnetization that ferromagnets undergo at high temperatures. The model we consider, involving the convergence in beliefs has many high level similarities though we do not address the question of what might be the analog of temperature in our model, how to take the thermodynamic limit, and if and how phase transitions may arise.

Another closely related model is the voter model studied in probability theory with its origins in the social sciences. Each agent lives on the vertex of the graph, has a belief which is a discrete variable, and is observable to its neighbors. Each agent changes its belief with a certain probability based on the observed beliefs of its neighbors. Another kind of belief propagation model is that described by Jackson (2007). In both cases, the beliefs are observable in contrast to our setting. Our communication graphs model the pattern of local interaction among agents and may arise through modes of social network formation studied in the field of social network theory [26, 15].

Linear update rules are often used in distributed systems, to achieve coherence among different agents or to share knowledge gathered individually. In a model that has been intensively studied, a number of sensors form a network, each of which measures a quantity such as temperature [4]. Neighbors communicate during each time step and make linear updates in a synchronous or asynchronous manner. The rate at which consensus is attained is studied. There is also a related body of work on Coordination and Distributed Control. A model of flocking has been considered in [6], where a group of birds, have a certain initial velocity, and the evolution of their velocities is governed by a differential equation wherein each bird modifies its velocity to bring it closer to that of its neighbors. The update rule involves a graph Laplacian. Some results are derived concerning the initial conditions that result in flocking behavior.

There are two connections to evolutionary theory that are worth mentioning. First, our proof of convergence exposes a natural coalescent process over words. Coalescent processes are, of course, widely used in modeling and making inferences about genetic evolution [16, 17]. Second, researchers have considered game-theoretic models of evolution [20] and more recent research in this tradition has addressed evolutionary

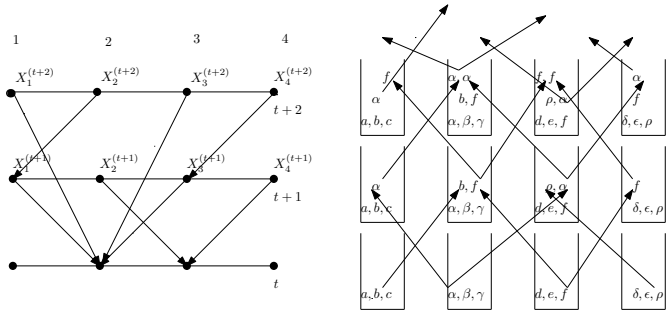


Fig. 1. A coalescent process obtained by tracing the origin of words backwards in time, and the associated memory stacks of agents 1 to 4 for time steps t to $t+2$. Each agent produces α at time $t+2$ due to coalescence to a single word α produced by agent 2 at time t.

games on graphs [24]. The question of how agents may learn an appropriate strategy for a coordination game on a graph has many high level similarities to the problem studied in this paper.

Finally, there have been a large number of models on achieving coherence in a linguistic population. Many of these rely on simulations. Among mathematical studies, two strands are worth noting. The model of language evolution proposed in [7] has many similarities with languages of agents evolving on a graph. But it is worth noting that in that model, if at each time step, the number of linguistic examples (observations) collected by each agent is bounded from above by a constant (independent of time), the community fails to achieve a consensus language. A second strand is the collection of results obtained in [23, 14]. While there are many synergies with that body of work, there is nothing that is directly comparable.

Convergence to a Shared Belief: Quantitative results

Let $\tilde{P}$ be the transition matrix on the state space $\tilde{S} = S \cup \hat{S}$, where for $i, j \in S := \{1, \ldots, n\}$ and $\hat{S} = \{\hat{1}, \ldots, \hat{n}\}$.

$$\tilde{P}(i \to j) = \tilde{P}(\hat{i} \to j) = \alpha P_{ij},$$

$$\tilde{P}(i \to \hat{i}) = \tilde{P}(\hat{i} \to \hat{i}) = 1 - \alpha.$$

Definition 1. *Let $T_{mix}(\epsilon)$ denote the mixing time of $\tilde{P}$, defined as the smallest t for which, for each specific choice of $v, w \in \tilde{S}$,*

$$\sum_{u \in \tilde{S}} |\tilde{P}^{(t)}(v \to u) - \tilde{P}^{(t)}(w \to u)| < \epsilon.$$

Here $\tilde{P}^{(t)}(b \to c)$ denotes the probability that a Markov Chain governed by $\tilde{P}$ starting in b lands in c at the t^{th} time step.

The following is the main result of this paper.

Theorem 1:

1. The probability that all agents produce the same word at times $T, T+1, \ldots$ tends to 1 as T tends to ∞. More precisely, if

$$\begin{aligned} \tau &= (4n/\alpha^2) T_{mix}(\frac{\alpha}{4}) \ln(4n/\alpha^2) \\ M &= e, \end{aligned}$$

then

$$\mathbb{P}[\forall_{\substack{t \geq T \\ u \in S}} X_u^t = X_1^T] > 1 - \frac{MnTe^{-\frac{T}{\tau}}}{1 - e^{-\frac{T}{\tau}}}. \qquad [3]$$

2. As time $t \to \infty$ all produced words converge (almost surely) to a word whose probability distribution is

$$\sum_{i=1}^{n} \pi_i \mathrm{b}_i^{(0)},$$

where $(\pi_1, \ldots, \pi_n)$ is the stationary distribution of the Markov chain whose transition matrix is P.

A Model of Memory. The evolution of the $B^{(t)}$ is a Markov chain. It can be seen that its only absorbing states are of the form $(\mathrm{b}_1^{(t)}, \ldots, \mathrm{b}_n^{(t)})^T$, where $\forall i, \mathrm{b}_i^{(t)} = \delta_w$, and δ_w is the point measure concentrated on some word $w \in \mathrm{W}$. Formally, δ_w is the measure on W, which assigns to a measurable set A the measure $\delta_w(A)$ according to the following rule.

$$\begin{aligned} \delta_w(A) &= 1 \quad \text{If } w \in A \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Therefore, if the Markov Chain were finite, a simple argument would suffice. In our case however, we have a Markov Chain whose state space is uncountably infinite. Thus in principle, its dynamics could be hard to analyze. Our proof is based on coalescent processes, which have also been extensively used to study biological evolution [16, 17]. In analyzing the evolution of beliefs, we trace the origin of words backwards in time and find that all surviving words, are copies of a single word produced at some point in time sufficiently far in the past. Observe that if the process had begun at time 0, the beliefs at time $t+1$ would be

Observation 1.

$$B^{(t+1)} = \sum_{i=0}^{t} \alpha(1-\alpha)^i P X^{(t-i)} + (1-\alpha)^{t+1} B^{(0)}. \qquad [4]$$

$X^{(t)} = (X_1^{(t)}, \ldots, X_n^{(t)})^T$ is a random vector whose entries are point measures, where $X_i^{(t)} = \delta(w_i^{(t)})$ and $w_i^{(t)}$ is chosen from the measure $\mathrm{b}_i^{(t)}$ on W, independent of the choice of other coordinates of the vector $X^{(t)}$. This observation, motivates a model of memory that we define in the following paragraph.

Let each agent's memory be modeled as a stack. At the top level of the stack of agent i are all the words heard at time t. Below this are all words heard at time $t-1$ and so on tracing backwards in time until the first words heard at an initial time 1. At the lowest level, corresponding to time 0, is the initial belief $\mathrm{b}_i^{(0)}$ which is a probability distribution on the set of words. We may imagine this to be a form of vestigial memory.

Let agent j be adjacent to agent i. We shall describe the process by which agent j produces word $w_j(t)$ and whereby also generates or produces $X_j(t)$ which is the point measure supported on $X_j(t)$. Let S_j be the stack held by agent j, and $S_j^{(t)}, \ldots, S_j^{(0)}$ be the levels in its stack from top to bottom. After j produces $X_j(t)$, i places $X_j(t)$, and all other $X_{j'}(t)$ produced by neighbors of i at time step t on the top of its stack. In order to describe the mechanism by which $X_j(t)$ is generated, let us introduce a geometric random variable Y where

$$\mathbb{P}[Y = i] = \alpha(1-\alpha)^i.$$

If $Y \leq t-1$, $X_j(t)$ is chosen to be the word produced by j' at time $t-1-Y$ (which is stored in S_{t-1-Y}) with probability $P_{jj'}$. If $Y \geq t$, $X_j(t)$ is chosen from the distribution in $\mathrm{b}_j^{(0)}$. This process has been illustrated in Figure . Note that in this model words are formal objects. While any two words present in the stack positions $S_j^{(t)}$ for $t = 1, 2, \ldots$ are considered distinct, there is a natural "parent-child" structure existing on the set of words. Under this scheme, let the probability distribution of $X_i^{(t)}$ be denoted $\tilde{\mathrm{b}}_i^{(t)}$. Denoting by $\tilde{B}^{(t)}$ the vector $(\tilde{\mathrm{b}}_1^{(t)}, \mathrm{b}_2^{(t)}, \ldots, \mathrm{b}_n^{(t)})$.

Observation 2. *A direct computation shows that in the model just described*

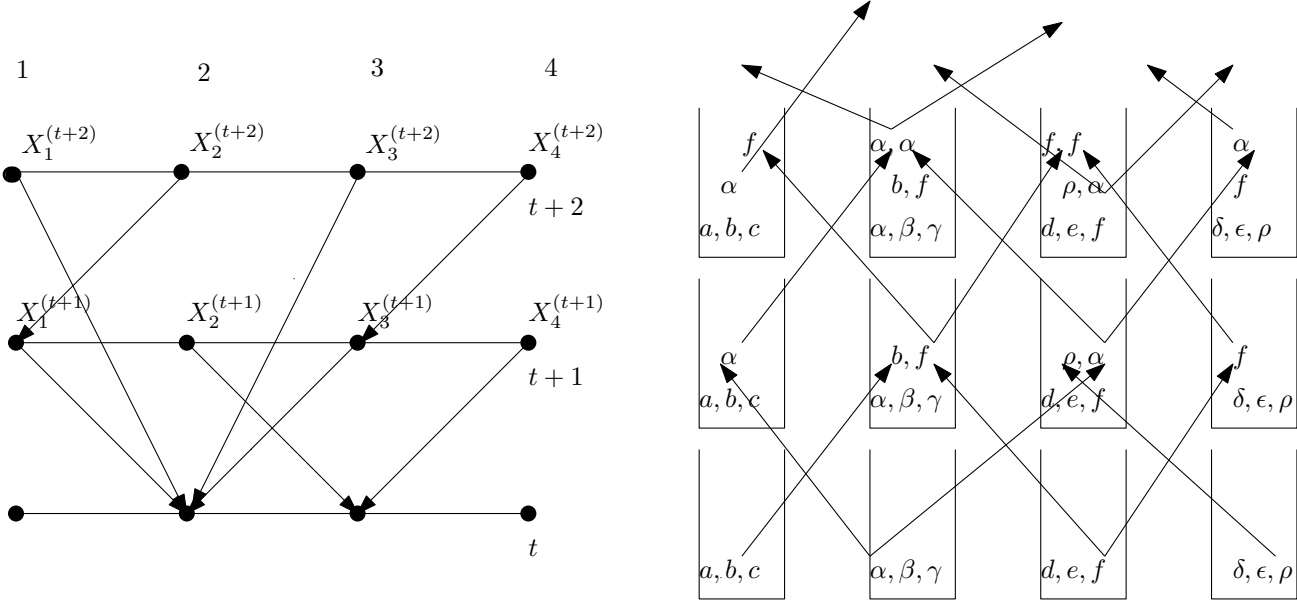
$$\tilde{B}^{(t+1)} = \sum_{i=0}^{t} \alpha(1-\alpha)^i P X^{(t-i)} + (1-\alpha)^{t+1} \tilde{B}^{(0)}. \qquad [5]$$

This along with the fact that the randomness used in the generation of $X_j^{(t)}$ is independent of the randomness in the

generation of all other words, tells us that the model of memory just described results is a system with the same dynamics as that introduced earlier. This particular model of memory may be viewed as an implementation of the ideas implicit in exemplar based accounts of linguistic behavior.

ACKNOWLEDGMENTS. This project was completed before the tragic death of Partha Niyogi. With his passing, the first author lost an inspiring guide and a wonderful friend.

1. A. Baronchelli, V. Loreto and L. Steels. In-depth analysis of the Naming Game dynamics: the homogeneous mixing case, Int. J. of Mod. Phys. C (in press, 2008).
2. A. Baronchelli, M. Felici, E. Caglioti, V. Loreto and L. Steels. Sharp transition towards shared lexicon in multi-agent systems, J. Stat. Mech. P06014 (2006).
3. R. R. Bush and F. Mosteller, Stochastic Models for Learning, Wiley, 1955.
4. D. Bertsekas and J. Tsitsiklis, Parallel and Distributed Computation: Numerical Methods, Prentice-Hall, 1989; republished in 1997 by Athena Scientific.
5. J. N. Tsitsiklis, D. P. Bertsekas and M. Athans, Distributed Asynchronous Deterministic and Stochastic Gradient Optimization Algorithms, IEEE Transactions on Automatic Control, Vol. 31, No. 9, 1986, pp. 803-812.
6. F. Cucker, S. Smale, Emergent behavior in flocks, IEEE transactions on Automatic Control, May 2007.
7. F. Cucker, S. Smale, D. X. Zhou, Modelling Language evolution, preprint
8. B. de Boer, Evolution of Speech and its Acquisition, Adaptive Behavior 13(4) 281292
9. J. Bybee, Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. Language Variation and Change 14. 261-290.
10. P. Diaconis and D. Stroock, Geometric bounds for eigenvalues of Markov Chains, Annals of Applied Probability, 1, (1991).
11. M.D. Hauser (1996), The evolution of communication, MIT Press, Cambridge, MA.
12. M. Jackson, The Economics of Social Networks, Chapter 1 in Volume I of Advances in Economics and Econometrics, Theory and Applications, 2006.
13. N. Moshtagh and A. Jadbabaie, Distributed geodesic control laws for flocking of nonholonomic agents. IEEE Transactions on Automatic Control, to appear, April 2007.
14. S. Kirby, Function Selection and Innateness: the Emergence of Language Universals. Oxford University Press. Innateness and culture in the evolution of language.
15. J. Kleinberg, Navigation in a Small World. Nature 406 (2000), 845.
16. R. R. Hudson, Gene Genealogies and the Coalescent Process. Oxford Surveys in Evolutionary Biology 1991, 7: 1-44
17. Kingman, J.F.C. On the Genealogy of Large Populations. Journal of Applied Probability 1982, 19A:27-43
18. Mark Liberman, The "Lexical Contract": Modeling the emergence of word pronunciations, http://www.ldc.upenn.edu/myl/abm/index.html
19. Lieberman, E., Hauert, C., and Nowak, M, Evolutionary Dynamics on Graphs, Nature, 433, 2005.
20. J. M. Smith, Evolution and the Theory of Games, Cambridge University Press 1982
21. C. Yang, Knowledge and Learning in Natural Language, Oxford University Press, 2002.
22. P. Niyogi, The Computational Nature of Language Learning and Evolution. MIT Press, April 2006.
23. M. A. Nowak, and N.L. Komarova (2001), Towards an evolutionary theory of language, Trends in Cognitive Sciences 5 (7), pp. 288295.
24. M. A. Nowak, and K. Sigmund, Evolutionary Dynamics of Biological Games, Science 6 February 2004: Vol. 303. no. 5659, pp. 793 - 799
25. Ohtsuki, H., Hauert, C., Lieberman, E., and Nowak, M, A Simple Rule for the Evolution of Cooperation on Graphs, Nature, 441, 2006.
26. D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks, Nature, 393:440-442 (1998)
27. L. Steels. The puzzle of language evolution. Kognitionswissenschaft, 8(4), 1999.
28. Steels, L., and McIntyre, A. Spatially distributed naming games, In ECAL 97.


1
2
3
4
$X_1^{(t+2)}$
$X_2^{(t+2)}$
$X_3^{(t+2)}$
$X_4^{(t+2)}$
$t+2$
$X_1^{(t+1)}$
$X_2^{(t+1)}$
$X_3^{(t+1)}$
$X_4^{(t+1)}$
$t+1$
t
f
α
a, b, c
α, α
b, f
α, β, γ
f, f
ρ, α
d, e, f
α
f
δ, ϵ, ρ