Advances in Complex Systems
Vol. 15, Nos. 3 & 4 (2012) 1150017 (16 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0219525911500172

World Scientific
www.worldscientific.com

UNDERSTANDING THE ORIGINS OF MORPHOLOGICAL DIVERSITY: THE LINGUISTIC NICHE HYPOTHESIS

RICK DALE[*,‡] and GARY LUPYAN[†,§]

[*]*Cognitive and Information Sciences,*
University of California, Merced,
Merced, CA, 95343, USA

[†]*Department of Psychology, University of Wisconsin,*
Madison, Wisconsin, 53706, USA
[‡]*rdale@ucmerced.edu*
[§]*lupyan@wisc.edu*

Received 9 May 2011
Revised 20 September 2011
Accepted 14 October 2011
Published 13 March 2012

Human language is unparalleled in both its expressive capacity and its diversity. What accounts for the enormous diversity of human languages [13]? Recent evidence suggests that the structure of languages may be shaped by the social and demographic environment in which the languages are learned and used. In an analysis of over 2000 languages Lupyan and Dale [25] demonstrated that socio-demographic variables, such as population size, significantly predicted the complexity of inflectional morphology. Languages spoken by smaller populations tend to employ more complex inflectional systems. Languages spoken by larger populations tend to avoid complex morphological paradigms, employing lexical constructions instead. This relationship may exist because of how language learning takes place in these different social contexts [44, 45]. In a smaller population, a tightly-knit social group combined with exclusive or almost exclusive language acquisition by infants permits accumulation of complex inflectional forms. In larger populations, adult language learning and more extensive cross-group interactions produce pressures that lead to morphological simplification. In the current paper, we explore this learning-based hypothesis in two ways. First, we develop an agent-based simulation that serves as a simple existence proof: As adult interaction increases, languages lose inflections. Second, we carry out a correlational study showing that English-speaking adults who had more interaction with non-native speakers as children showed a relative preference for over-regularized (i.e. morphologically simpler) forms. The results of the simulation and experiment lend support to the *linguistic niche hypothesis*: Languages may vary in the ways they do in part due to different social environments in which they are learned and used. In short, languages adapt to the learning constraints and biases of their learners.

Keywords: Language change; social structure; morphology; agent-based simulation.

1. Introduction

Languages differ greatly in their degree of morphological complexity [17]. At the one extreme are languages in which semantic distinctions are made almost exclusively through lexical means. At the other extreme are polysynthetic languages in which a relatively small set of lexical items are combined with a large set of affixes to make semantic distinctions in ways that have been compared to chemical compounds [36]. In the middle lies a continuum of morphological specification. For example, in English the past-tense is generally marked by adding *-ed* to the end of a verb stem (a simple morphological encoding). In comparison, the Peruvian language Yagua uses 5 inflections on verbs to denote levels of temporal remoteness: For example, *-jásiy* is used to mark an event that happened just recently, and another inflection, *-jay*, is used to communicate an event that occurred a week to a month ago [34]. In English a speaker can, of course, simply add more descriptive content through lexical means: I ate the cheese *a few hours ago*. However, morphologically encoded semantic distinctions tend to be (1) obligatory (an English-speaker cannot simply omit a verb-tense), and (2) more idiosyncratic. Although English morphological expressions of tense do contain some irregular forms (*walk/walked*, but *eat/ate*), the simple morphology leaves little room for irregularities and context-sensitive paradigms of the type that occur in more morphologically rich languages like Yagua.

What factors drive languages, over a historical time scale, to employ more or less specified inflectional systems? At least part of the answer may lie in the ways that a given language is learned and used [44, 45]. The grammatical structure of languages adjusts as a function of the communicative needs and learning constraints of the speaking community. This general idea that grammatical structure is constrained by the community learning and speaking the language has been proposed by a number of authors [9, 11, 27, 29, 43–46].

Consider the two languages just described. Yagua is spoken by about six-thousand people, and English by hundreds of millions [16]. Yagua is considerably more morphologically elaborated than English. Though there are exceptions to this pattern, these two languages serve to exemplify the general trend identified in analysis of over 2000 languages: Languages spoken by smaller populations, and over smaller geographic areas, tend to employ more complex inflectional morphology. Languages spoken by larger populations and over larger geographic areas tend to have sparser inflection and tend to employ lexical over inflectional devices [25]. Following the terminology of [46], we term languages spoken in small communities as occupying the *esoteric* niche, and languages spoken in large communities (with correspondingly larger numbers of adult learners) as occupying the *exoteric* niche.

One way to explain the relationship between social and grammatical structure– what we have termed the *linguistic niche hypothesis* — is by appealing to different learning skills of children and adults. Consider that, by definition, unlearnable languages cannot exist [9], and all natural languages are subject to the learning constraints of infants (insofar as languages not learnable by infants will be unlikely to be passed on to the subsequent generation). However, only some languages are additionally subject to the learning constraints of adults. The majority of languages are learned completely or almost completely by infants (consider also that half of the world's languages have fewer than 7000 speakers [16]). A numeric minority are learned primarily as non-native languages by a large proportion of speakers (e.g. English has 70% L2 speakers, [16]). Language acquisition during infancy (despite a lack of consensus accounts of underlying mechanisms) appears to be a relatively robust and fast process [1]. Infants transition from a few early vocalization types [33] to syntactic skill in a matter of a few years [41]. Second-language (L2) learning, by contrast, appears to pose substantial problems, especially for learning phonology and syntax: Individuals learning an L2 as adults can speak it for decades without measurable improvements in grammatical fluency [31].

Consider a language that is being spoken in the exoteric niche. As a language spreads over a larger area (e.g. as a result of colonization), it tends to accumulate more L2 learners for reasons of intermarriage, trade, cross-cultural exchange, migration, etc. Insofar as complex morphology poses a challenge to L2 learners, their learning is incomplete, and the languages they pass on to their children are simplified (see [25, 26] for further explication). The result is that over historical time, grammatical paradigms difficult for L2 learners to master become simplified [11, 27, 43, 46]. In particular, larger populations and corresponding increase in L2 learning may bring about a pressure on languages to drop morphological inflections, because these pose learning challenges and often are informationally redundant [27].

Why would complex morphological systems arise in languages to begin with? We theorized [25] that such systems may have arisen as an adaptation to the esoteric niche. That is, although complex morphological paradigms are dispreferred in the exoteric niche due to the learning difficulty they pose for adults, the same paradigms may paradoxically facilitate infant language acquisition. One reason is that because morphological markings of relational roles, number, evidentiality, possibility, possession, demonstratives, etc., make these distinctions obligatory, this added redundancy may help the infant learner connect the language to goings-on in the world around her. What appears to be redundant information to adult learners, may provide infants with additional cues to assist language learning, permitting them to rely less on extralinguistic cues and context to extract meaning (see [25] for discussion).

The linguistic niche hypothesis is consistent with much prior work exploring the origins of linguistic diversity [11, 27, 35, 43, 46]. However, at present, there is no computational demonstration of how differences in the language learning community can lead to differences in structure of language over historical time (though for related models see [5, 15, 23, 30]). The growing application of population dynamics [32], statistical mechanics [2, 4, 24], network analysis and dynamics [5, 7, 22], game theory [19], and agent-based modeling [12, 21, 40] to language change and evolution offers a wide range of possibilities for computational exploration (as evidenced by this volume, and see [3, 8, 20] for previous review). The purpose of the current paper is to develop a simple agent-based simulation of the dynamics predicted by the linguistic niche hypothesis, and to test one of the predictions with human data. In Sec. 2, we present an agent-based model that serves as a simple existence proof of the linguistic-niche hypothesis. The simulation makes the qualitative prediction that adult learners who are exposed to different languages, cultures, and a larger number of non-native speakers of the language ought to have different metalinguistic biases. These biases relate to the probability that an adult learner would embrace certain kinds of linguistic expressions after exposure to a large and more complex social structure. In Sec. 3, we substantiate this qualitative prediction with an empirical study.

2. Simulations

An agent-based framework was developed based on recent work in Chater *et al.* [6]. In this framework, agents are represented as bit vectors that encode particular grammatical features. In our case, we define an agent as a bit vector of M messages, $A = (0, 1, 1, \ldots)$, each dimension of which represents a particular message. The simplest way of encoding "morphology" in this system is to have the bit vector encode 1 on the messages that have "inflection", and 0 on those that do not. Some example agents are displayed in Fig. 1 (lower).

This simple representation forms the basis of this agent-based simulation. By connecting these agents, and imposing learning and communication constraints (see below), idealized linguistic change can be explored under various constraints. As iterations of the agent-based simulation unfold, an agent A's morphological encoding can be given a score as the proportion of messages that have inflection bits

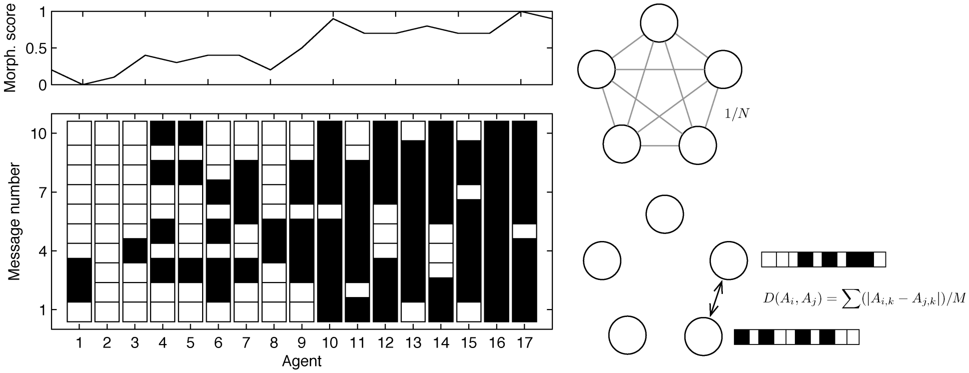

Fig. 1. Lower: A community of agents (bit-vectors) showing a range of morphological encoding (black = message is morphologically marked, white = not marked). Top: The morphological score (proportion of a vector containing 1s) for each of these example bit-vector agents. The higher the morphological score, the more the agents are encoding their messages with a morphological marker. Right: The social network structure is fully connected (top), and on an iteration of a simulation run, any two agents interact by comparing their bit vectors (bottom).

turned on (see Fig. 1, top). In the initial simulation below, we further detail how communication and linguistic change are implemented.

2.1. *Simulation 1: When adults learn, inflections are dropped*

In simulation 1, we assume that all agents can communicate with all others with equal probability. Given N agents in a population, the probability at any given point that an agent will communicate with another is $1/N$. On the first iteration of the simulation, agent vectors are set randomly with 0s and 1s. Upon each subsequent iteration of the simulation, we randomly select pairs of agents to communicate. This is done by simply comparing their bit vectors using a normalized Hamming distance: $D(A_i, A_j) = \sum(|A_{i,k} - A_{j,k}|)/M$. Agents are deemed to successfully communicate and do not change their language when $D(A_i, A_j) < \epsilon_1$. ϵ_1 is the lower bound on communication, and when agent codes differ only by this much, they are deemed to be members of the same linguistic group, and will not change their bit-vector encoding.

However, if two agents A_i and A_j have $D(A_i, A_j) \geq \epsilon_1$, but also reveal a smaller difference than some upper-bound of communication $D(A_i, A_j) < \epsilon_2$, then they will carry out an adjustment of their coding. Because simulation 1 is meant to capture the learning tendencies of adults who find complex morphology difficult to pick up, the agent with the highest morphological encoding score randomly has one of its on-bits set to 0, creating a bias against complex morphology. In the case that $D(A_i, A_j) \geq \epsilon_2$, the two agents do not successfully communicate and do not carry out any modification of their morphological coding. These simple communication rules form the basis of adult interaction in these simulations, and are not unlike basic communication rules in other work [6].

A single run of our simulation is composed of many iterations. Each iteration of a run is a bout of communication in which each agent has a turn communicating with one other agent. 100 separate runs of 500 iterations were carried out across a range of the parameters N (population) and M (messages).[a]

Results are shown in Fig. 2 (left), and reflect general trends of this simple "adult" system. When population is small, language systems rapidly stabilize in a particular morphological configuration. When population increases, languages become moving targets, and gradually lose inflectional on-bits. It can be shown that the lower bound of communication reflects an attractor for morphological coding in a growing population of learners: $\lim_{N \to \inf} \bar{A} = \epsilon_1$. In other words, in this simple

[a] We chose N and M parameters based on piloting of early simulations. These parameters can only be interpreted relatively. When N is larger, there are more agents in the populations. We do not intend for these parameters to reflect actual population values or demographic characteristics in real human societies — only the presence of different distributions over social networks and learning. More realistic parameter values may be carried out, as we describe below, if one takes into account social network structure and perhaps more dynamic features of cross-cultural interaction (e.g. between-group trade and interaction). Using small population sizes in existence proofs of this kind is common practice [6].

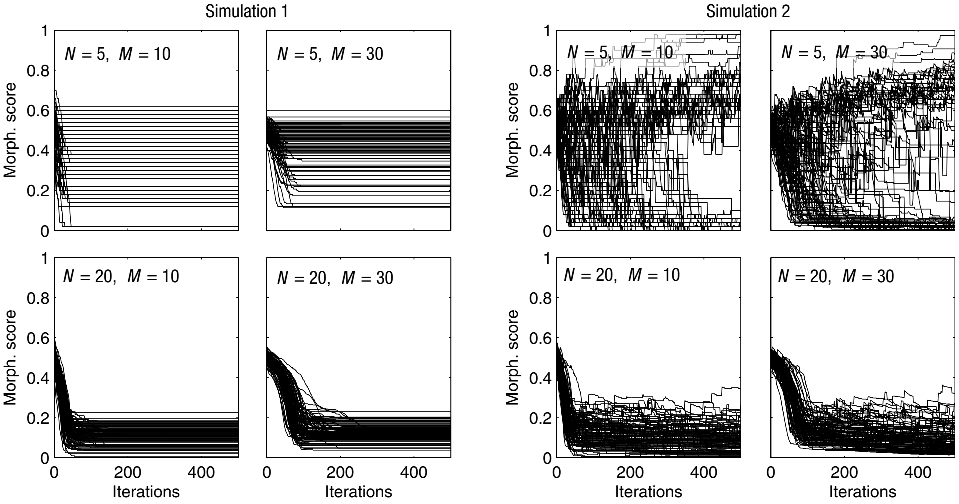

Fig. 2. Left panel: Results of the simulation with only "adult" agents. As above, morphological score is equal to the proportion of an agent's vector containing 1s (e.g. the proportion of messages morphologically marked). A line reflects the mean morphological score across N agents of a given run. In small groups ($N = 5$), languages can stabilize on a range of morphological complexity. Larger groups involve descent towards near-total loss of inflection. Right panel: When agents are occasionally recycled, and "infant" agents learn from adult agents, high inflection becomes an attractor for language change. However, as group size increases, loss of inflection again dominates. In both panels M represents the number of messages (dimensionality of bit vectors). In both $\epsilon_1 = 0.1, \epsilon_2 = 0.4$. In the second simulation, the death rate is 0.05 (probability of one death per iteration).

simulation, as populations grow, morphological encoding approach a score that is reflective of the lower bound of successful communication (the communicative "slack" that is permissible across agent "dialects").

The primary purpose of simulation 1 was to set up an agent-based framework for subsequent exploration. Its results are, perhaps, easily anticipated. It is important to note that in "esoteric" agent networks, stable inflecting strategies (in these simple systems) are capable of serving as equilibrium points because the population can descend upon a consistent code quickly enough to achieve the lower bound of communication ϵ_1. In communities composed of only adults, complex inflectional forms either reduced or remained stable. To demonstrate that complex morphologies can accumulate over time, we simulate the addition of "infant" agents into the population.

2.2. *Simulation 2: Adding children promotes inflection*

In the second simulation, the framework developed in simulation 1 was modified by adding gradual replacement of adult agents with "infant" agents. With some probability at each iteration, a random agent might "die" and be replaced with an "infant" that had its inflectional bit-vector set to undetermined values (akin to [6]). When this happened, that infant would engage in communication at the next iteration, and would set its bits in a manner often referred to as "frequency boosting", a pattern seen in real child learning [38]. This learning rule is simple, and results in an amplification of an already-existing encoding strategy. Let A_0 denote the new agent, A_i the randomly chosen teacher agent, and $\bar{A}_i$ this agent's mean encoding score. The frequency boosting works in the following way: The kth element of A_0 is set to the same value of A_i if that kth value in A_i is equal to the mode in the bit vector (Eq. (1)). Otherwise, the other bit values are randomly selected from a uniform distribution (Eq. (2)):

$$A_{0,k} = A_{i,k}, \quad \text{if } A_{i,k} = H\left(\bar{A}_i - \frac{1}{2}\right) \tag{1}$$

$$= H\left(U(0,1) - \frac{1}{2}\right), \quad \text{otherwise.} \tag{2}$$

H here is the step function and returns 0 if its input is less than or equal to 0, and 1 if its input is greater than 0. Because $\bar{A}_i$ lies between 0 and 1, subtracting 0.5 returns 0 if the most common encoding strategy is none (mostly 0s, $\bar{A}_i \leq 0.5$) or 1 if the most common strategy is morphological encoding (mostly 1s, $\bar{A}_i > 0.5$).

When running the same set of parameters N and M as described in simulation 1, the result is considerably different from the first simulation (see Fig. 2, right). With small groups, there are two distinct attractor states for morphological score: high and low. Across the 100 runs (of 500 iterations), a large proportion of the runs began to take on more complex encoding strategies, while another proportion descended towards loss of inflection. This emerges simply from "infant" agents frequency boosting the already-existing coding strategy of current adults. Importantly, when the simulation moves from an esoteric ($N = 5$) to a more exoteric ($N = 20$) niche, adult learning and interactions dominate, and the stable morphological code again approaches ϵ_1.

2.3. *Simulation 3: Child bias for inflection in learning*

Simulation 2 was based merely on "infant" agents boosting the prevalent strategy of the adults who interact with them. There was no bias for the infant agents to favor inflections, as discussed in the introduction. To implement this next, we slightly modified Eq. (2) above to favor inflection during frequency boosting: $H(U(0,1) - 1/2 + b)$. This parameter b simply increases the likelihood that when the infant agent's encodings are randomly set, the agent's learning rules now favor encoding $(= 1)$. For example, if $b = 0.5$, then there is an added 50% likelihood (given $U(0,1) - 1/2$) that a given element will become an on-bit. We carried out the same simulation as the previous one, this time with 2000 iterations (again, for 100 runs). Results are shown in Fig. 3 with $b = 0.5$, $N = 5$, $M = 10$, and ϵs the same as above. The strongest attractor state in this simulation is maximal encoding $(= 1)$ — accounting for a full 40% of all the simulation runs. Some runs of the simulation were still stable

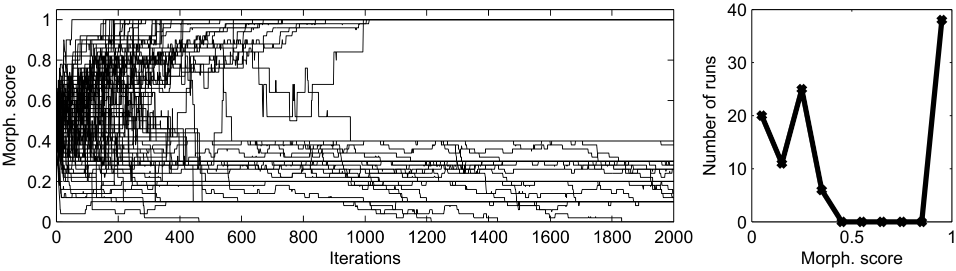

Fig. 3. Left: When the "infant" agents have a bias ($b = 0.5$) for greater morphological encoding due to the hypothesized benefit of overspecification for infant learning, complex morphology systems become the strongest attractors. Intermediate levels of coding are still possible. As above, morphological score is equal to the proportion of an agent's vector containing 1s (e.g. the proportion of messages morphologically marked). Right: A histogram of the observed end-point bit-vector states. The most stable attractor is maximal encoding, though other states are possible.

at lower encoding scores, but only half that number (20%) descended to a score of 0 (see Fig. 3, right). In separate simulations not shown here, if we assume that children occasionally create novel grammaticalized forms, then the lower encoding states become more unstable, and trend towards more elaborated systems in esoteric (small network) communities.

2.4. *Simulation 4: Incoming population of adults*

Imagine that a small community speaking language L experiences a sudden influx of outsiders who try to learn L. L now becomes increasingly subject to the learning constraints of these new adult learners. This scenario can be readily modeled in the current framework. We ran a version of simulation 3 for 500 iterations. The first 250 iterations had 5 agents, a small group that permitted the emergence of heavily coded bit vectors (see Fig. 4, iterations 1–250). In this simulation, at iteration 250, we introduced a larger number of novel adult agents who then proceeded to interface with the pre-existing community. This rapid emergence of adult interaction caused the high morphological scores to be unstable, and all the runs produced a precipitous drop in morphological complexity (see Fig. 4, iterations 251–500).

This introduction of "non-native" speakers in a growing interacting population leads to a simplification of the inflectional encoding of messages in the bit vectors. Thus, as the population of the language L grows, accumulating interactions with L2 learners who speak an approximation of the "original" version of L, the original group of speakers may be inclined to change their own linguistic code. Even when this change is slight, the changes accumulate from iteration to iteration. The linguistic niche hypothesis and the results of the simulation predict that something like this actually occurs. We provide an initial test of this prediction in a simple human study, presented next.

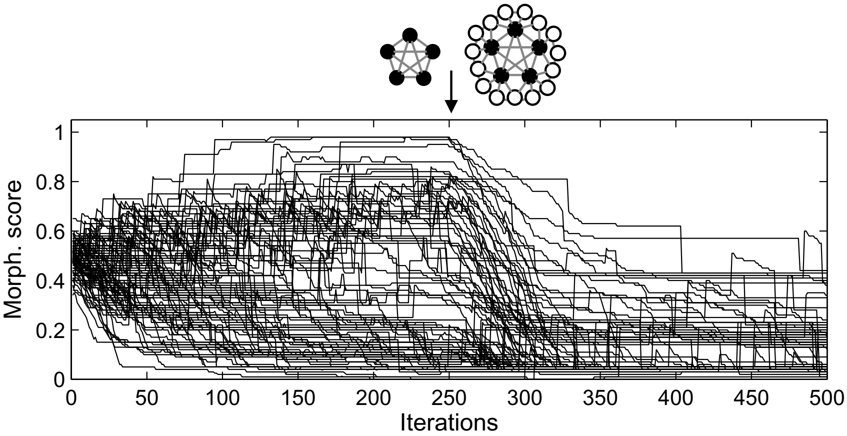

Fig. 4. 100 runs of 500 iterations with large influx of new adult speakers at iteration 250. For the first 250 iterations, $N = 5$, $\epsilon_1 = 0.1$, $\epsilon_2 = 0.4$, death rate = 0.05, $b = 0.3$. At iteration 250, N is set to 20. Importantly, these morphological scores for all 1–500 iterations (and all 100 runs) are computed only from the original $N = 5$ population. As above, morphological score is equal to the proportion of an agent's vector containing 1s (e.g. the proportion of messages morphologically marked).

3. Human Study

As noted in the previous paragraph, a growing population is likely to experience an influx of non-native speakers with whom the dominant linguistic group communicates. The correlation between L2 speakers and population size can be easily observed in natural languages [16]. According to the linguistic niche hypothesis and the results of simulation 4, we may observe these effects even *within* a language insofar as L1 speakers of that language may be exposed to different different proportions of L2 speakers. One interesting comparison is American and British English. As evidenced by n-gram analyses, American English is substantially more regular when it comes to the past-tense paradigm than British English, owing to greater regularization of irregular verb forms (see [28], Fig. 2G). Inspired by the observation made by Trudgill concerning apparent greater productivity (i.e. regularity) of derivational morphology in American versus British English, we decided to do our own informal analysis using Google search results from American and British websites as the dependent variable. As shown in Fig. 5, left, there is indeed a tendency for American English to generalize derivational patterns such as *pay/payee*, *retire/retiree* more than British English.

One common explanation for such differences in regularity/productivity is that American English is simply more innovative. If true, we might expect that just as novel regularizations are apparently welcomed in American English, so would occasional *irregularizations* (i.e. decreases in productivity) — both are innovations after all. Although a comprehensive test of this hypothesis is lacking at present, a telling example concerns one of the few verb forms which is at present

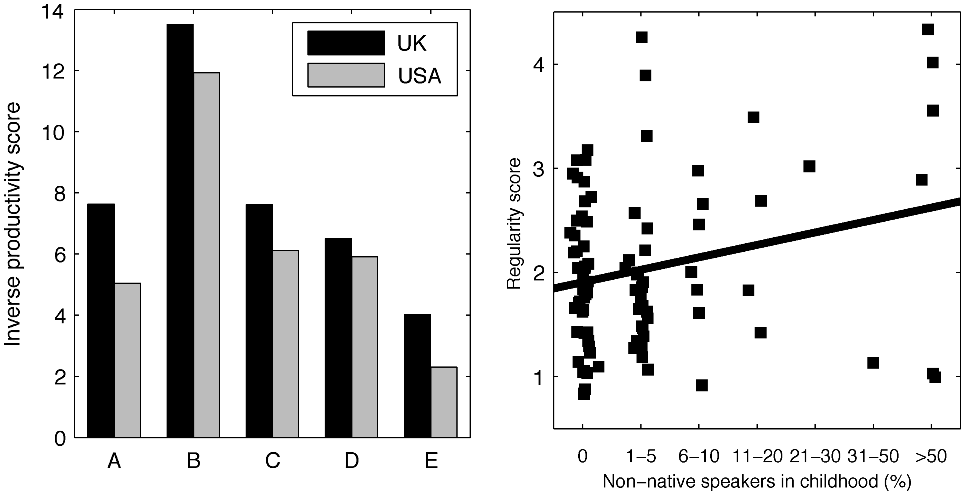

Fig. 5. Left: The productivity of several derivational suffixes in the US and UK. The inverse productivity score is equal to $\log(f_{\text{root}}/f_{\text{-ify}})$ where f is the number of returned Google hits of the word. Lower scores reflect greater productivity of the derivational forms. A = *ugly/uglify*, B = *city/citify*, C = *pressure/pressurize*, D = *pay/payee*, E = *retire/retiree*. Right: Endorsement rating of regularized forms of English verbs as a function of percentage of childhood non-native exposure. Note: The points have been plotted with some random jitter to avoid superimposition.

becoming irregular: the past tense of *to light* is shifting from *lighted* to *lit*. We performed a Google n-gram analysis (http://books.google.com/ngrams) examining the frequency trajectories of *light* and *lit*. In printed British English, *lit* overtook *lighted* in 1908. In contrast, in American English the switch did not take place until 1943. We also performed the light/lit comparison using the Google Trends dataset (http://www.google.com/trends), which tracks Google search keywords and is thus more up-to-date and reflects more casual language usage. It shows that in American English, the ratio of *lit* to *lighted* is 1.86 : 1. In British English, the ratio cannot be computed because there are insufficient occurrences of the term *lighted*. Thus, at least in this case, Americans appear to be *less* innovative when it comes to *irregularization*, once again trending towards greater regularity.[b]

The linguistic niche hypothesis offers an explanation: American English is more regular than British English because it has more non-native learners for whom deviations from dominant grammatical paradigms pose a greater learning problem than for L1 learners (80% of English speakers are L1 speakers in the US versus 95% in the UK, [16]). We sought to test this prediction in a human study involving *only* American English speakers. We recruited human participants to provide judgments about the past-tense of several English verbs that currently have both

[b] The use of *lighted* relative to *lit* in American English peaks at Christmas, apparently owing to references to lighting candles and other ornaments, but as revealed by a Google trends analysis, within the past 5 years this usage is becoming increasingly rare.

regular and irregular forms. We predicted that individuals who learned English in a more exoteric context (e.g. having more contact with non-native speakers) would be more likely to prefer regularized forms relative to individuals who learned English without exposure to L2 learners [42].

3.1. *Methods*

We recruited 95 participants from the United States online via Amazon Mechanical Turk [39]. These participants were queried on their opinions regarding regularization of the forms *lit/lighted*, *snuck/sneaked*, *sped/speeded*, *wed/wedded*, and *bent/bended*. Participants were presented with identical sentence frames for both irregular and regular verb forms (e.g. *They sneaked around/They snuck around*), and asked to rate acceptability of each verb on a scale from 1 (completely unacceptable) to 5 (sounds perfect). A "regularity score" was calculated for each participant from their average score on this scale for the (over)regularized verbs. In addition to gathering the participants' ratings on these sentences, we also asked them about their level of education, and social environment, including whether their parents were born in the US, what proportion of their friends and close acquaintances were non-native speakers of English, and what proportion of such speakers were friends during childhood. These latter two questions used an ordinal scale: (1) 0%, (2) 1–5%, (3) 6–10%, (4) 11–20%, (5) 21–30%, (6) 31–50%, (7) greater than 50%. The great majority of participants fell into the first three response bins. In addition to this self-reported degree of exposure to non-native speakers, we also used US Census records to calculate an objective measure of prevalence of non-native speakers in each US state. This *native-speaker advantage* score was calculated by dividing the (log-transformed) number of households in each state reporting that only English is spoken in the home by the number of households reporting languages other than English spoken in the home. We reasoned that individuals growing up in states with greater native-speaker advantage scores would be less likely to be exposed to non-native English speakers (and, by extension, to varieties of English spoken by non-native speakers).

3.2. *Results*

One participant was excluded due to being an extreme outlier in verb judgments, and two others were excluded for failing to answer questions correctly. An additional 8 did not indicate in what state they grew up and thus are missing native-speaker advantage scores. Given the expected overall preference for the irregular versions of forms such speeded/sped, all participants preferred the irregular forms over the regular forms, paired t-test: $t(91) = 17.96$, $p < 0.0005$; the preference for the irregular forms was also significant in a by-item analysis, $t(4) = 11.40$, $p < 0.0005$. Importantly, the endorsement of regular forms was predicted by the participants' self-reported proportion of friends/acquaintances they knew while growing up who were not native English speakers, Pearson's $r = 0.23$, $p = 0.025$ (see Fig. 5, right).

This relationship remained significant when education was partialed out, $r = 0.24$, $p = 0.024$ (education by itself did not predict acceptability ratings). The relationship was also significant in an ordinal logistic regression, controlling for education: $odds\!-\!ratio = 0.78$, $p = 0.028$. The *current* proportion of non-native English speakers known by the participant was not a reliable predictor of sentence acceptability, $p > 0.2$.

The native-speaker advantage, computed from US Census records, was significantly correlated with mean regularity, $r = -0.23$, $p = 0.035$, and remained significant when controlling for education, $r = -0.24$, $p = 0.031$. A multiple regression analysis controlling for education showed that the native-speaker advantage score remained a significant predictor of endorsement of regular forms controlling for the self-reported proportion of native-speakers known in childhood, $b = -0.48$, $t = 2.07$, $p = 0.042$. The latter predictor was no longer significant with native-speaker advantage score in the model, $t < 1$.

This simple and admittedly preliminary study provides evidence that the social environment in which a language is learned is associated with detectable systematic differences on acceptability judgments: Individuals who reported knowing more non-native speakers when they were growing up tended to more strongly endorse (over-)regularized forms of past-tense verbs: a result consistent with the prediction from the linguistic niche hypothesis. The present results leave open many questions such as whether the observed association is due to participants' direct experience with non-native English speakers or their experience to a linguistic milieu variously shaped by non-native speakers. The present results, however preliminary, are valuable in that they connect typically separated avenues of inquiry in the language sciences: The mechanisms hypothesized to create diachronic changes may be acting in measurable ways in synchronic patterns of behavior. The linguistic niche hypothesis may serve as an explanatory bridge between these patterns of change at different levels of social scale, from social groups to individual learners.

4. Conclusion

In this paper we developed an agent-based framework, and ran a series of simulations that serve as a basic existence proof of the linguistic niche hypothesis. When simple agents following simple learning rules are combined in social networks, social dynamics have large effects on the structure of resultant languages. At the level of individual agent behavior, simulation 4 suggests that adult agents in a growing social context may be biased towards morphological simplification. We provided some preliminary evidence for this in a human study of English verb regularization. The model shows that (a simplified notion of) linguistic typology can be constrained by subtle learning constraints accumulating over extended periods of time; the human data suggest that these subtle learning constraints may be observable in human subjects.

The current approach has a number of limitations (which we are addressing in ongoing work). The simulations implement morphological encoding in a very simple way, and are based on small social networks. Motivated by previous work [6], the simulations serve as a computational framework in which the linguistic niche hypothesis can be explored and gradually scaled up. Although simplistic, the assumptions we made in our simulations are well within the scope of complexity common to many formal analyses of language in the relevant literature (e.g. form-reference mappings as in [32]). A specific example of such scaling up is encoding syntactic rules in the agents' grammars, such as whether determiners (e.g. English *the*) appear before or after their nouns [10]. In this way, grammatical rule systems could be integrated in the current simulation. In addition, it may be possible to integrate it with explorations of other network properties, such as small-world structure [22]; these structures may be potentially mapped onto quantitative data from natural linguistic typology [18] as well as artificial grammar-learning experiments.

In the current paper, we have implemented a simple computational framework for testing predictions of the linguistic-niche hypothesis [25], previously articulated in a variety of forms, [11, 27, 43, 46]. When agents of varying cognitive constraints converge in social networks, the resulting linguistic systems have equilibria that are partly a function of the structure of that social network. In the case of simulation 4, this may be an emergent property of the social network. In this simulation, added adult members of a community broaden the linguistic environment, and create opportunities for slight linguistic change within each round of interaction. The new stable patterns having reduced morphological encoding result from the cumulative effect of this communicative "sampling" in a larger and more densely packed social network.

The empirical data we present are also subject to simplifying assumptions and are best viewed as a starting point for a full-scale investigation. Despite these shortcomings, the fact remains that acceptability judgments of English sentences are predictable from such seemingly irrelevant factors as the self-reported proportion of non-native speakers in the speaker's social circle during childhood, but not at present times, and the proportion of households using languages other than English in the US state where the participant grew up. These observed correlations, although requiring further explication, may derive from the interaction between our subjects, who are members of the dominant linguistic group, and non-native English speakers. The effects of these interactions may relate to recent experimental findings suggesting that significant change in artificial languages can occur through coordination between pairs of participants (see e.g. [14, 37]), affording another possible avenue for investigating these linguistic tendencies.

In sum, we presented here a simple computational instantiation of the linguistic niche hypothesis: the claim that languages adapt to the social environment in which they are learned and used. Insofar as adults and children differ in their learning biases, languages "optimized" for adult and child acquisition come to have different structures. Our computational framework provides an initial existence proof of this hypothesis, showing, for example, how languages composed purely of L1 learners can stabilize at high levels of morphological complexity and how introducing L2 learners into a population composed of L1 learners can lead to morphological simplification. The empirical data provide additional evidence, however preliminary, that the impact of social structure on language is observable even within a single generation, making the results potentially relevant to understanding language change on a range of timescales.

References

[1] Bloom, P., *How Children Learn the Meanings of Words* (The MIT Press, Cambridge, MA, 2000).

[2] Blythe, R. and Croft, W., The speech community in evolutionary language dynamics, *Lang. Learn.* **59** (2009) 47–63.

[3] Cangelosi, A. and Parisi, D., *Simulating the Evolution of Language* (Springer London, 2002).

[4] Castellano, C. and Loreto, V., Statistical physics of social dynamics, *Rev. Mod. Phys.* **81** (2009) 591.

[5] Castelló, X., Toivonen, R., Eguluz, V. M. and Miguel, M. S., Modelling bilingualism in language competition: The effects of complex social structure, *Proceedings of the 4th Conference of the European Social Simulation Association* (2007) 581–584.

[6] Chater, N., Reali, F. and Christiansen, M., Restrictions on biological adaptation in language evolution, *Proc. Natl. Acad. Sci.* **106** (2009) 1015.

[7] Choudhury, M., Mukherjee, A., Basu, A., Ganguly, N., Garg, A. and Jalan, V., Language diversity across the consonant inventories: A study in the framework of complex networks, in *Proceedings of the EACL 2009 Workshop on Cognitive Aspects of Computational Language Acquisition* (Association for Computational Linguistics, 2009), pp. 51–58.

[8] Christiansen, M. and Kirby, S., Language evolution: Consensus and controversies, *Trends Cogn. Sci.* **7** (2003) 300–307.

[9] Christiansen, M. H. and Chater, N., Language as shaped by the brain, *Behav. Brain Sci.* **31** (2008) 489–508.

[10] Christiansen, M. H. and Dale, R., The role of learning and development in language evolution: A connectionist perspective, in *Evolution of Communication Systems: A Comparative Approach*, Oller, D. K. and Griebel, U. (eds.) (MIT Press, Cambridge, MA, 2004), pp. 91–109.

[11] Dahl, Ö., *The Growth and Maintenance of Linguistic Complexity* (John Benjamins Publishing Co, 2004).

[12] De Boer, B., Self-organization in vowel systems, *J. Phonetics* **28** (2000) 441–465.

[13] Evans, N. and Levinson, S. C., The myth of language universals: Language diversity and its importance for cognitive science, *Behav. Brain Sci.* **32** (2009) 429–448.

[14] Galantucci, B. and Garrod, S., Experimental semiotics: A new approach for studying the emergence and the evolution of human communication, *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems* (2010), pp. 1–13.

[15] Gong, T., Ke, J., Minett, J. and Wang, W., A computational framework to simulate the coevolution of language and social structure, in *Artificial Life IX: Proceedings of*

the 9th International Conference on the Simulation and Synthesis of Living Systems (2004), pp. 158–164.

[16] Gordon, R. G., *Ethnologue: Languages of the World*, 15th edn. (SIL International, 2005).

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

[18] Holman, E. W., Schulze, C., Stauffer, D. and Wichmann, S., On the relation between structural diversity and geographical distance among languages: Observations and computer simulations, *Linguist. Typol.* **11** (2007) 393–422.

[19] Jaeger, G., Evolutionary stability conditions for signaling games with costly signals, *J. Theor. Biol.* **253** (2008) 131–141.

[20] Jaeger, H., Baronchelli, A., Briscoe, E., Christiansen, M. H., Griffiths, T., Jaeger, G., Kirby, S., Komarova, N., Richerson, P. J., Steels, L. and Triesch, J., What can mathematical, computational and robotic models tell us about the origins of syntax? in *Biological Foundations and Origin of Syntax*, Bickerton, D. and Szathmary, E. (eds.), Vol. 3 (Cambridge, MA: MIT Press, 2009), pp. 385–410.

[21] Kirby, S., Natural language from artificial life, *Artif. Life* **8** (2002) 185–215.

[22] Klemm, K., Eguíluz, V. M., Toral, R. and San Miguel, M., Nonequilibrium transitions in complex networks: A model of social interaction, *Phys. Rev. E* **67** (2003) 026120.

[23] Komarova, N. and Levin, S., Eavesdropping and language dynamics, *J. Theor. Biol.* **264** (2010) 104–118.

[24] Loreto, V., Baronchelli, A., Mukherjee, A., Puglisi, A. and Tria, F., Statistical physics of language dynamics, *Journal of Statistical Mechanics: Theory and Experiment* **2011** (2011) P04006.

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

[26] McWhorter, J., The world's simplest grammars are creole grammars, *Linguist. Typol.* **5** (2001) 125–166.

[27] McWhorter, J., What happened to english? *Diachronica* **19** (2002) 217–272.

[28] Michel, J., Shen, Y., Aiden, A., Veres, A., Gray, M., Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J. *et al.*, Quantitative analysis of culture using millions of digitized books, *Science* **331** (2011) 176.

[29] Nettle, D., Explaining global patterns of language diversity, *J. Anthropol. Archaeol.* **17** (1998) 354–374.

[30] Nettle, D., Using social impact theory to simulate language change, *Lingua* **108** (1999) 95–117.

[31] Newport, E. L., Maturational constraints on language learning, *Cognitive Sci.* **14** (1990) 11–28.

[32] Nowak, M., Komarova, N. and Niyogi, P., Evolution of universal grammar, *Science* **291** (2001) 114.

[33] Oller, D., The emergence of the sounds of speech in infancy, *Child Phonology* **1** (1980) 93–112.

[34] Payne, D. L. and Payne, T. E., Yagua, *Handbook of Amazonian Languages* **2** (1990) 249–474.

[35] Perkins, R., *Deixis, Grammar, and Culture* (J. Benjamins Publishing Company, 1992).

[36] Sapir, E., *Language: An Introduction to the Study of Speech* (Harcourt, Brace and company, 1921).

[37] Selten, R. and Warglien, M., The emergence of simple languages in an experimental coordination game, *Proc. Natl. Acad. Sci.* **104** (2007) 7361.

[38] Singleton, J. and Newport, E., When learners surpass their models: The acquisition of American Sign Language from inconsistent input* 1, *Cognitive Psychol.* **49** (2004) 370–407.

[39] Sprouse, J., A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory, *Behavior Research Methods* (2011), pp. 1–13.

[40] Steels, L. and Belpaeme, T., Coordinating perceptually grounded categories through language: A case study for colour, *Behav. Brain Sci.* **28** (2005) 469–489.

[41] Tomasello, M., *Constructing a Language: A Usage-Based Theory of Language Acquisition* (Harvard University Press, Cambridge, MA, 2005).

[42] Trudgill, P., *Dialects in Contact* (Blackwell Oxford, 1986).

[43] Trudgill, P., *Sociolinguistics: An Introduction to Language and Society,* 4th edn. (Penguin (Non-Classics), 2001).

[44] Trudgill, P., *Linguistic and Social Typology* (Oxford University Press, 2002), pp. 707–728.

[45] Trudgill, P., *Sociolinguistic Typology: Social Determinants of Linguistic Structure and Complexity* (Oxford University Press, 2011).

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