Demographic shifts, inter-group contact, and environmental conditions drive language extinction and diversification

Marco Tulio Pacheco Coelho[1,2], Hannah J. Haynie[1,3], Claire Bowern[4], Robert K. Colwell[2,5,6], Simon J. Greenhill[7,8], Kathryn R. Kirby[7,9], Thiago F. Rangel[2], Michael C. Gavin[1,7]

[1] Department of Human Dimensions of Natural Resources, Colorado State University, Fort Collins, CO, USA
[2] Departamento de Ecologia, ICB, Universidade Federal de Goias, Goiania, Goias, Brazil
[3] Department of Linguistics, University of Colorado at Boulder, Boulder, CO, USA
[4] Department of Linguistics, Yale University, New Haven, CT, USA
[5] Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT USA
[6] University of Colorado Museum of Natural History, Boulder, CO USA
[7] Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena, Germany
[8] CoEDL (ARC Centre of Excellence for the Dynamics of Language), Australian National University, Canberra, Australia
[9] Department of Ecology and Evolutionary Biology, University of Toronto, Ontario, Canada

Humans currently collectively use thousands of languages[1,2]. The number of languages in a given region (i.e. language "richness") varies widely[3–7]. Understanding the processes of diversification and homogenization that produce these patterns has been a fundamental aim of linguistics and anthropology. Empirical research to date has identified various social, environmental, geographic, and demographic factors associated with language richness[3]. However, our understanding of causal mechanisms and variation in their effects over space has been limited by prior analyses focusing on correlation and assuming stationarity[3,8]. Here we use process-based, spatially-explicit stochastic models to simulate the emergence, expansion, contraction, fragmentation, and extinction of language ranges. We varied combinations of parameter settings in these computer-simulated experiments to evaluate the extent to which different processes reproduce observed patterns of pre-colonial language richness in North America. We find that the majority of spatial variation in language richness can be explained by models in which environmental and social constraints determine population density, random shocks alter population sizes more frequently at higher population densities, and population shocks are more frequently negative than positive. Language diversification occurs when populations split after reaching size limits, and when ranges fragment due to population contractions following negative shocks or due to contact with other groups that are expanding following positive shocks. These findings support diverse theoretical perspectives arguing that language richness is shaped by environmental and social conditions, constraints on group sizes, outcomes of contact among groups, and shifting demographics driven by positive innovations, such as new subsistence strategies, or negative events, such as war or disease.

Humans currently use over 7000 different languages[1,2]. These languages form the basis of all human communication, underpinning social and political groups and contributing to both cooperation and conflict. Linguistic diversity is also an indicator of broader patterns of cultural diversity, including diverse worldviews that serve as the foundation of all human beliefs, norms, and knowledge systems. Linguistic diversity takes on many forms, including phylogenetic language diversity (i.e. the length of branches that span taxa on a phylogenetic tree) and linguistic disparity (i.e. variation in expression of a linguistic trait within a clade)[3]. We focus on language richness, which is a measure of the number of distinct languages spoken in a given region. Languages are spread unevenly across the planet. In some regions hundreds of languages are spoken, whereas in others only one predominates. For example, in pre-colonial North America far more languages were spoken on the west coast (e.g., in what is now California) than in the northeast (e.g., in what is now Newfoundland and Labrador) (see Fig 2a). However, despite the importance of language diversity, much is still unknown about the processes that drive language diversification. Here we use process-based simulation modeling to uncover the mechanisms driving language diversification in North America.

Toward a causal understanding of language diversification

Empirical research to date has identified associations between environmental, geographical, and sociocultural variables and spatial patterns in language richness. Language richness is distributed along a

strong latitudinal gradient, with far more languages, and languages with smaller geographical ranges, in the tropics than toward the poles[7,9–12]. Therefore, it should come as no surprise that many climatic and ecological factors that also exhibit clear latitudinal patterns correlate with language richness. These factors include net primary productivity, rainfall, temperature, mean growing season, and biodiversity[4,6,10,13–16]. Geographically, mountains—and topographic complexity more generally, rivers, habitat heterogeneity, and the size and distance to nearest landmasses of islands have all been shown to have some association with richness[9,12,17–23]. Sociocultural factors that influence the size and organization of human societies, including patterns of population movement, subsistence strategies, and political complexity, have also been linked to language richness patterns[3,18,24,25].

Most prior empirical research, however, has relied on correlational approaches that face substantial methodological challenges. Most notable is the well-known adage that correlation does not imply causation. Latitude, for example, correlates with language richness, but crossing the artificial "lines in the sand" that latitude represents obviously has no bearing on the languages that individuals speak. Instead, other factors that also correlate with latitude, such as environmental conditions, resource availability, and potential population density, may have mechanistic links with language diversification and extinction processes. In other cases, the associations that studies uncover may be attributed to different processes that are difficult to differentiate using correlation alone. For example, mountains may increase richness by serving as barriers to human movement, leading to increased rates of population fragmentation. Alternatively, mountains might increase habitat heterogeneity, with different resources available at different altitudes, which in combination with niche partitioning could contribute to language diversification[19]. Prior research has also operated under the assumption that the factors associated with language richness, and by implication the mechanisms driving diversification, do not change from one region to another. However, recent research has demonstrated the non-stationarity of factors associated with language richness in North America[8], highlighting that the mechanisms causing language extinction and diversification can differ among locations. In addition, previous studies have often excluded processes known to shape language richness patterns over time, most notably the effects of language extinctions. Overall, while prior studies have played an important role in highlighting the potential links between richness and various factors, we are still far from a comprehensive mechanistic understanding of the processes involved.

To address these shortcomings, we used process-based simulation models. These models use clearly defined rules and interpretable parameters to test the degree to which specific mechanisms reproduce observed patterns[26]. These simulation models serve as *in silico* experiments, in which certain parameters can be altered while holding others constant in order to ascertain the effects each parameter has on the final outcomes. By isolating the effects of particular processes on richness patterns, these simulation approaches bring us closer to a causal understanding of diversification. Here, we developed a spatially explicit stochastic simulation model that simulates the emergence, expansion, contraction, and extinction of languages.

As far as we know, only two prior studies[8,27] have used process-based simulation models to test the relative effects of previously proposed processes driving language diversification. The first found that over half of the spatial variation in language richness in Australia could be explained by three simple processes: (i) humans move to fill unoccupied spaces, (ii) population densities are determined by environmental conditions, and (iii) human groups have a maximum population size, which when exceeded leads to the emergence of new groups speaking new languages (Fig. 1a). The second study followed up by demonstrating that this simple model does not generalize to other locations, including the fact that the model explains less than 20% of spatial variation in North America[8].

Here we build on the prior simulation models by adding complexity in two fundamental ways. First, we allow the processes of diversification to vary across space. Second, we experiment with a more

comprehensive range of processes that have been proposed to shape language richness patterns but have remained untested.

Simulation modeling to date has not considered the effects of changes in human populations over time. Relatively rapid changes in population sizes, which we will refer to as *shocks*, may be positive (i.e., increased population size) or negative (i.e. decreased population size), and could be limited to a particular group (i.e., "group shocks") or felt by all groups within a given region (i.e., "regional shocks") (Fig. 1b). These shocks correspond to a variety of hypothetical processes that have been proposed as drivers of human population changes and ethnolinguistic diversity. Group shocks might include, but are not limited to, innovations that promote population increases (e.g., new subsistence strategies, such as the development of agriculture), or internal turmoil that leads to population decreases. Regional shocks may include, but are not limited to, changing environmental conditions that affect subsistence (e.g., drought), regional conflicts or warfare, or the emergence of particular environmental resources (e.g., domesticable species or diseases) that can affect all human populations in the region.

Shocks lead to changes in potential population sizes that might cause range expansion or contraction. Range contraction can lead to fragmentation and diversification through splitting (cladogenesis) (Fig. 1c). Alternatively, range contraction might lead to a process that prior models have not considered: language extinction. Shock-driven range expansion leads to contact among groups speaking different languages. Although contact has long been considered a major driver of all forms of language diversity[28], prior simulation models have not considered the diverse potential outcomes of contact. Following contact, groups and their languages may remain relatively intact, or the expanding group may replace (extinguish), displace, or fragment the populations and languages they contact[29] (Fig. 1c).

Our model starts with the empty geographical domain of North America. We measured time using artificial algorithmic cycles, which were not intended to represent any historical time, or event. At each algorithmic cycle, cladogenesis occurred as new languages emerged when there was empty space surrounding existing languages (Fig. 1a). Languages emerged from a randomly selected cell and spread to neighboring cells. Initial population sizes were sampled from a theoretical distribution that assumed a higher probability of sampling smaller group sizes, with an upper limit represented by an adjustable standard deviation. For each cell we calculated carrying capacity using prior research[30] that modeled global population density (people per km^2), based on the direct and indirect effects of productivity, topography, precipitation seasonality, distance to coast, resource ownership norms, and residential mobility. A population and the language it spoke could spread to the total number of neighboring cells that would support the given population size based on the carrying capacity of cells.

We explored the effects of different probabilities of occurrence of group and regional shocks, as well as the impact of the proportion of shocks that were negative versus positive. The positive and negative shocks in our model are akin to many different processes that have occurred over human history that altered population sizes and language richness patterns. For example, much has been written about innovations in subsistence practices, particularly the advent and spread of agriculture and related changes in political organization, that increased the population sizes that human societies could maintain and subsequently impacted language richness patterns[18,31]. Other shocks, such as disease pandemics, can have negative impacts on human populations, and researchers have proposed links between disease and language diversification processes[32].

How much of the population was gained or lost due to shocks was determined based on a sample drawn from a uniform distribution with an adjustable minimum and maximum limit. Population loss either led to extinction (if loss was 100%) or to randomized range contraction until the adjusted geographical distribution supported the reduced population size based on the cells' carrying capacity. The stochasticity of range contraction could cause range fragmentation, and the resulting fragments of the original language

range were considered independent languages (Fig. 1c). In other words, range fragmentation could drive cladogenesis.

Positive shocks caused population size increases, which led to range expansion. In cases where range expansion led to contact with existing populations speaking other languages, the probability of colonizing occupied territory was weighted with a higher probability given to colonizing territories occupied by smaller populations. Colonization of occupied territory could lead to the extinction of the colonized population, or to the fragmentation of their ranges and language cladogenesis (Fig. 1c). Because processes affecting language extinction and diversification might vary in space[8], we also explored non-stationarity in the probability of group shocks by allowing a higher shock probability in locations with higher population densities. The model ran until reaching a predefined number of algorithmic cycles that were necessary to stabilize spatial patterns of language richness.

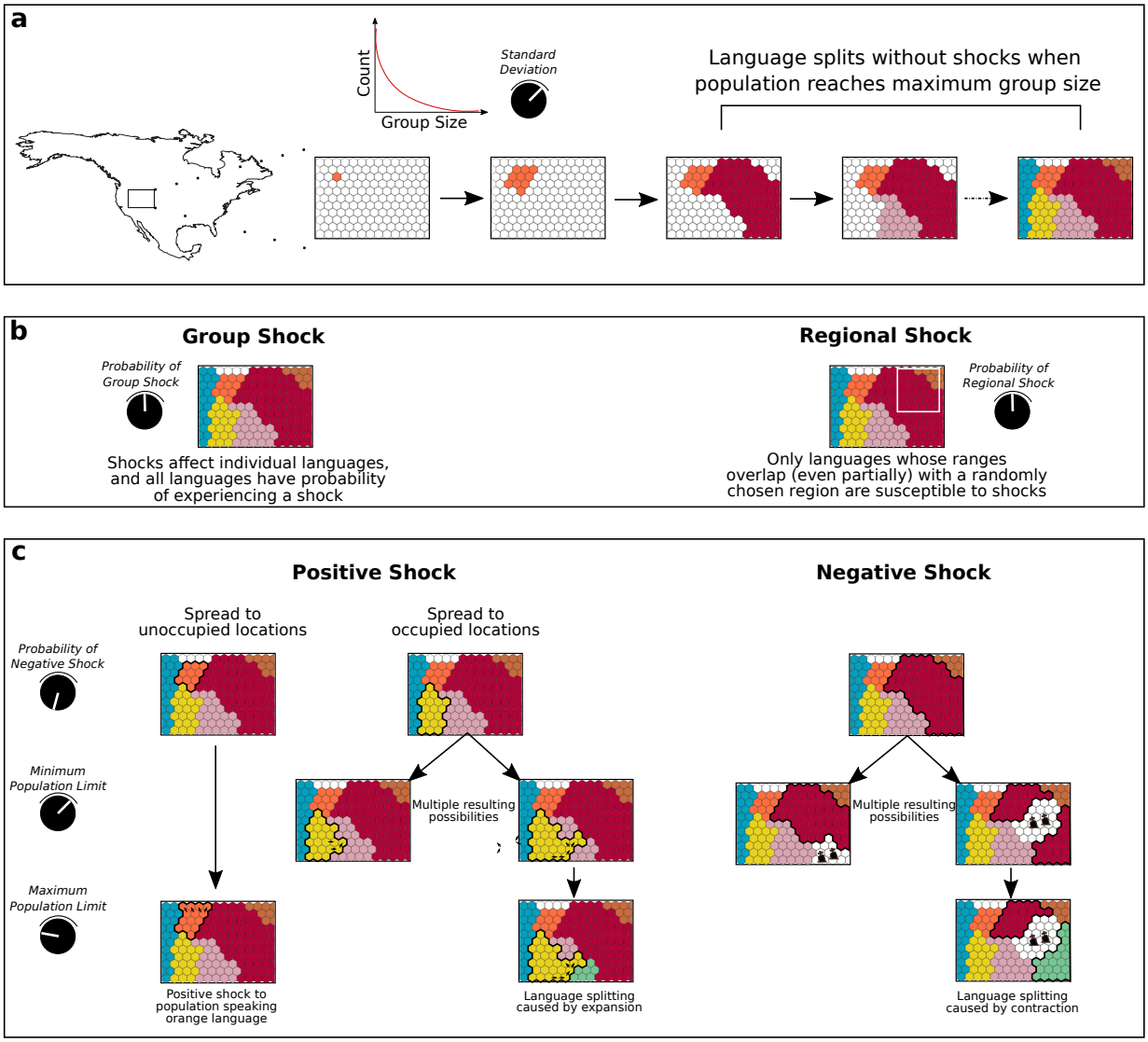

Figure 1. Process-based simulation models test the impacts of specific mechanisms on spatial patterns of language richness. (A) Models begin when an ancestral language emerges from a randomly selected cell. Population size is defined by a sample drawn from a normal curve with mean of zero and adjustable standard deviation. A population, and its language, spreads to the number of cells that support its size given underlying carrying capacities (see Methods). In each algorithmic cycle, if empty space occurs at the border of an extant

language range, a new language emerges from a randomly selected cell. (B) At each cycle, all extant languages are susceptible to positive or negative group and/or regional shocks, which vary in probability across different iterations of the model (see Methods), and cause increases or decreases in population size. (C) Shocks drive range expansion or contraction that may cause extinctions (not pictured) or range splitting and the emergence of new languages.

We used a factorial design to understand the role of each adjustable parameter on emergent patterns of language diversity, by running ~100,000 simulations, each with a different parameter combination. Our models explored the following parameters: the probability of group, regional, and negative shocks, the minimum and maximum limits for population decrease and growth, the threshold used to represent the slope of the linear relationship between the probability of group shock and population density, and the standard deviation of the theoretical distribution used to sample group size. This factorial parameter exploration allowed us to identify parameter sets which better fit the empirical data, but also allowed us to explore the effects of modifications to one parameter while holding others constant. We inspected the effect of each parameter modification on the emerging spatial pattern of language diversity and the total number of languages that emerged in the domain.

Results and Discussion

To evaluate the results of our simulation we compared the number and spatial pattern of languages in the simulated analyses to that of the observed pattern. We found that population shocks drive diversification, and as shocks become more frequent, the total number of languages increases (Fig S1, S2). However, the spatial patterns produced with more frequent shocks do not resemble the observed pattern, indicating that frequent shocks alone are not a major driver of richness patterns. Instead, our results suggest that when group shocks occur with a higher probability in locations with greater population densities, the simulation has a better fit to the observed spatial patterns of language diversity (Fig S3). This result suggests that drivers of population shocks and subsequent changes in language richness patterns are more likely at higher population densities. Our finding is consistent with recent research that has found support for the surplus theory for the origins and spread of agriculture, which argues that productive environmental conditions permit higher population densities and increase the likelihood of agricultural innovations[30]. The spread of politically complex agriculturalist societies has been linked to regional reductions in language richness[31]. More generally, higher population densities have also been linked to higher innovation rates, more disease spread, and greater rates of war and conflict, all of which can have substantial impacts on population dynamics and language richness [33–36].

We also find that the similarity between the simulated and observed patterns of language diversity increases when negative group shocks are more frequent than positive group shocks (Fig S4). Although positive shocks in our model can lead to range fragmentation and increases in richness, they can also cause language extinction. However, when positive shocks are too frequent, we tend to see the spread of a small number of languages each with a relatively large population and a large geographical distribution. We find that some positive shocks, that may mimic key social, political, or subsistence innovations, are needed to produce observed patterns of language richness. However, negative group shocks, driven by events such as pandemics and internal conflicts, need to occur at even higher frequencies in order to reproduce observed spatial patterns in language richness. Although our model does not purport to recreate specific historical events, but rather studies the effects of general processes, we note that some key positive shocks have been proposed for pre-colonial North America. For example, positive shocks occurred in the form of two centers for the origins of agriculture in what are the present-day southeastern United States and central Mexico[37]. Also, evidence of pre-colonial negative shocks, in the form of conflict and disease, have been reported from different regions (e.g.,[38–40]).

The population size of groups at time of origin in our model also has an impact on the fit between empirical and simulated patterns. When the size of emerging groups has a larger standard deviation, we

observe a lower total number of languages (**Fig. S5**). The fit between empirical and simulated spatial patterns of language diversity is highest when the standard deviation varies between 7000 and 8000 speakers (**Fig. S5**). This result also resembles historical population patterns in North America. During the pre-colonial era, the continent supported many hunter-gatherer societies. However, groups that relied heavily on agriculture did exist in several regions [41]. Agriculture, and the often-higher level of political complexity associated with it, tended to promote larger group sizes[42]. Other parameters did not have any notable impact on model performance (see supporting information).

Based on the factorial parameter exploration and the visual inspection of the effect of each parameter in isolation, we selected the three parameters that were the most impactful on language richness patterns and explored these parameter spaces by means of MCMC (see Methods). We found that the fit between empirical and simulated patterns is best when negative shocks occur at a higher frequency (62%) than positive shocks (38%), when standard deviation of maximum population sizes is set to about 7000, and when the slope of the linear relationship between probability of a group shock and population density is about 0.02 (**Fig. S10**). These parameter values represent the mean values of the parameter landscape explored by the MCMC sampler, which shows that the Gibbs sampler spent most of its time in these regions of parameter space that clearly has a better fit to empirical data (Fig. S10). We ran 200 replicates with the best parameter set to compute the mean explanatory power and the mean total number of languages. Our model explains a mean of 63% of the variation between empirical and simulated spatial patterns of language diversity. The model produces a mean of 759 languages in North America (**Fig 2**; observed = 414 languages).

Our model predicted more languages than expected most notably in the western part of the continent, especially in what is currently western Mexico (see model residuals Fig. 2e). Predicted language richness was lower than expected in a few locations, such as the northeastern U.S., but these positive residuals tended to be smaller and less frequent than the negative residuals. The size and location of model residuals may be shaped in part by limitations of the observed language range map. Language ranges both before and after colonial contact have been in constant flux due to dynamic social-ecological conditions. Although the observed map represents the best approximation available of language ranges at the time of European contact, the exact dates that individual polygons intend to represent vary depending on data availability (see Methods). In some locations, colonial impacts may have dramatically impacted language richness before any available recording was made. Conflicts and disease brought by Europeans caused substantial population loss[43], which could have reduced language richness in impacted regions. Therefore, locations with high negative residuals may represent areas in which the observed map underrepresents language richness due to colonial impacts.

Any mechanisms missing from our simulation models could also contribute to the size of residuals. For example, our model may be under- or over-predicting richness by not accounting for lags in the time needed for positive or negative shocks to take effect. We might expect that some positive shocks, such as new subsistence strategies, would take a longer time to influence population densities and ultimately language richness patterns. However, some negative shocks, such as disease or warfare might have a substantial impact on populations within relatively short time periods. Our model currently does not account for variation in the time needed for shocks to take effect. This could be particularly important if the effects of some shocks increase the probability of others. For example, the innovation of agriculture has been shown to have variable effects on population size and density over time[42], and our model includes mechanisms through which these changes would influence spatial patterns in language richness. However, agricultural societies also have a higher probability of increased political complexity, and greater complexity can allow larger groups of people to maintain social cohesion and one language[18]. Our model currently would depict these two innovations (agriculture and political complexity) via the occurrence of separate positive shocks, but these events may not be strictly independent. Future modeling

efforts that more accurately accounts for time lags in some shocks and possible additive effects of shocks over time may be able to further improve the fit with the observed language range map.

In addition, some previously proposed mechanisms for language diversification and homogenization in North America are not represented in our simulations. For example, prior research points to the importance of temperature and precipitation constancy as key factors linked to high language richness on the west coast of the continent, in the exact regions where our analysis produces the biggest residuals[8]. Climatic constancy has been suggested as a proxy for ecological risk[8], and societies may buffer against ecological risk by developing more expansive social networks that provide more stable access to resources and also reduce the probability of language diversification[6]. Future work that seeks to incorporate additional mechanisms such as these may be able to improve fit in regions with the largest residuals from the current model.

Unlike previous correlational studies that could, at best, imply the role of particular mechanisms, we explicitly modeled five interacting general mechanisms: (i) population densities shaped by social and environmental conditions[30]; (ii) groups spreading out to establish geographical distributions based on limits to their population sizes and the carrying capacity of the land they occupy; (iii) the potential population size of groups varying within a broad distribution and affected by negative (e.g. war, conflict, disease, drought) and positive (subsistence or political innovation, improved environmental conditions) shocks; (iv) negative group shocks that occur at higher probabilities when population densities are higher, leading to range contraction that can cause extinction or range fragmentation and cladogenesis; and (v) positive group shocks that also occur more frequently at higher densities, leading to range expansion and contact with other groups, which may in turn face extinction or range fragmentation and cladogenesis. We conclude that a combination of these processes can simulate prominent aspects of the variation in the geographical distribution of precolonial languages in North America.

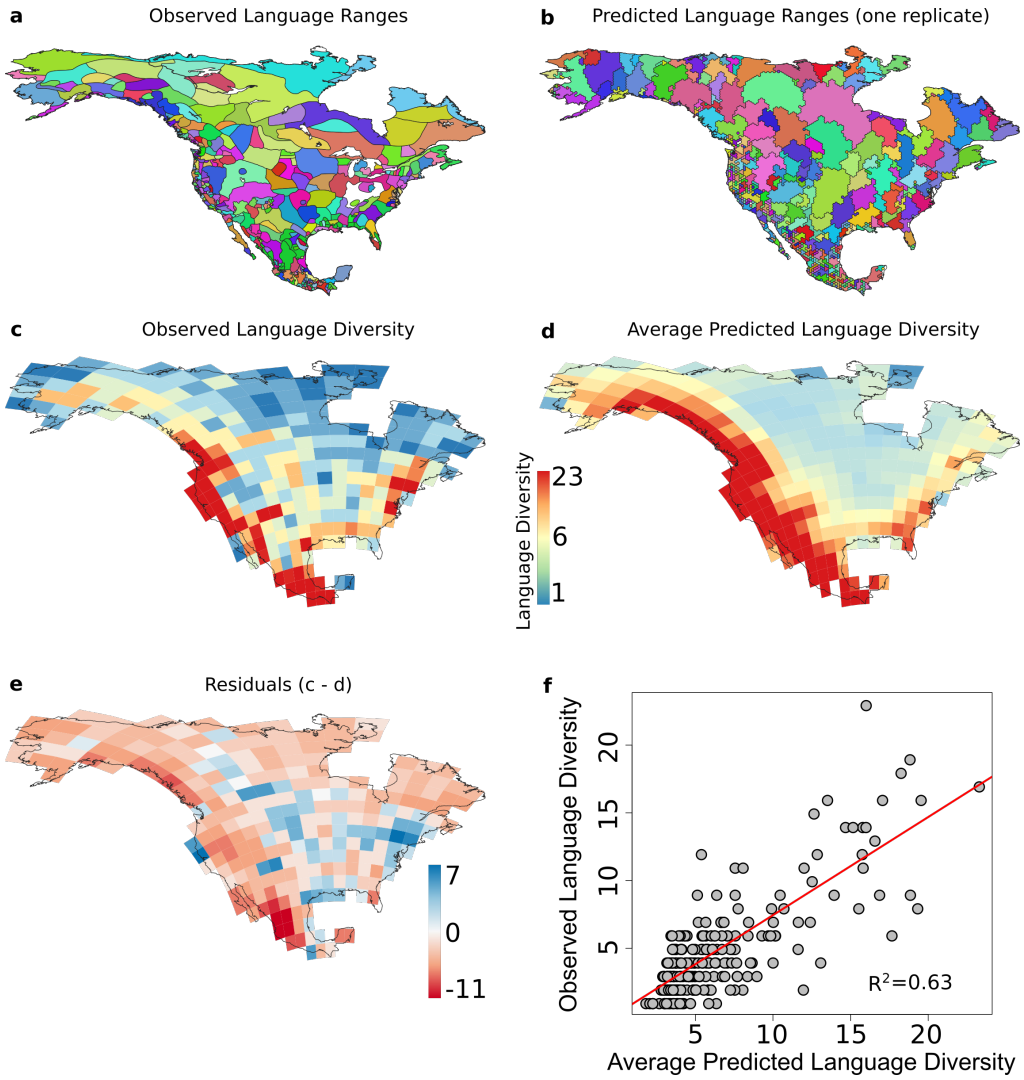

Figure 2. Observed and predicted patterns of language richness in North America**.**

Methods

Observed language richness patterns

Languages may be defined by mutual intelligibility, but in practice identifying boundaries between languages faces many challenges. We relied on classifications of languages from established literature and previously published North American language range maps (see [44] for full details of the methods and metadata used to create the observed language range map). We used a large set of source maps to assess language range locations. We developed a set of ranked priorities to assess alternative sources and resolve conflicting boundary placements (see [44]). Although we followed conventions of previous North American maps in depicting language ranges as non-overlapping polygons, we recognize this approach may not capture all the complexity inherent in various types of settlement and land use, language contact, and bilingualism. We also emphasize that this approach is unable to erase the colonial legacy of prior language mapping projects, or correct errors in prior language range representations.

We used the best approximations available of language ranges immediately prior to European colonial incursion. However, data availability varies, and therefore the date individual polygons purport to represent also varies (see [44]).

Simulation Assumptions

Our model was spatially explicit. We assumed that carrying capacity for forager groups varied across the geographical space. We measured time using artificial algorithmic cycles, which were not intended to represent any historical time or event.

Geographical space

Over the domain of North America we built an equal area hexagonal gridded map of 0.5x0.5° in which we simulated language range dynamics. For each hexagonal cell we calculated the carrying capacity for individuals based on estimations of population density (people per km^2) for forager societies[30]. Kavanagh et al.[30] fitted a piecewise structural equation model with direct and indirect effects of productivity, topography, precipitation seasonality, distance to coast (i.e. access to marine resources), resource ownership (i.e. whether resource are owned or not), and residential mobility (i.e. average distance travelled per residential move) to empirical forager societies. Their model explained 77% of the variation in population density among observed foraging societies and was used to estimate population density at global scale at 0.5x0.5° resolution. Thus, our hexagonal gridded map was set to the same resolution to avoid the interpolation of population density data to higher resolutions. In addition, cells must be large enough to encompass a group of individuals, but smaller than most observed language ranges in North America, which was consistent with our grid resolution. On top of the hexagon grid map, a rectangular grid cell was used to compute simulated and empirical language diversity (number of languages per geographical cell). This rectangular gridded map was built in an equal area resolution of 300x300km^2 to ensure that grid cells were small enough to capture the variation in language diversity across space. The same resolution was used to characterize language diversity in North America and in Australia in previous studies[8,27].

Language Emergence and Expansion

Languages originated in the geographical space with a predefined population size (see model dynamics), and their geographical distribution was defined by the number of hexagonal grid cells necessary to sustain the group's population. Thus, groups tended to have broader geographical distributions in regions with lower carrying capacity (e.g. deserts) because more geographical cells were necessary to support the population. However, the geographical distribution of languages was not fixed, because we also modelled stochastic events that could lead to increases or decreases in populations, which would in turn affect their geographical distribution. All language groups were susceptible to these random population shocks that could cause geographical range contraction, expansion (to occupied or unoccupied locations), splitting and extinction.

Population Shocks

Over the course of language evolution and spread over the geographical domain, groups in our simulations were susceptible to population shocks that could occur at the level of language groups (i.e. group shocks), or at regional level (i.e. regional shocks) (Fig. 1b). Group shocks correspond to a wide variety of different hypothetical processes that could affect a group's

population and language ranges. These might include, but are not limited to, innovations by the group that promote population increases, or internal turmoil that leads to population decreases. Contrary to group shocks, regional shocks occur at a regional level and not to a specific group. These regional shocks correspond with different hypothetical processes that affect a given location and cause changes in human populations in that location. These processes might include, but are not limited to, changing environmental conditions that affect subsistence, or the presence/emergence of particular environmental resources (e.g. domesticable species or diseases) that can affect the human populations in the region. Group shocks and population shocks can be either positive or negative and cause respectively increase or decrease in group size and group geographical distribution.

Model dynamics

Our model started with an empty geographical domain. In the first algorithmic cycle, one hexagonal cell was randomly selected from the domain. An ancestral language, from which all languages descended, emerged from the randomly selected cell and spread to neighboring cells. A population size for the emerging language was defined as the absolute number sampled from a normal curve with mean equal to zero and an adjustable standard deviation. This theoretical distribution assumed that there was a higher probability of sampling smaller group sizes than larger ones with an adjustable upper-tail controlled by the standard deviation. In addition, for this study, we did not sample group size from the empirical distribution of forager groups as we did in previous studies[27], because the processes we modeled allowed group size to increase or decrease over the course of the simulation.

Based on the sampled population size and the carrying capacity of hexagonal cells where the population was located, the population and its language could spread to the total number of neighboring cells that would support its given population size. If the initial, randomly chosen cell had a carrying capacity that was equal to or greater than the ancestral language population size, the ancestral language would occupy a single hexagonal cell. However, if the carrying capacity of the initial cell was smaller than the ancestral population size, the population would spread to neighboring cells and stop spreading once its total population was supported (Fig. 1a).

After the ancestral language emerged and spread over empty space it was susceptible to group and/or regional shocks that were controlled respectively by the probability of group shocks *P(GS)* and the probability of regional shocks *P(RS)*. Group shocks could affect all extant populations (*i.e. P(GS)* was tested for all living languages), whereas regional shocks affected one region with an adjustable impact area (Fig S9). If a regional shock occurred, all languages inside a randomly defined region were affected by population shocks. When group and/or regional shocks occur, shocks can be positive or negative. The probability of positive shock *P(PS)* was equal to 1 - probability of negative shock *P(NS).* Population shocks had a 50% chance of being positive or negative when *P(NS)* was set to 0.5, but our model allowed for an adjustable higher or lower probability of positive or negative shocks when *P(NS)* was lower or higher than 0.5.

If a language was affected by a negative shock, the language lost part of its population. How much population was lost was determined based on a sample drawn from a uniform distribution with an adjustable minimum (minimum population limit, *MinPL*) and maximum (maximum population limit, *MaxPL*) limit that varied from *MinPL* to 100%. Because a language's range

size was defined by the number of hexagonal cells that could support its population size, population loss led to distributional range contraction (Fig. 1c). To simulate range contraction one random hexagonal cell was selected at the border of the language range, and if the population loss was higher than the population occurring in the randomly selected cell, that cell was lost and assumed to be empty (i.e., not colonized by a language). This range contraction spread to neighboring cells until all the population that had been lost in the shock was subtracted from the original population size. If *MaxPL was* set to 100%, the entire population was lost, and the language went extinct. In addition, the stochasticity of range contraction could lead to range fragmentation that could represent, but was not limited to, internal conflict leading to group fragmentation. When range fragmentation occurred, the resulting fragments of the original language range were considered independent languages (Fig. 1c). Thus, each language was assumed to have a strictly continuous distribution, so that range fragmentation always resulted in language splitting.

If a positive shock affected a language group, its population increased. How much the population increased was also determined based on a sample drawn from a uniform distribution with minimum and maximum limits according to adjustable *MinPL* and *MaxPL.* If a positive shock occurred, the language range expanded. If there were empty cells at the border of the language affected by the shock, the language range expanded its distribution to empty cells to support the increasing population (Fig. 1c). However, if the expanding language group was surrounded by geographical cells that were already occupied by other language groups, then the expanding group spread over occupied locations, coming into conflict with a neighboring language group. Because languages can be surrounded by multiple language groups with different characteristics, the language group with an increasing population was assigned a higher probability of spreading its range to locations that were occupied by languages with smaller populations. Thus, the probability of colonizing occupied territory was weighted by neighboring language populations with a higher probability of colonizing territories occupied by smaller populations. The stochasticity of language expansion to occupied locations could also cause the fragmentation of the language range(s) that were colonized by the expanding population. Because a language has only continuous distributions in our simulation, these language fragmentations caused by the expansion of neighboring populations also caused language splitting (Fig. 1c). Finally, language extinction could also occur during positive shocks, when the expansion of a language range colonized all the geographical cells that were previously occupied by another language group. Thus, the conquered language was assumed to go extinct.

The model initialized another algorithmic cycle (i.e., time step) after the probability of group and environmental shocks was tested. In new cycles, each new language emerged in a randomly selected cell at the border of any living language (e.g,. in the second cycle a new language emerged at the border of the ancestral language) and expanded its distribution to neighboring cells until it had occupied enough territory to support its sampled group size. Subsequent steps were repeated as described above. Thus, at any algorithmic cycle (Fig S10), new languages emerged when there was empty space surrounding living languages (Fig 1a), and group and regional shocks could affect all living languages. The model ran until reaching a number of algorithmic cycles that were necessary to stabilize spatial patterns of language diversity. The stabilization of the spatial pattern of language diversity was measured by the serial correlation between the number of languages in each grid cell over time steps.

Non-stationary population shocks

Because processes affecting the emergence of spatial patterns of language diversity might have variable effects in different regions of the geographical domain[8], population shocks were modelled as non-stationary processes. In our simulation, group shocks were modelled with higher probability in locations with higher population density. We assumed a linear relationship between probability of shocks and population density. The intercept of this relationship was set to zero. By definition, *P(GS)* varied from zero to one. Thus, the slope of the linear relationship was defined as a ratio between 1 and an adjustable *threshold* to set all values above the threshold to equal 1. Therefore, if y = slope*x, we set y to be the probability of shock and x to be population density. To make sure that y was never above 1 the slope was calculated as 1/*threshold*. So, if 40 people per km^2 was the threshold, the slope was 0.025 (y = 0.025*40 = 1). Alternatively, if 1000 people per km^2 was the threshold, the slope was 0.001.

Parameter exploration

To understand the role of each adjustable parameter on emergent patterns of language diversity, we ran 96,768 simulations, each with a distinct combination of parameter settings. This factorial design consisted of running the model with all possible combinations of parameter values. Probability parameters (i.e., *P(GS), P(RS)* and *P(NS)*) and minimum and maximum limits for population decrease and growth (i.e., *MinPL* and *MaxPL*) had clearly defined limits (i.e., from 0 to 1 and from 0 to 100%) and were varied to represent low, medium and high variations of these parameters. *P(GS)* and *P(RS)* higher than 0.04 (i.e., either higher than 4% chance that a given group present at a given time step experienced a shock, or higher than 4% chance a given region experienced a shock at a given time step) produced more than 2000 languages in North America, a number many times greater than reality. Thus, the maximum limits of these parameters were set to 0.04. We explored values of the threshold used to represent weak, medium and strong relationships between *P(GS)* and population density. Finally, the standard deviation of the theoretical distribution used to sample group size had no clearly defined maximum values. Here we explored values of the standard deviation that could generate maximum group sizes that were observed in empirical foraging populations.

The factorial parameter exploration allowed us to identify parameter sets that better fit the empirical data, but also allowed us to explore the effects of modifications to one parameter while holding others constant (see Supporting information). Thus, to better understand the effect of each parameter in isolation, we inspected the effect of each parameter modification on the emerging spatial pattern of language diversity and the total number of languages that emerged in the domain. This approach gave us a better understanding of which parameters most affected language diversity, allowing us to robustly fit the most important parameters instead of all adjustable parameters.

To fit the most important parameters we also explored the parameter space using a Gibbs sampler Markov Chain Monte Carlo simulation[45]. We designed the Gibbs sampler to explore the landscape of parameter values approximating the posterior distribution of the model, given the empirical data. The summary statistic used in the Gibbs sampler was the goodness of fit (f) of the model, measured as the similarity between simulated and empirical patterns. Here, f was defined as:

$$f = \frac{R^2 + \left(1 - \frac{|ObsNLang - PredNLang| + 1}{Max(ObsNlang, PredNlang) + 1}\right)}{2}$$

where R^2 is the percentage of variation explained when comparing the empirical and simulated pattern of language diversity, *ObsNLang* is the observed total number of languages in North America, and *PredNLang was* the predicted total number of languages in North America. Thus, *f* varied from zero to one, and the maximum fit would be observed only when R^2 equaled one and *PredNlang* was equal to *ObsNlang (i.e. 414 languages)*.

We ran the Gibbs sampler for 11,000 iterations and assumed 1,000 iterations as a burn-in period. Each iteration replicated the stochastic model 100 times. The MCMC chains were tested for convergence following the Heidelberg and Welch's convergence diagnostic[46], which consists of a two-step convergence diagnostic. First, the diagnostic evaluates if the chain is a stationary distribution by comparisons of multiple subdivisions of the first half of the chain, to the latter 50% portion of the chain. If the chain passes the stationarity test, then the diagnostic calculates a 95% confidence interval of the mean value of the chain. Half of the width of the confidence interval is then compared to the mean value of the chain (i.e. half-width test). If the ratio between the half-width and the mean is lower than a critical value (usually 0.1), then the chain passes the test. We used one parameter set, assuming the mean values of each parameter after running the MCMC and the convergence statistics, to produce the average language diversity observed in each rectangular cell of North America and the average total number of languages that emerged in the geographical domain.

Literature Cited

1. Hammarström, H., Forkel, R. & Haspelmath, M. *Glottolog 4.4*. (Max Planck Institute for the Science of Human History, 2021).
2. Eberhard, D. M., Simons, G. F., & Fennig, C. D. Ethnologue, 24th edition. (SIL International, 2021). *Online version: http://www. ethnologue.com.*
3. Gavin, M. C. *et al.* Toward a mechanistic understanding of linguistic diversity. *BioScience* **63**, 524–535 (2013).
4. Gorenflo, L. J., Romaine, S., Mittermeier, R. A. & Walker-Painemilla, K. Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. *PNAS* **109**, 8032–8037 (2012).
5. Greenhill, S. J. Demographic correlates of language diversity. in *The Routledge handbook of historical linguistics* 557–578 (Routledge, 2014).
6. Nettle, D. *Linguistic diversity*. (Oxford University Press, 1999).
7. Gavin, M. C. & Stepp, J. R. Rapoport's Rule Revisited: Geographical Distributions of Human Languages. *PloS one* **9**, e107623 (2014).
8. Pacheco Coelho, M. T. *et al.* Drivers of geographical patterns of North American language diversity. *Proceedings of the Royal Society B* **286**, 20190242 (2019).
9. Hua, X., Greenhill, S. J., Cardillo, M., Schneemann, H. & Bromham, L. The ecological drivers of variation in global language diversity. *Nature communications* **10**, 2047 (2019).
10. Collard, I. & Foley, R. Latitudinal patterns and environmental determinants of recent human cultural diversity: do humans follow biogeographical rules? *Evolutionary Ecology Research* **4**, 371–383 (2002).
11. Nichols, J. *Linguistic diversity in space and time*. (University of Chicago Press, 1992).
12. Mace, R. & Pagel, M. A latitudinal gradient in the density of human languages in North America. *Proceedings of the Royal Society of London B: Biological Sciences* **261**, 117–121 (1995).
13. Moore, J. L. *et al.* The distribution of cultural and biological diversity in Africa. *Proceedings of the Royal Society of London B: Biological Sciences* **269**, 1645–1653 (2002).

14. Nichols, J. Modeling ancient population structures and movement in linguistics. *Annual review of anthropology* **26,** 359–384 (1997).
15. Nettle, D. Explaining global patterns of language diversity. *Journal of anthropological archaeology* **17**, 354–374 (1998).
16. Sutherland, W. J. Parallel extinction risk and global distribution of languages and species. *Nature* **423**, 276–279 (2003).
17. Nettle, D. Language diversity in West Africa: An ecological approach. *Journal of Anthropological Archaeology* **15**, 403–438 (1996).
18. Currie, T. E. & Mace, R. Political complexity predicts the spread of ethnolinguistic groups. *PNAS* **106**, 7339–7344 (2009).
19. Stepp, J. R., Castaneda, H. & Cervone, S. Mountains and biocultural diversity. *Mountain Research and Development* **25**, 223–227 (2005).
20. Gavin, M. C. & Sibanda, N. The island biogeography of languages. *Global Ecology and Biogeography* **21**, 958–967 (2012).
21. Axelsen, J. B. & Manrubia, S. River density and landscape roughness are universal determinants of linguistic diversity. *Proceedings of the Royal Society B: Biological Sciences* **281**, 20133029 (2014).
22. Cashdan, E. Ethnic Diversity and Its Environmental Determinants: Effects of Climate, Pathogens, and Habitat Diversity. *American Anthropologist* **103**, 968–991 (2001).
23. Amano, T. *et al.* Global distribution and drivers of language extinction risk. *Proc. R. Soc. B* **281**, 20141574 (2014).
24. Nettle, D. Linguistic diversity of the Americas can be reconciled with a recent colonization. *Proceedings of the National Academy of Sciences* **96**, 3325–3329 (1999).
25. Renfrew, C. World linguistic diversity. *Scientific American* **270**, 104–111 (1994).
26. Gotelli, N. J. *et al.* Patterns and causes of species richness: a general simulation model for macroecology. *Ecology Letters* **12**, 873–886 (2009).
27. Gavin, M. C. *et al.* Process-based modelling shows how climate and demography shape language diversity. *Global Ecology and Biogeography* **26**, 584–591 (2017).
28. Thomason, S. G. & Kaufman, T. *Language contact, creolization, and genetic linguistics*. (Univ of California Press, 1992).
29. Bellwood, P. The dispersals of established food-producing populations. *Current Anthropology* **50**, 621–626 (2009).
30. Kavanagh, P. H. *et al.* Hindcasting global population densities reveals forces enabling the origin of agriculture. *Nature human behaviour* **2**, 478 (2018).
31. Diamond, J. & Bellwood, P. Farmers and Their Languages: The First Expansions. *Science* **300**, 597–603 (2003).
32. Fincher, C. L. & Thornhill, R. A parasite-driven wedge: infectious diseases may explain language and other biodiversity. *Oikos* **117**, 1289–1297 (2008).
33. Shennan, S. *Pattern and Process in Cultural Evolution*. (University of California Press, 2009).
34. Boserup, E. *The conditions of agricultural progress*. (Aldine Publishing Company, 1965).
35. Wolfe, N. D., Dunavan, C. P. & Diamond, J. Origins of major human infectious diseases. *Nature* **447**, 279–283 (2007).
36. Acemoglu, D., Fergusson, L. & Johnson, S. Population and conflict. *The Review of Economic Studies* **87**, 1565–1604 (2020).
37. Larson, G. *et al.* Current perspectives and the future of domestication studies. *Proceedings of the National Academy of Sciences* **111**, 6139–6146 (2014).
38. LeBlanc, S. A. & LeBlanc, S. A. *Prehistoric warfare in the American Southwest*. (University of Utah Press, 1999).
39. Martin, D. L. & Harrod, R. P. *The bioarchaeology of violence*. (University Press of Florida, 2012).
40. Martin, D. L. & Goodman, A. H. Health conditions before Columbus: paleopathology of native North Americans. *Western Journal of Medicine* **176**, 65 (2002).
41. Doolittle, W. E. Agriculture in North America on the eve of contact: a reassessment. *Annals of the Association of American Geographers* **82**, 386–401 (1992).
42. Gignoux, C. R., Henn, B. M. & Mountain, J. L. Rapid, global demographic expansions after the origins of agriculture. *Proceedings of the National Academy of Sciences* **108**, 6044–6049 (2011).
43. Denevan, W. M. The pristine myth: the landscape of the Americas in 1492. *Annals of the Association of American Geographers* **82**, 369–385 (1992).

44. Haynie, H. J. & Gavin, M. C. Modern language range mapping for the study of language diversity. *SocArXiv,* https://doi.org/10.31235/osf.io/9fu7g (2019).
45. Gelman, A. *et al. Bayesian Data Analysis*. (CRC Press, 2013).
46. Plummer, M., Best, N., Cowles, K. & Vines, K. CODA: convergence diagnosis and output analysis for MCMC. *R news* **6**, 7–11 (2006).