Linguistic diversity as a factor influencing disease spread

Sihan Chen[a,1] and Antonio Benítez-Burraco[b]

This manuscript was compiled on March 15, 2025

We test the hypothesis in the literature that linguistic diversity hinders disease spread: people speaking different languages contact less. We first analyze COVID-19 transmission data in countries differing in linguistic diversity, as well as U.S. counties differing in linguistic diversity. We controlled for factors such as development level, geography, and climate. We then simulate the disease transmission in societies varying in size, degree of contact, and network density - three sociopolitical factors impacting language diversity. We found evidence supporting that language diversity correlates with reduced transmission rates.

language diversity | disease transmission | language evolution | mathematical simulation | data analysis

Humans speak more than 7000 languages belonging to different language families (1). The reasons for this astounding diversity remain elusive, although different factors have been suggested to trigger and maintain language diversity. Languages adapt to the physical and social environment in which they are spoken, which are very diverse (2). Typically, the sound pattern of a language is shaped by factors such as temperature or humidity, which impact sound production and transmission (3–5). Likewise, the number of speakers, their modes of living, the type of social network they form, or the status of the language can favor complexity or simplicity in specific domains, like morphology or syntax (6–10), thus rendering different types of languages (11), and ultimately fostering diversity. As a general rule, ecological stability promotes population isolation, which in turn stimulates language diversity (12)). By contrast, ecological risk promotes population contact, which typically erodes diversity (13, 14). Still, high levels of linguistic diversity can be maintained within densely populated areas if languages are strong identity markers (15).

In his 2016 book, Andrea Moro hypothesized that language diversity might have served to mitigate the problems of population growth in eras in which technology was inadequate to provide enough food supplies and particularly, health standards to large assemblages of people. This can make even more sense in view of evidence that the spread of cultural innovations can favor the emergence and spread of diseases (16). Following Moro's intuition, in this paper we aim to test the specific possibility that language diversity acts as a protective barrier to disease spread.

We have followed two different, but still complementary approaches. First, we collected epidemiological data on the recent COVID-19 pandemics and compared the spread of the disease in two different scenarios: macro and micro. In the first case, we compare the global spread of COVID-19, specifically, between world regions that show opposite linguistic landscapes, but are as similar as possible in terms of their geography, climate, and living conditions, and more specifically, the factors known to impact the spread of COVID-19. The risk that a pathogen will spread within a population is mostly influenced by population-level factors, such as population density, facility of movement, and public health responses (Sands and others 2016), but also by individual susceptibility to infection, which mostly depends on living conditions and their environmental correlates (17, 18). In the case of COVID-19, factors known to have an impact on the transmission of the disease are air pollution, temperature, humidity, solar radiation, population density, and socioeconomic activities such as trade, and certainly, health condition and health policies (19, 20). Accordingly, in our research we selected areas of the world with different Linguistic Diversity Index (LDI), but with similar altitude, latitude, climate, population density, road density, and Human Development Index (HDI). In the micro scenario, we compared the spread of COVID-19 at a local level, specifically between US counties that show opposite linguistic landscapes, but are as similar as possible in terms of the physical and social factors known to affect the spread of COVID-19.

Significance Statement

Humans speak more than 7000 languages. This notable diversity results from the effect of environmental, social, cognitive, and language-internal factors interacting complexly. One of such factors has been claimed to be a protective effect of diversity to disease: people speaking different languages contact less. We tested this possibility and found that worldwide language diversity negatively correlates with Covid-19 diffusion rates. At a local level, this pattern can be reversed depending on societal conditions. We further modelled the transmission of a disease in different types of societies hypothesized to speak typologically diverse languages. We found that disease spreads differently in different societies. Overall, we found support to the view that language diversity might be sensitive to disease dynamics among human groups.

Author affiliations: [a]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02144; [b]Department of Spanish, Linguistics & Theory of Literature (Linguistics), University of Seville, Spain

ABB conceived the paper. SC conducted the data analyses. SC and ABB analyzed the results. SC and ABB wrote and approved the final manuscript.

The authors declare no conflicts of interests.

[1]To whom correspondence should be addressed. E-mail: sihanc@mit.edu

www.pnas.org/cgi/doi/10.1073/pnas.XXXXXXXXXX

Type	Size	Network size	Contact with other societies
1	small	tight	low
2	small	tight	high
3	small	loose	low
4	small	loose	high
5	large	loose	low
6	large	loose	high

Table 1. The six basic types of society interactions outlined in (8), characterized by their size, network typology, and contact between different communities.

Second, to clarify the mechanism of language diversity on disease spread, we implemented a mathematical model to test if a pathogenic vector behaves differently in different types of societies, in view of the principal role played by sociopolitical diversity in triggering and shaping language diversity. Two basic types basic types of societies can be posited: esoteric and exoteric (after 21). Esoteric societies typically consist of small and tightly knit human groups that maintain reduced contact with other groups. By contrast, exoteric societies are larger, consist of looser social networks, and are involved in higher rates of inter-group contact and cultural exchange. These two types of societies have been shown to speak languages with different, somehow opposite typological features (11). Nevertheless, the exoteric vs. esoteric distinction could be seen as too basic, since it subsumes, as noted, different sociopolitical features. By this reason, in our analyses, we have considered instead Trudgill(8) 's senary societal typology, which results from the differential interaction between 3 main factors: community size (small vs large), network topology, (tight vs loose), and degree of contact (low vs high) (See Table 1). According to Trudgill, these 6 societal types account for the typological diversity of human languages (seemingly together with selected physical factors impacting mostly, as noted on their phonological features). In truth, Trudgill's types 1 and 6 can be seen as strongly esoteric or strongly exoteric societal types, respectively. Therefore, we have modeled whether a pathogenic vector spreads differentially in these six types of societies, with the ultimate aim of testing if changing one or several of the 3 societal features highlighted by Trudgill results in changes in the dynamics of the disease, and particularly if moving towards esoteric-like societies accelerates disease spread, as these societies are typically less linguistically diverse.

Results

Epidemiological analysis. To investigate disease dynamics in different linguistic landscapes, we have gathered epidemiological data on Covid-19 spread in selected world regions, as well as counties in the United States from the John Hopkins Coronavirus Resource Center*. We calculated the mean daily transmission rate for each country in the global analysis and for each U.S. county in the U.S. analysis (See Materials and Methods for more details). We filtered out countries and U.S. counties with drastically different transmission rate, in order to avoid these data points skewing the results. To quantify linguistic diversity, in the global analysis, we used the Linguistic Diversity Index (LDI) published by Ethnologue (1), whereas in the U.S. Analysis, we pulled data from the American Community Survey (ACS) published by the U.S. Census Bureau and calculated the entropy (H) of the language distribution in each county. A higher LDI and a higher H entails higher linguistic diversity. We also controlled for other factors that might also impact disease spread, such as physical environment and social development level (See Table 2 for a list of the variables being controlled).

<table>
<tr><th>Category</th><th>Global Analysis</th><th>U.S. Analysis</th></tr>
<tr><td rowspan="4">Physical environment</td><td colspan="2">Mean annual temperature (MAT)</td></tr>
<tr><td colspan="2">Mean annual precipitation (MAP)</td></tr>
<tr><td>Elevation</td><td>N.A.</td></tr>
<tr><td colspan="2">Absolute value of latitude</td></tr>
<tr><td rowspan="5">Social development</td><td rowspan="3">HDI</td><td>Education attainment</td></tr>
<tr><td>Income</td></tr>
<tr><td>Life expectancy</td></tr>
<tr><td colspan="2">Road density</td></tr>
<tr><td colspan="2">Population density</td></tr>
</table>

Table 2. Other factors in each epidemiological analysis. Acronym: HDI - Human Development Index.

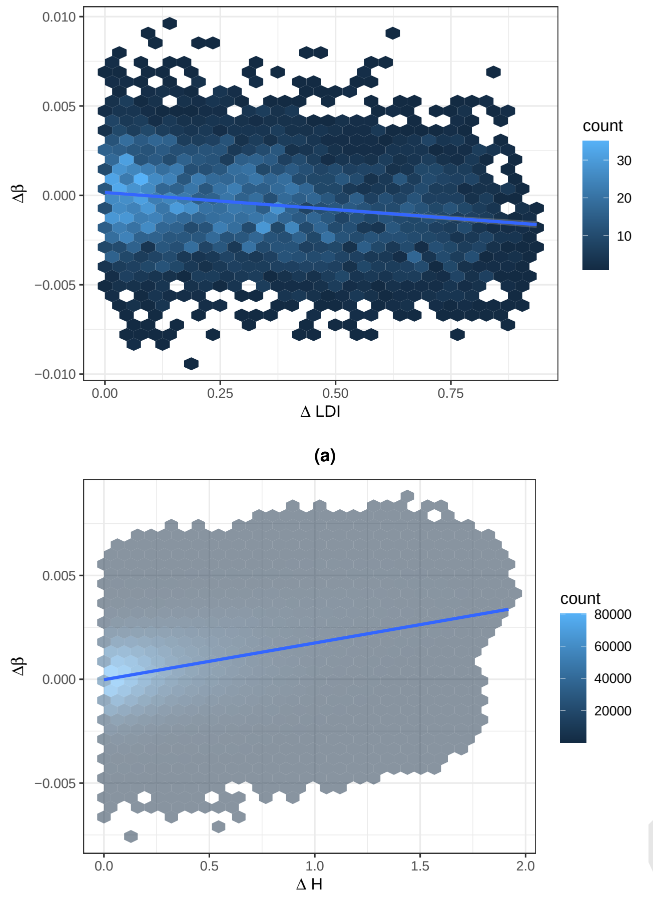
After filtering out countries and U.S. counties with missing data and out-of-distribution data, we had 101 countries in the global analysis and 2739 U.S. counties in the U.S. analysis. Then, we calculated the difference in each metric between country pairs in the global analysis (5050 pairs in total) and county pairs in the U.S. analysis (3749691 pairs). We then conducted a linear regression between the difference in transmission rate ($\Delta\beta$) and the difference in linguistic diversity (ΔLDI or ΔH) along with the differences in other factors listed in Table 2.

Global Analysis. Figure 1a illustrates the relationship between the difference in LDI (ΔLDI) and the difference in disease transmission rate ($\Delta\beta$), along with the linear regression line. The results from the linear regression is shown in Table S1. Most importantly, we found a significant, negative effect of ΔLDI on $\Delta\beta$ (slope estimate =-2.665*10^{-4}, p = 0.021), meaning that there is a negative effect of language diversity on disease transmission, and this effect is not due to confounding factors in the physical environment or social development.

U.S. Analysis. Figure 1b shows the relationship between the difference in linguistic diversity (ΔH) and the difference in disease transmission rate ($\Delta\beta$), along with the linear regression line. The results from the linear regression is shown in Table S2. Contrary to the global analysis, among counties within the United States, we found a significant, positive effect of ΔH on $\Delta\beta$ (slope estimate 1.416*10^{-3}, p $< 2 * 10^{-16}$). This result implies a positive effect of language diversity on disease transmission, and this effect is not due to confounding factors in the physical environment or social development level.

Why would the two analyses manifest opposite patterns? It could be possibly due to the difference in population dynamics between other countries and the USA. Globally, linguistic diversity predominantly corresponds to a constellation of esoteric societies that rarely communicate with each other,

[*] https://coronavirus.jhu.edu/



(a)

(b)

Fig. 1. The difference in the Covid-19 transmission rate ($\Delta\beta$) plotted against language diversity differences. **(a)** Global country pairs (ΔLDI). **(b)** U.S. county pairs (ΔH). The shades of blue represent the number of datapoints in each hexagon. The blue line is the linear regression line.

and in this case, linguistic diversity would serve as a barrier to disease spread. In contrast, in the U.S., due to immigration and the dominance of English as a *lingua franca*, people speaking different languages in fact communicate with each other frequently and live predominantly in densely-populated areas[†], and in this case, linguistic diversity would in fact positively correlate with disease transmission.

In fact, populations speaking different languages roughly correspond to the first societal type in (8) (See Table 1): they are relatively small with a tight social network, and contact between different societies are relatively low. In contrast, in U.S., the situation roughly corresponds to the sixth societal type in Table 1: they are relatively large with a loose social network, but contact between different societies are relatively high, possibly due to the existence of a *lingua franca*.

The two epidemiological analyses suggest that linguistic diversity does have an effect on disease transmission, and the effect seems to be modulated by societal structure. To investigate how different societal structures impact disease spread, in the next section, we simulated the disease spread in the six society types in (8) (Table 1).

[†] https://www.census.gov/library/stories/2024/04/where-do-immigrants-live.html

Simulations. In this analysis, we implemented a modified version of Susceptible, Exposed, Infectious, or Recovered (SEIR) model (22, 23) to simulate disease transmission in different population groups. The model (Figure 2) assumes a community with population N_k consists of four groups: a group that is uninfected and thus susceptible for the disease (henceforth the **susceptible group** S_k), a group that is exposed to the disease but not yet infectious (the **exposed group** E_k), a group that is infectious and showing symptoms (the **symptomatic infectious group** I_i), and a group that is recovered (the **recovered group** R_k) from the disease. Following (24), we assumed a given number of people in each society commute daily between their home society and other societies during a given time interval. In particular, we modeled every day between 6am and 6pm, the number of people travel from society i to society j and back to society i is given by q_{ij}. Disease is then transmitted through contact of the exposed group and the infectious groups between societies. See Equations 6 - 9 and Table 4 in the Materials and Methods section for the equations and parameters.

In our analysis, the six societal types in (8) corresponds to the six scenarios in Table 3. We assume in each scenario, there are four different **societies** with different native languages. Within each societies, there are four different **communities**, which are the basic unit of our analysis. To simulate different societal types, we varied the population in a community, the communication rate between communities within the same society (the intra-society communication rate), and the communication rate between communities in different societies (the inter-society communication rate). We assumed initially in each scenario, 1 person in 1 community is exposed to the disease while everyone else is healthy, and we simulated the disease transmission for 100 days.

Scenario	Population / community	Intra-comm.	Inter-comm.
1	250	$[1, 1.5]$	$[0.1, 0.15]$
2	250	$[1, 1.5]$	$[1, 1.5]$
3	250	$[0.1, 0.15]$	$[0.1, 0.15]$
4	250	$[0.1, 0.15]$	$[1, 1.5]$
5	1000	$[0.1, 0.15]$	$[0.1, 0.15]$
6	1000	$[0.1, 0.15]$	$[1, 1.5]$

Table 3. Summary of parameters for each scenario to simulate the disease transmission dynamics in different types of societies in (8) and Table 1. Acronyms: Intra-comm. - intra-society communication rate; inter-comm. - inter-society communication rate. The communication rate between community pairs is randomly sampled from their corresponding ranges.

The simulation results for Scenarios 1 and 2 are presented in Figure S1.In both cases, the number of susceptible individuals slowly decreases in the first stage, as the number of exposed individuals was low. Then, as more and more individuals are exposed to the disease and later became infectious, the number of susceptible individuals sharply decreases. At the same time, the number of recovered individuals sharply increases as the number of exposed and infectious individuals reach their peak. In the end, everyone in the population is infected with the disease and eventually recovers.

To quantify the disease spread speed in these two scenarios, we measured the **half-time** in each community: the time it

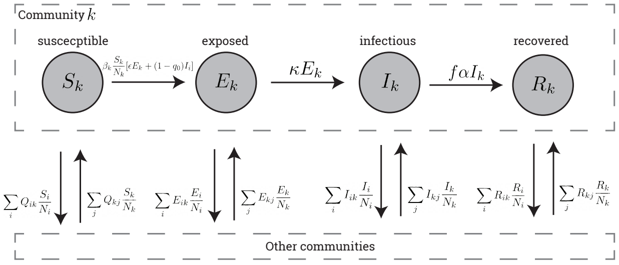

Community k
suscecptible
exposed
infectious
recovered
S_k
E_k
I_k
R_k
$\beta_k \frac{S_k}{N_k}[\epsilon E_k + (1-q_0)I_i]$
κE_k
$f\alpha I_k$
$\sum_i Q_{ik}\frac{S_i}{N_i}$
$\sum_j Q_{kj}\frac{S_k}{N_k}$
$\sum_i E_{ik}\frac{E_i}{N_i}$
$\sum_j E_{kj}\frac{E_k}{N_k}$
$\sum_i I_{ik}\frac{I_i}{N_i}$
$\sum_j I_{kj}\frac{I_k}{N_k}$
$\sum_i R_{ik}\frac{R_i}{N_i}$
$\sum_j R_{kj}\frac{R_k}{N_k}$
Other communities

Fig. 2. Mathematical setup of our modified SEIR model from (24). Each community with population N_k consists of a susceptible group (S_i), an exposed group (E_i), an infectious group (I_k), and a recovered group (R_k). The population change in each group comes from 1) the disease converting a person from one group to another and 2) people traveling from one community to another. Quantities on arrows indicate the rate of change in the population of each group. The definition and value of each parameter can be found in Table 4 in Materials and Methods.

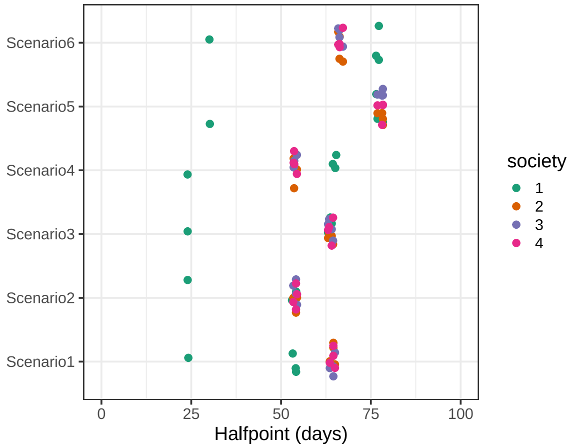


Fig. 3. The time it takes (unit: days, x-axis) for the susceptible population to drop below 50% of the original population for each community in each scenario (y-axis). Communities are represented by color.

takes for the disease to infect half its population. The shorter the half-time is, the faster the disease spreads. The results are shown in Figure 3. In both scenarios, it takes roughly 25 days for half the population in Community 1, where the disease originates, to be exposed or infected with the disease. In Scenario 1, it takes around 53 days for the disease to infect half the population within the same society, but it takes longer for it to do so in other societies (approximately 64 days). In contrast, in Scenario 2, it takes about the same time for the disease to infect half the population in other societies ($\tilde{5}$3 days). The results for other scenarios are presented in Figures S2 and S3.The half points for them are also shown in Figure 3.

A comparison between scenarios 1 and 2, as well as between scenarios 3 and 4, and scenarios 5 and 6, enables to discern the impact of contact on disease spread. As a rule, lower contact rates slow down disease transmission between societies, but has no impact on disease spread within societies. Comparing Scenario 3 with Scenario 5 (and Scenario 4 with Scenario 6), we see that the population size does also have an effect on disease transmission, As a rule, population size slows down the spread of the disease, which makes sense: it takes longer for one exposed individual to transmit disease in a larger population, compared to in a smaller population. In contrast to the effect of contact, intra-society and inter-society transmission are similarly impacted by population size. Finally, comparing Scenario 1 with Scenario 3 (and 2 and 4) allows to determine the effect of social network structure on disease transmission. As a rule, network density has no impact on inter-society transmission, but increase intra-society transmission.

When the three factors are considered together, the slowest transmission rates are observed in Scenario 5 (large isolated human groups forming loose networks), whereas the fastest rates are found in Scenario 2 (small groups forming dense networks and incurring in frequent contacts with other groups).

Discussion

In this paper, we tested an idea raised in Moro (25) that language diversity might act as a protective barrier to slow disease spread. We adopted a two-pronged approach to evaluate such an idea. We first analyzed real-world data, comparing the spread of COVID-19 in regions with opposite linguistic landscapes, but with similar geographical and sociopolitical conditions. In a macro analysis using worldwide data, we observed an inverse correlation between disease transmission and linguistic diversity. Nevertheless, in a micro analysis focusing on a single country (US), we detected a positive correlation between disease transmission and linguistic diversity.

We further conducted a simulation on disease spread on different types of societies that are hypothesized to favor different types of languages and ultimately, diverse degrees of linguistic diversity, as outlined in (8). To test this, we varied population size, and contact rates between and within human groups, which are the three main factors accounting for the societal diversity with an impact on linguistic diversity. Our simulation suggested a variable effect of language diversity on disease spread, as previously showed by real-world data. In general, large population sizes, loose social networks, and reduced contact with other groups slow disease spread. And vice versa. If we regard contact, specifically, as the robustest proxy of linguistic diversity (with low contact rates resulting in high linguistic diversity, as discussed by e.g. (12, 26, 27)), and if we focus on Scenarios 1-2 in our experiment (since the size of human groups has been small for most of our history and we have lived forming dense social networks), our simulation suggests that increasing linguistic diversity could have certainly contributed to slow down the spread of diseases, even if it had no positive effect on disease transmission within groups. This would be in line with the negative correlation we have found between worldwide linguistic diversity and worldwide COVID-19 transmission, certainly under the view that the actual distribution of world languages is a picture (and a consequence) of our deep past.

Our finding of a positive correlation between COVID-19 transmission and linguistic diversity within US makes sense under the view that in this case, the pattern of linguistic diversity is a picture (and a consequence) of our modern world: a multicultural society comprising speakers of many different languages, who are loosely integrated within a very large human group, who move frequently and distantly, and who have extensive contact with people of other cultural groups. Scenario 6 in our simulation is the closest one to that, and in such scenario disease spreads faster.

Equating our simulation to the real world would certainly be a crude simplification. Scenario 1 in the simulation might be a confident proxy of real esoteric societies speaking Type S languages, but present-day linguistic diversity is not properly a constellation of isolated Type S languages, but a a complex admixture of Type S and Type X languages. Likewise, US linguistic diversity is an admixture of many languages of different types, but it has resulted not from prolonged isolation, but from the quick accretion of different languages within a strongly exoteric society. All this certainly explains some mismatches between the simulation and the correlations we found between linguistic diversity and the spread of COVID-19. Hence, while the simulation shows the spread of a disease within pure societal types (in turn impacting on language types and patterns of linguistic diversity), the real world analyses reflect the effect of a complex admixture of societal types (in turn resulting in a complex admixture of language types and patterns of language diversity) on the spread of a disease. Nevertheless, with these limitations in mind, our findings give support to the view that a barrier effect to disease spread could be one of the factors promoting language diversity.

Materials and Methods

Comparative epidemiological analyses.

Global analysis In our global analyses, we compared disease spread in countries differing in language diversity, controlling for other factors that might impact disease spread: physical environment and social development level.

The linguistic diversity in each country is measured by the **Linguistic Diversity Index** (LDI, [28]): a metric reflecting both the number of living languages in a country and the relative proportion of these languages. An LDI of 1 implies that 2 people selected at random in a country would have different mother tongues, whereas an LDI of 0 means that everyone has the same mother tongue. The data is provided in Ethnologue ([1]). In the data, the most linguistically diverse country/region is Papua New Guinea, with an LDI of 0.988, whereas the least linguistically diverse regions are St. Helena and Saint Pierre & Miquelon, each with an LDI of 0, as only English and French are spoken in each territory, respectively. The physical environment in each country is measured by four metrics: **mean annual temperature** (MAT) in 2022, **mean annual precipitation** (MAP) in 2022, **elevation** (with benign climate and rough geography acting, as noted, as triggers of linguistic diversity, but also being proxies of the individual susceptibility to disease transmission), and the absolute value of **latitude** (i.e. 40 degrees north and 40 degrees south are treated in the same way). The MAT and MAP data are obtained from Geospatial Data v0.2 published by the Global Data Lab (DS2 in Table S3). The elevation data is sourced from Wikipedia (DS3 in Table S3). The latitude data for each country is taken from World Bank API, using the R ([29]) package `worldbank`([30]). The social development level in each country is measured by the **Human Development Index** (HDI), a composite index subsuming key dimensions of human development: health (assessed by life expectancy at birth), education (assessed by expected years of schooling for children of school entering age, as well as by mean of years of schooling for adults aged 25 years and more), and well-being (assessed by gross national income per capita). The HDI data for each country is also obtained from the Global Data Lab (DS4 in Table S3). For each country, we also considered a **road density index** (i.e. km of road per 100 km^2 of land area), which we obtained from World Bank (DS5 in Table S3), and **population density**, also from World Bank (DS6 in Table S3). We wish to regard these three indices as confident proxies of the main factors eroding linguistic diversity (road density and average years of schooling, according to Bromham ([31])), but also of the population-level factors impacting on disease transmission.

The speed of disease spread was quantified by the **transmission rate** (β), which was estimated according to the method shown in ([32]), namely:

$$\beta = \frac{-\log\left(1-\frac{I_N}{I}\right)}{T\frac{S}{N}} \qquad [1]$$

In Equation [1], I_N stands for the new infections since the previous sampling, I for the number of individuals that are already infected, T the sample interval, S the number of susceptible individuals, and N the population. We pulled the Covid-19 data from the Johns Hopkins Coronavirus Resource Center, specifically the cumulative number of confirmed cases in each country over time (see DS1 in Table S3 in Supporting Information for more details). The data was updated daily between 2020 and 2023 (hence $T = 1$ day in Equation [1]). I was the current number of cumulative cases. I_N was calculated by subtracting the current number of cumulative cases with the number of cumulative cases logged in the previous day. N was taken from the 2022 population data published by the World Bank (DS7 in Table S3), and S was calculated by subtracting the number of infections from the population. We then averaged the β value throughout the entire 4-year period for each country. The estimated β values are presented in Figure **??**: most countries share a similar range, except five countries (United States, North Korea, Japan, Thailand, and China). To prevent these outlier βs from skewing our results, we removed these five countries from the analysis.

We gathered all the aforementioned data together and removed countries with missing data. This left us with 101 countries in the dataset. To ensure comparability across metrics, we normalized all the physical environment and social development metrics by subtracting each metric's mean from its values and dividing by its standard deviation. Then, we computed the difference in each metric for every country pair, resulting in 5050 country pairs[‡].

We did a linear regression between the difference in LDI (ΔLDI) and the difference in COVID-19 transmission rate $\Delta\beta$. To control for the influence of physical environment and social development, we also entered these metrics as predictors. In R syntax, the regression is:

$$\begin{aligned}\Delta\beta \sim \Delta\text{LDI} + \Delta\text{MAP} + \Delta\text{MAT} + \Delta\text{elev.}+\\ \Delta\text{lat.} + \Delta\text{HDI} + \Delta\text{popu.density} + \Delta\text{road.density}\end{aligned} \qquad [2]$$

If the intuitions in Moro ([25]) are true, we should expect a negative relation between ΔLDI and $\Delta\beta$ in Equation [2].

U.S. Analysis In our fine-grained analyses, following a similar approach, we compared the spread dynamics of COVID-19 in US counties with various language diversity. We also control for other factors that might impact disease spread: physical environment and social development level.

The speed of disease spread was calculated in a similar way to the previous analysis, according to Equation [1], except we calculated a β for each county. A histogram of β values calculated for each U.S. county is presented in Figure S6.Most of the values range between 0.06 and 0.08. To avoid extremely low β values skewing the analysis, we excluded counties with β lower than 0.06.

The language data was taken from the American Community Survey (ACS) published by the U.S. Census Bureau (DS8 in Table S3). The data is presented in 5 categories: English only, Spanish, other Indo-European languages, Asian and Pacific-Islander languages, other languages. As an approximation, we assume each person falls into one and only one of these five categories. The linguistic diversity in each US county was then measured by the **entropy** of distribution of categories in each county (Equation [3]). Entropy is minimal when everyone in a county falls into one category (i.e. they all only speak Spanish)

[‡] Each country pair appeared twice in our dataset (e.g. Afghanistan - Albania and Albania - Afghanistan) and all the differences in metrics were the same value but with different signs. We only kept the country pairs that resulted in a positive difference in LDI.

and is maximal when people in a county are equally distributed across these five categories.

$$H = -\sum_{i=1}^{5} p(k)\,\log p(k) \qquad [3]$$

The physical environment in each county was measured by three metrics: MAT in 2022, MAP in 2022, and latitude. MAT and MAP are obtained from the National Oceanic and Atmospheric Administration (DS9 and DS10 in Table S3). Latitudes of each county is taken from the US Census Bureau (DS11). The social environment in each county was approximated by the following metrics: development level, population density, and road density. However, due to lack of US county level HDI data, we replaced HDI with three separate metrics: education profile, income profile, and life expectancy. The education profile data was taken from ACS educational attainment data published by the US Census Bureau (DS12). The data is grouped into 7 categories: less than 9th grade, 9th to 12th grade (no diploma), hgh school graduate, some college (no degree), associate's degree, bachelor's degree, and graduate or professional degree. Each county has its own distribution of populations into these 7 categories. The **educational dissimilarity** of two counties $d_e(c_i, c_j)$ was calculated by subtracting the **cosine similarity** of these two counties' education profile from 1. If two counties have the exact same educational attainment profile, the cosine similarity will be 1, and therefore their dissimilarity will be 0. Similarity, we took the income profile data from the ACS dataset "income in the past 12 months", published by the US Census Bureau (DS13). The data is grouped into 10 categories, ranging from "less than \$10,000" to "\$200,000 or more". Similar as before, we calculated the **income dissimilarity** of two counties $d_i(c_i, c_j)$ by subtracting the cosine similarity of these two counties' income profile from 1. The life expectancy data for each US county was drawn from US Centers of Disease Control and Prevention (CDC; See DS14). The population density data was calculated by dividing the population of a county, taken from US Census Bureau (DS15), divided by the area of the county (in squared miles, data also taken from US Census Bureau; See DS11). The road density data was calculated in a similar way, by dividing the length of roads in a county (taken from a dataset published by US Census Bureau and compiled by Erin Davis; See DS16), divided by the area of the county.

Similar to the global analysis, we gathered all the aforementioned data together and removed US counties with missing data. We normalized all the physical and social environment metrics by the following formula (Equation 4, so that every metric ranges between 0 and 1. We then computed the difference in metrics between each county pair, along with each pair's educational and income similarity, resulting in 3749691 pairs.

$$r_i = \frac{i - \min i}{\max i - \min i} \qquad [4]$$

We did a linear regression between the linguistic diversity ΔH and $\Delta\beta$. We also entered MAT, MAP, latitude, educational profile dissimilarity, income profile dissimilarity, life expectancy, population density, and road density as predictors. In R syntax, the regression is:

$$\begin{aligned}&\Delta\beta \sim \Delta H + \Delta\text{MAP} + \Delta\text{MAT} + \Delta\text{lat.} + \\ &\quad d_e + d_i + \Delta\text{life.expec} + \Delta\text{popu.density} + \Delta\text{road.density} \qquad [5]\end{aligned}$$

Similar to the global analysis, if the intuitions in Moro (25) are true, we should expect a negative relation between ΔH and $\Delta\beta$ in Equation 5.

Simulations.

Mathematical setup The model used in this study is a modified version of the one developed in Lee & Jung (24). Their model is modified from the classic Susceptible, Infectious, or Recovered (SIR) model (22) in order to predict pathogen transmission between different communities. In their model, each **society** (N_k) consists of five **groups**: a group that is uninfected and thus susceptible for the disease (henceforth the **susceptible group** S_k), a group that is exposed to the disease but not yet infectious (the **exposed group** E_i), a group that is infectious but not showing symptoms (the **asymptomatic infectious group** A_k), a group that is infectious and showing symptoms (the **symptomatic infectious group** I_k), and a group that is recovered (the **recovered group** R_k) from the disease. A given number of people in each society commute daily between their home society and other societies during a given time interval. In particular, they modeled every day between 6am and 6pm, the number of people travel from society i to society j and back to society i is given by q_{ij}. Disease is then transmitted through contact of the exposed group and the infectious groups between societies.

In our model, the linguistic landscape consists of four different **societies**, which can be more or less tightly interconnected. In turn, each society consists of four **communities**, which can also be more or less tightly interconnected. These communities are the basic unit of analysis in our study. We also assume people being exposed to the pathogen do not isolate. To simplify the implementation, we removed the asymptomatic group from each community and assumed every infectious individual is symptomatic. The modified spatiotemporal model from Lee & Jung (24) is given by the following equations, in the k^{th} community (in our study k =1, 2, ..., 16):

$$\frac{dS_k}{dt} = \sum_i Q_{ik}\frac{S_i}{N_i} - \sum_j Q_{kj}\frac{S_k}{N_k} - \beta_k\frac{S_k}{N_k}[\epsilon E_k + (1-q_0)I_k] \qquad [6]$$

$$\begin{aligned}\frac{dE_k}{dt} &= \sum_i Q_{ik}\frac{E_i}{N_i} - \sum_j Q_{kj}\frac{E_k}{N_k} + \\ &\quad \beta_k\frac{S_k}{N_k}[\epsilon E_k + (1-q_0)I_k] - \kappa E_k \qquad [7]\end{aligned}$$

$$\frac{dI_k}{dt} = \sum_i Q_{ik}\frac{I_i}{N_i} - \sum_j Q_{kj}\frac{I_k}{N_k} + \kappa E_k - \alpha I_k \qquad [8]$$

$$\frac{dR_k}{dt} = \sum_i Q_{ik}\frac{R_i}{N_i} - \sum_j Q_{kj}\frac{R_k}{N_k} + f\alpha I_k \qquad [9]$$

Following Lee & Jung (24), the commuting function for people living in community i commuting to community j, $Q_{ij}(t)$ (unit: days), is given by:

$$Q_{ij} = \begin{cases} \frac{\pi q_{ij}}{T}\sin\left(\frac{2\pi}{T} \mod (t - t_s, \Delta t)\right), & t_s \le \mod (t, \Delta t) \le t_e \\ 0, & \text{otherwise} \end{cases} \qquad [10]$$

where q_{ij} is the number of persons commuting from community i to community j ($q_{ii} = 0$). The definitions and values of other parameters are taken from Lee & Jung (24) and are listed in Table 4.

Table 4. Parameters used in our study. The values are predominantly taken from Lee & Jung (24).

Symbol	Definition	Value
β_k	Pathogen transmission rate per capita in community k	0.68
ϵ	Infectivity reduction rate for exposed groups	0
q_0	Contact reduction rate via isolation	0 (assume no isolation)
α	Recovery rate for infectious groups	1/6
κ	Progression rate from exposed to infectious	1/1.9
f	Survivability rate	1
t_s	Commute start time (unit: day)	0.25 (e.g. 06:00)
t_e	Commute ending time (unit: day)	0.75 (e.g. 18:00)
T	Commute duration (unit: day)	0.5 (i.e. $t_e - t_s$)
Δt	Unit time	1

All the simulations are conducted in MATLAB (Version R2024b, Mathworks, Inc., Natick, MA) with the ode45 function, which solves ordinary differential equation systems based on Runge-Kutta (4,5) formula ([33], [34]). The relative tolerance in our simulation is set to be 10^{-6}.

Scenarios We simulated disease transmission under 6 different scenarios by entering different commute rate between communities, different commute rates between societies, and different community population sizes (Table 3). We set the population of a small community to be 250 and a large community to be 1000. We set the commute rate between communities within a society with a tight network to be randomly sampled between 1 and 1.5. This corresponds to that for a community of the size of New York City (8 million people), up to 720000 people travel to a different community daily and return to their home community at night. Regarding the rate between communities within a society with a loose network, we scaled down it by a factor of 10. We choose the factor 10 by estimating the difference in number of travelers each year between three major cities in Europe: Madrid, Barcelona, and Lisbon. These 3 cities are similar in size in their metropolitan area and comparable in economic status, with an important difference: Madrid and Barcelona share an official language (Spanish), but Madrid and Lisbon do not (Spanish and Portuguese). Hence, we obtain the factor by estimating the relative traffic between Madrid and Lisbon (low contact between societies) and Madrid and Barcelona (high contact between societies). The main means of traffic between Madrid and Barcelona is high-speed rail, and in 2023, the number of passengers in the line is approximately 13.8 million[§]. The main means of traffic between Madrid and Lisbon is by plane, and in 2022, the number of passengers in the Madrid-Lisbon route is approximately 1.5 million (DS17 in Table 3). We also simulate the disease transmission dynamics in the other four types of societies in ([8]). Similarly, we set the commute rate between communities belonging to different societies to be between 1 and 1.5 as a high contact rate, and between 0.1 and 0.15 as a low contact rate.

1. DM Eberhard, GF Simons, CD Fennig, eds., *Ethnologue: Languages of the World.* (SIL International, Dallas, Texas), Twenty-seventh edition edition, (2024) Online version: http://www.ethnologue.com/.
2. G Lupyan, R Dale, Why are there different languages? the role of adaptation in linguistic diversity. *Trends cognitive sciences* **20**, 649–660 (2016).
3. C Everett, DE Blasi, SG Roberts, Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proc. Natl. Acad. Sci. United States Am.* **112**, 1322–1327 (2015).
4. SG Roberts, Robust, causal, and incremental approaches to investigating linguistic adaptation. *Front. psychology* **9**, 166 (2018).
5. T Wang, S Wichmann, Q Xia, Q Ran, Temperature shapes language sonority: Revalidation from a large dataset. *PNAS nexus* **2**, pgad384 (2023).
6. K Sinnemäki, Complexity in core argument marking and population size in *Language complexity as an evolving variable*, eds. G Sampson, D Gil, P Trudgill. (Oxford University Press, Oxford), pp. 126–140 (2009).
7. G Lupyan, R Dale, Language structure is partly determined by social structure. *PloS one* **5**, e8559 (2010).
8. P Trudgill, *Sociolinguistic Typology: Social Determinants of Linguistic Complexity.* (Oxford University Press, Oxford), (2011).
9. D Nettle, Social scale and structural complexity in human languages. *Philos. transactions Royal Soc. London. Ser. B, Biol. sciences* **367**, 1829–1836 (2012).
10. D Gil, Tense-aspect-mood marking, language-family size and the evolution of predication. *Philos. transactions Royal Soc. London. Ser. B, Biol. sciences* **376**, 20200194 (2021).
11. S Chen, et al., Linguistic correlates of societal variation: A quantitative analysis. *PloS one* **19**, e0300838 (2024).
12. D Nettle, Explaining global patterns of language diversity. *J. anthropological archaeology* **17**, 354–374 (1998).
13. G Sankoff, Linguistic outcomes of language contact in *The handbook of language variation and change*, eds. JK Chambers, P Trudgill, N Schilling-Estes. (Blackwell, Hoboken), pp. 638–668 (2004).
14. S Romaine, Contact and language death in *The handbook of language contact*, ed. R Hickey. (Blackwell, Hoboken), pp. 320–339 (2010).
15. R Van Gijn, et al., The social lives of isolates (and small language families): the case of the northwest amazon. *Interface focus* **13**, 20220054 (2022).
16. P Pooladvand, JR Kendal, MM Tanaka, How cultural innovations trigger the emergence of new pathogens. *Proc. Natl. Acad. Sci.* **121**, e2322882121 (2024).
17. MJ Toole, RJ Waldman, Prevention of excess mortality in refugee and displaced populations in developing countries. *Jama* **263**, 3296–3302 (1990).
18. P Farmer, *Infections and inequalities: The modern plagues.* (Univ of California Press), (2001).
19. S Metelmann, et al., Impact of climatic, demographic and disease control factors on the transmission dynamics of covid-19 in large cities worldwide. *One Heal.* **12**, 100221 (2021).
20. Z Gong, et al., Natural and socio-environmental factors in the transmission of covid-19: a comprehensive analysis of epidemiology and mechanisms. *BMC Public Heal.* **24**, 2196 (2024).
21. A Wray, GW Grace, The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua* **117**, 543–578 (2007).
22. R Ross, An application of the theory of probabilities to the study of a priori pathometry.—part i. *Proc. Royal Soc. London. Ser. A, Containing papers a mathematical physical character* **92**, 204–230 (1916).
23. ON Bjørnstad, K Shea, M Krzywinski, N Altman, The seirs model for infectious disease dynamics. *Nat. Methods* **17**, 557–558 (2020).
24. J Lee, E Jung, A spatial–temporal transmission model and early intervention policies of 2009 a/h1n1 influenza in south korea. *J. theoretical biology* **380**, 60–73 (2015).
25. A Moro, *Impossible languages.* (MIT Press), (2016).
26. H Skirgård, et al., Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Sci. Adv.* **9** (2023).
27. J Nichols, *Linguistic Diversity in Space and Time.* (University of Chicago Press), (1992).
28. JH Greenberg, The measurement of linguistic diversity. *Language* **32**, 109 (1956).
29. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria), (2024).
30. M Mücke, *worldbank: Client for World Banks's 'Indicators' and 'Poverty and Inequality Platform (PIP)' APIs*, (2025) https://cran.r-project.org/web/packages/worldbank/index.html.
31. L Bromham, et al., Global predictors of language endangerment and the future of linguistic diversity. *Nat. Ecol. amp; Evol.* **6**, 163–173 (2021).
32. C Kirkeby, T Halasa, M Gussmann, N Toft, K Græsbøll, Methods for estimating disease transmission rates: Evaluating the precision of poisson regression and two novel methods. *Sci. reports* **7**, 9496 (2017).
33. JR Dormand, PJ Prince, A family of embedded runge-kutta formulae. *J. computational applied mathematics* **6**, 19–26 (1980).
34. LF Shampine, MW Reichelt, The matlab ode suite. *SIAM journal on scientific computing* **18**, 1–22 (1997).

[§] source: https://data.cnmc.es/transporte-y-postal/transporte/conjuntos-de-datos/indicadores-trimestrales-del-transporte-0