Reali, F., Chater, N. & Christiansen, M.H. (2014). The paradox of linguistic complexity and community size. In E.A. Cartmill, S. Roberts, H. Lyn & H. Cornish (Eds.), The evolution of language: Proceedings of the 10th International Conference (pp. 270-277). Singapore: World Scientific.

THE PARADOX OF LINGUISTIC COMPLEXITY AND COMMUNITY SIZE

FLORENCIA REALI

Psychology Department, Universidad de los Andes, Bogotá, DC 11001000, Colombia

NICK CHATER

Behavioural Science Group, Warwick Business School, University of Warwick, Coventry, CV4 7AL, UK

MORTEN H. CHRISTIANSEN

Department of Psychology, Cornell University, Ithaca, NY 14853, USA
Santa Fe Institute, Santa Fe, NM 87501, USA

It has been observed that languages with huge numbers of speakers tend to be structurally simple while small communities can sometimes develop languages with great structural complexity. Paradoxically, an apparent opposite pattern appears to be observed in relation to non-structural properties of language such as number of content words. These apparent contradictory patterns pose a challenge for cultural evolution approaches to language evolution. In this paper, we use computational simulations to investigate the hypothesis that the opposite effects of linguistic community size on linguistic structure and vocabulary depend on a single factor: ease of learning. We created a population of simulated agents arranged on a network, such that agents connected by a link on the network are able to communicate and potentially pass linguistic conventions to one another. Each agent can both invent entirely new conventions and replicate conventions that they have previously generated themselves or learned from other agents. Linguistic conventions are divided into two categories *Easy* and *Hard* to learn, depending on how many times an agent needs to hear a convention in order to learn it. The simulation results show that when the population is small, Hard conventions represent a sizable proportion of the total linguistic inventory. As population size increases the number of easy-to-learn properties increases whereas the frequency of those that are hard to learn decreases systematically. The results suggest that the size of a linguistic community can potentially have opposite effects on the richness of different aspects of the language as a function of the ease of learning of different language properties.

1. Introduction

It has often been observed (e.g., Lupyan & Dale, 2010; Trudgill, 2011; Wray & Grace, 2007) that the properties of human languages appear to be influenced, in part, by the size and degree of isolation of the linguistic community. Thus, small, isolated linguistic communities oftentimes develop languages with great structural complexity, elaborate and opaque morphology, rich patterns of agreement, and many irregularities (Lupyan & Dale, 2010; Haspelmath et al. 2008; Trudgill, 2011; Wray & Grace, 2007). By contrast, languages with large communities of speakers, such as Mandarin or English, appear to be structurally simpler. The causal role of the size of the linguistic community is, moreover, further indicated by the historical tendency toward structural simplification as languages gain an ever-larger community of speakers (MacWhorter, 2002).

But an apparently opposite pattern appears to be observed in relation to non-structural properties language: languages with large linguistic communities tend to have larger vocabularies of content words. For example, the vocabulary of wide-spread languages, such as English, appears to have grown rapidly in historical times, and are typically estimated to have many hundreds of thousands of words, including those with highly specialized and technical meanings. Tens of thousands of words are known by typical English speakers (Goulden, Nation & Read, 1990). Despite their frequent spectacular structural complexity, languages spoken by small bands of hunter-gatherers are typically assumed to have smaller vocabularies, although reliable data for such languages are difficult to gather (cf. Pawley, 2006).

These apparently contradictory patterns pose a challenge for theories based on the cultural evolution of language. Various theorists have convincingly suggested mechanisms for the erosion of complexity in larger language communities (e.g., Dale & Lupyan, 2012). But why do such arguments for simplification not also apply to the lexicon? One possibility is that structural and lexical aspects of language are learned by very different mechanisms, so that the gradual modification of these two aspects of language, from generation to generation, is governed by distinct principles. For example, adult-child interactions might be the primary vehicle for grammatical regularization; and adult-adult interactions might be the primary vehicle for lexical innovations. Moreover, there may be differential impacts of language contact on structural and lexical aspects of language.

While not denying that such factors may play a role, we focus here on a more parsimonious alternative: that the very same learning mechanisms can yield opposite relationships between population size and lexical vs. structural complexity depending on a single parameter: ease of learning. The crucial difference between structural and lexical aspects of language, we suggest, is that structural aspects of language are difficult to learn and require many instances of the relevant structural feature to be encountered before learning is possible. Words, by contrast, can be learned rapidly (indeed, during the vocabulary spurt, children may learn as many as 10 words each day), and require few exposures.

To illustrate this scenario, we divide properties of language, as a first approximation, into two basic categories—*Easy* and *Hard*—requiring, respectively, few or many exposures to be acquired by a new speaker. Easy properties of the language can rapidly be transmitted across the linguistic community. As a linguistic community grows in size, so does the number of members who can spontaneously modify or invent new Easy properties (such as lexical items) that may subsequently spread across the community. Hence, large communities will end up with large inventories of easy features. By contrast, Hard properties of the language require many exposures to learn, so that propagating such properties across the population is more difficult. A modification or innovation of a Hard property by a particular speaker will not spread easily across large linguistic communities, where speakers tend only to have minimal interactions with a great number of speakers, rather than repeated interactions with a small number speakers. Thus, in large linguistic communities, typical interactions between individual speakers will be too limited to transmit the Hard linguistic property successfully.

Can these intuitions be made precise by computer simulation? To test this, we created a population of simulated agents. Agents are arranged on a network, so that agents connected by a link on the network are able to "converse" and hence, potentially pass linguistic "conventions" to one other. Each agent is not only able to "invent" entirely new conventions but can also replicate conventions that they have previously generated themselves or learned from other agents (i.e., agents to which they are connected by links in the network). When an agent produces a convention (whether novel or a replication), it propagates that convention to one of its neighbors.

To capture the dynamics of individuals interacting with one another, either conversing by way of old conventions or inventing new ones, we use a modified version of the Chinese restaurant process (Jordan, 2005). This is a widely used probabilistic model defining the frequency distribution over a potentially limitless number of types (e.g., linguistic conventions, words, categories). It embodies the assumption that the "rich-get-richer"—the probability of a token of an existing type is proportional its current frequency (i.e., the chance of the new diner sitting at a table is proportional to the number of diners already at that table), while also allowing the creation of new types (i.e., a diner being seated at previously unoccupied table).

In the current framework, we view each agent as corresponding to a "restaurant" with a finite, but infinitely extendable, number of "tables," i.e., conventions. Each time the agent generates a convention, it chooses an existing convention with a probability proportional to the number of previous tokens of that convention; this is equivalent to seating each new customer in the restaurant at a table in proportion to the number of customers already seated at that table. But it is also possible that an entirely novel convention will be generated (a new table in the restaurant is created, and the new customer becomes the first person sitting at that table). This occurs with probability 1/(M+1) (where M is the number of current restaurant customers).

However, following this scheme precisely is, of course, not appropriate for the present task, as each agent will be generating conventions (i.e., properties of the language) entirely independently, and not sharing those conventions with the rest of the linguistic community. A simple extension of the Chinese restaurant process can deal with this: for each agent, the probability of generating an existing convention is determined by the sum of the number of times that it has, itself, previously generated that convention, added to the sum of the number of times that it has received that convention from a neighboring agent. Thus, in this model, agents tend not merely to generate what they have generated before; but also to generate what they have "heard" (and learned) from neighboring agents.

As the simulation progresses, agents will invent conventions, and pass them on to each other. Thus, initially the number of conventions used by the agents (i.e., the complexity of the language) will gradually increase. It will not increase indefinitely, however, as we introduce a rule for eliminating "unused" conventions. We assume that agents have a limited capacity to store convention-tokens. Within the Chinese restaurant process, this means that the restaurant for generating language has limited seating. When the limit is reached, the agent starts forgetting. That is, after a threshold number of conventions have been seated (or waiting to be seated) then whenever a new convention comes in, an existing one must leave. This will occasionally leave a table completely empty, and if so, that table is deleted and the agent loses the corresponding convention.

So far, we have not distinguished between Easy conventions (which can be learned from another agent by minimal exposure—these correspond to lexical items) and Hard conventions (which require multiple exposures—these correspond to structural properties of the language). As a first approximation, we make the simplest of distinctions between them: Easy conventions can be learned by an agent from a *single* exposure. Once a convention has been generated by a neighbor, an agent can immediately generate that convention. Hard conventions can only be learned from *two* exposures: only when an agent has encountered two examples of the exact same convention from its neighbors (whether the from same or different neighbor), will this convention be seated at a new table (representing that convention in the agent).

2. Simulations

Agents are represented as nodes in a random graph. Specifically, we use Gilbert random graphs (Bollobás, 2001), *G(n,p)*, where n is the number of nodes (i.e., the population size) and p is the probability that a link connects a pair of nodes (i.e., agents), making them neighbors. As n increases, so does the number of neighbors that an agent has on average (even for a fixed value of p). In the current simulations, we used a fixed value of $p = .5$. However, the value of n was systematically varied to explore the effect of population size.

On a given iteration, each agent “utters” one convention to one of its neighbors, who is randomly picked from the set of all its neighbors in the graph. The convention produced by the agent can be either part of its repertoire (conventions that have been previously generated or learned by the agent) or invented anew. Conventions are divided into two types: Easy and Hard to learn conventions. Each time an agent “invents” a new convention, that convention is randomly defined to belong to one of these two categories with probability 0.5.

We use an extension of the Chinese restaurant stochastic sampling process to model an agent’s selection of a convention to generate. The probability of choosing a given convention, c, is proportional to the number of c tokens that it has previously generated or heard from its neighbors. More precisely, the probability of selecting an already used convention is defined as,

$$P(convention = c) = \mathrm{t_c}/(M + 1) \tag{1}$$

where t_c is the number of tokens of convention c that are part of the agent´s repertoire and M is the number of convention tokens that the agent has stored in memory, thus $\sum t_c = M$. The probability of inventing a convention anew is defined as,

$$P(convention = anew) = 1/(M + 1) \tag{2}$$

The value of M increases over subsequent iterations until it reaches the maximum number of tokens, M_{max}, that an agent can store in memory. M_{max} is therefore a variable parameter in our current framework, capturing the idea that cognitive constraints affect the cultural evolution of language (Christiansen & Chater, 2008). When the M_{max} limit is reached, the agent starts forgetting convention-tokens: whenever a new token is generated or heard from a neighbor, an existing token is deleted at random from the agent´s memory.

Agents can learn conventions from neighbors. The *learned* convention becomes part of the agent’s repertoire and can be sampled during its own production. In the current simulations, Easy conventions are defined as those that are learned from only a single exposure, whereas Hard conventions require at least two exposures to be learned.

We are interested in determining the number of Easy and Hard conventions that are actively used at the population level. Thus, a convention is considered “active” when it has been learned or generated by (at least) a minimum number of agents in the population at some point across iterations. We call this value the *active-convention criterion*.

2.1. *Implementation and results*

A single run of our simulation is composed of many iterations. During each iteration, communication involves letting each agent "utter" one convention to another agent randomly selected from its pool of neighbors. When a convention is generated anew there is a 50-50 chance that it will be assigned to the Hard or Easy categories. Five separate runs of 1000 iterations were carried out across a range of the parameters *n* (population size) and M_{max} (maximum number of convention-tokens that an agent can store in memory) and the *active-convention criterion* (the minimum number of agents that must have generated or learned a convention for it to be considered "active"). At the end of each run, the number of *active conventions* that remained part of agents´ memory (tables in the restaurant) was counted. We computed both the absolute and relative number of active conventions in each of the two categories, Easy and Hard, as population size increases. All simulations were implemented using R (R Development Core Team, 2008).

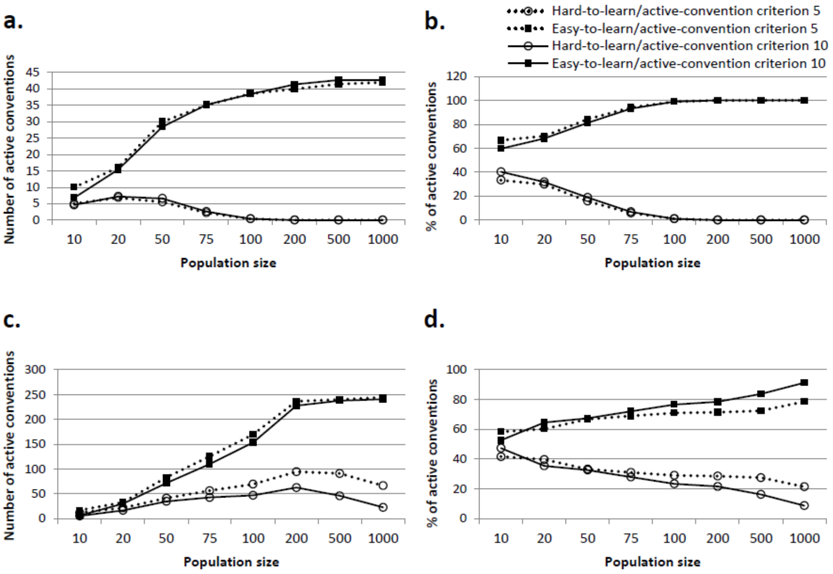

Figure 1. Absolute number (left panels) and relative proportion (right panels) of active conventions after 1000 iterations, obtained for increasing values of population sizes (displayed in the *x-axis*). Top panels (**a**. and **b**.) display the results corresponding to a value M_{max}=100 (the maximum number of convention-tokens that an agent can store in memory), while bottom panels (**c**. and **d.**), display the results corresponding to a value M_{max}=500. Filled squares correspond to Easy-to-learn conventions, and open circles correspond to Hard-to-learn conventions. Results displayed in dashed lines correspond to active-convention criterion = 5 agents, and solid lines to active-convention criterion = 10 agents.

Results are shown in Fig. 1, reflecting a general trend towards an increasing frequency of Easy conventions compared to Hard conventions as the population size increases. When the population is small, Hard conventions represent a sizable proportion of the total number of conventions. As population size increases and the overall number of active conventions grows, the absolute and relative number of Hard conventions decreases. Even though the number of conventions vary for different combinations of parameters, both the absolute and relative patterns remain the same across the different conditions, suggesting a robust effect of population size on the proportion of Hard vs. Easy to learn conventions.

3. Conclusion

We have shown that the size of a linguistic community can potentially have opposite effects on the richness of different aspects of the language. Linguistic innovations that are relatively easy to learn (such as new lexical items or modifications to existing ones) will increase in number as a linguistic community grows, because the number of potential innovators increases, and innovations can spread more rapidly. By contrast, small linguistic communities favor linguistic innovations that are hard to learn (such as, we suggest, structural changes in the language) because they require multiple interactions between individual speakers (or the innovation will not be transmitted successfully). In small communities, agents tend to have repeated interactions with a small number of speakers because they have fewer neighbors on average. Thus, the differential effects of population size on structural complexity and vocabulary size can be accommodated within a cultural evolution approach where the evolution of language is shaped by cultural transmission constrained by cognitive and interactional constraints (Christiansen & Chater, 2008)

It is likely, of course, that many additional forces have shaped the relative development of different aspects of linguistic complexity (Trudgill, 2011). One factor that may partly underlie the Easy/Hard distinction considered here concerns the degree to which properties of language can be learned independently. Perhaps an additional reason that learning a lexical item is relatively easy is that word meanings can, to a considerable degree, be learned independently of one another. By contrast, structural aspects of language may interlock in more complex ways, making the propagation of such linguistic innovations more difficult.

More broadly, it is interesting to speculate whether other aspect of linguistic and cultural evolution may be subject to the pressures described here. For example, perhaps an increase in community size might be associated with a reduction in the prevalence of complex dances, music, skills, rituals, or beliefs, but an increase in the prevalence of simpler variants. Of course, such effects may, to some extent, be counteracted by the ability of people to self-assemble into small specialist groups, to innovative and propagate cultural forms of high complexity. In the absence of the ability for people to self-organize in this way, our simulations raise the possibility that language and culture might become unrelentingly simpler, at the structural level, as human societies become increasingly interconnected.

References

Bollobás, B. (2001). *Random graphs* (2nd ed.). Cambridge: Cambridge University Press.

Christiansen, M. H. & Chater, N. (2008). Language as shaped by the brain. *Behavioral & Brain Sciences, 31*, 489-558.

Dale, R. & Lupyan, G. (2012). Understanding the origins of morphological diversity: the linguistic niche hypothesis. *Advances in Complex Systems, 15, 3 & 4*, 1150017.

Goulden, R., Nation, P. & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, *11*, 341-363.

Haspelmath, M., Dryer, M., Gil, D. & Comrie, B. (2008). *The world atlas of language structures online*. Munich: Max Planck Digital Library.

Jordan, M.I. (2005). Dirichlet processes, Chinese restaurant processes and all that. *Tutorial presentation at the NIPS Conference*. Vancouver, British Columbia, Canada.

MacWhorter, J. (2002). What happened to English? *Diachronica, 19,* 217-272.

Lupyan, G. & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS One, 5, e8559.*

Pawley, A. (2006). On the size of the lexicon in preliterate language communities: Comparing dictionaries of Australian, Austronesian and Papuan languages. In J. Genzor & M. Buckov (Eds.), *Favete Linguis: Studies in honour of Viktor Krupa* (pp. 171-191). Bratislava: Institute of Oriental Studies.

R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Trudgill, P. (2011). *Sociolinguistic typology: Social determinants of linguistic structure and complexity*. Oxford: Oxford University Press.

Wray, A. and Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua 117,* 543-578.