UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS (ICB)
PROGRAMA DE PÓS-GRADUAÇÃO EM ECOLOGIA E EVOLUÇÃO

CHRISTIELLY MENDONÇA BORGES

# **Teoria e métodos ecológicos e evolutivos aplicados a dados humanos:** de diversidade biocultural à propagação de doenças

GOIÂNIA

2022

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS

# TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

# E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

**1. Identificação do material bibliográfico**

[ ] Dissertação [ X ] Tese

**2. Nome completo do autor**

Christielly Mendonça Borges

**3. Título do trabalho**

Teoria e métodos ecológicos e evolutivos aplicados a dados humanos: de diversidade biocultural à propagação de doenças

**4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)**

Concorda com a liberação total do documento [ X ] SIM [ ] NÃO[1]

**[1]** Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
**a)** consulta ao(à) autor(a) e ao(à) orientador(a);
**b)** novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

**Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.**

Documento assinado eletronicamente por **CHRISTIELLY MENDONÇA BORGES**, **Discente**, em 12/05/2022, às 11:43, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Thiago Fernando Lopes Valle De Britto Rangel**, **Professor Titular-Livre Magistério Superior**, em 12/05/2022, às 13:20, conforme horário oficial de Brasília, com

fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2899820** e o código CRC **96E88EDF**.

**Referência:** Processo nº 23070.020933/2022-17 SEI nº 2899820

CHRISTIELLY MENDONÇA BORGES

# **Teoria e métodos ecológicos e evolutivos aplicados a dados humanos:** de diversidade biocultural à propagação de doenças

Tese apresentada ao Programa de Pós-Graduação em Ecologia e Evolução, do Instituto de Ciências Biológicas, da Universidade Federal de Goiás (UFG), como requisito para a obtenção do título de Doutora em Ecologia e Evolução.

Área de concentração: Ecologia e Evolução

Linha de pesquisa: Macroecologia e Ecologia Evolutiva

Orientador: Dr. Thiago Fernando Lopes Valle de Britto Rangel

GOIÂNIA

2022

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Borges, Christielly Mendonça
Teoria e métodos ecológicos e evolutivos aplicados a dados humanos [manuscrito] : de diversidade biocultural à propagação de doenças / Christielly Mendonça Borges. - 2022.
181 f.: il.

Orientador: Prof. Dr. Thiago Fernando Lopes Valle de Britto Rangel.
Tese (Doutorado) - Universidade Federal de Goiás, Instituto de Ciências Biológicas (ICB), Programa de Pós-Graduação em Ecologia e Evolução, Goiânia, 2022.
Bibliografia. Anexos. Apêndice.
Inclui mapas, gráfico, tabelas.

1. dialetos brasileiros. 2. diversidade biocultural. 3. evolução de línguas. 4. macroecologia humana. 5. saúde pública. I. Rangel, Thiago Fernando Lopes Valle de Britto, orient. II. Título.

CDU 574

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE CIÊNCIAS BIOLÓGICAS

**ATA DE DEFESA DE TESE**

Ata Nº **115** da sessão de Defesa de Tese de **Christielly Mendonça Borges** que confere o título de **Doutora em Ecologia e Evolução**, na área de concentração em **Ecologia e Evolução.**

Aos **vinte e cinco dias do mês de abril do ano de dois mil e vinte e dois (25/04/2022),** a partir das **8h30min**, por **videoconferência, seguindo portaria CAPES no. 36 de 16 de março de 2020 e recomendação da UFG**, realizou-se a sessão pública de Defesa de Tese intitulada **"Teoria e métodos ecológicos aplicados à dados humanos: de diversidade biocultural à propagação de doenças"**. Os trabalhos foram instalados pelo Orientador, **Professor Doutor Thiago Fernando Lopes Valle de Britto Rangel** (DECOL/UFG); com a participação dos demais membros da Banca Examinadora: **Dr. Marco Túlio Pacheco Coelho** (Swiss Federal Institute for Forest, Snow and Landscape Research - WSL) , membro titular externo; **Professor Doutor Bruno Vilela de Moraes e Silva** (IBio/UFBA), membro titular externo; **Professor Doutor Thiago Costa Chacon (**UnB), membro titular externo, **Professor Doutor José Alexandre Felizola Diniz-Filho** (DECOL/UFG), membro titular interno. Durante a argüição os membros da banca **fizeram** sugestão de alteração do título do **trabalho, conforme explicitado abaixo**. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Tese tendo sido a candidata **aprovada** pelos seus membros. Proclamados os resultados pelo **Professor Doutor Thiago Fernando Lopes Valle de Britto Rangel**, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos **vinte e cinco dias do mês de abril do ano de dois mil e vinte e dois (25/04/2022)**.

TÍTULO SUGERIDO PELA BANCA

Teoria e métodos ecológicos e evolutivos aplicados a dados humanos: de diversidade biocultural à propagação de doenças

Documento assinado eletronicamente por **Thiago Fernando Lopes Valle De Britto Rangel**, **Professor Titular-Livre Magistério Superior**, em 26/04/2022, às 15:59, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **José Alexandre Felizola Diniz Filho**, **Professora do Magistério Superior**, em 26/04/2022, às 16:46, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **MARCO TÚLIO PACHECO COELHO**, **Usuário Externo**, em 27/04/2022, às 08:20, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Bruno Vilela de Moraes e Silva**, **Usuário Externo**, em 27/04/2022, às 10:31, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Thiago Costa Chacon**, **Usuário Externo**, em 29/04/2022, às 14:45, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2850201** e o código CRC **B31171E4**.

**Referência:** Processo nº 23070.020933/2022-17 SEI nº 2850201

*The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom*

Isaac Asimov

*Eu não nasci rodeada de livros e, sim, rodeada de palavras.*

Conceição Evaristo

# AGRADECIMENTOS

Muitos eventos catastróficos aconteceram desde que eu ingressei no doutorado em 2018. O levante da extrema-direita no Brasil e no mundo, a pandemia de COVID-19 e consequentemente uma quarentena de dois anos, o desmonte da ciência brasileira, o aumento significativo do desmatamento no Brasil, o aumento imoral de garimpos ilegais em terras indígenas, e por fim, o início de uma guerra com potencial de expansão mundial. Esses são os mais calamitosos, são os que me tiraram o sono e até dias de trabalho. Confesso que passei tardes assistindo a CPI da COVID-19, ou me inteirando dos assuntos supracitados. Por vezes, trabalhar nesta tese me parecia ignóbil e fútil.

No entanto, trabalhar também era meu único refúgio desses problemas. Talvez seja irônico para alguns colegas, mas trabalhar me ajudava a lidar com a ansiedade que eu sentia. E assim construí esta tese. Entre momentos de extrema preocupação e também de alienação ao resto do mundo, tijolo por tijolo. Felizmente, não caminhei sozinha nesta jornada. Se chegou minha hora de brilhar, muitas pessoas vão brilhar comigo, por isso aqui faço questão de agradecer à todas.

À minha mãe, **Rosângela**, minha **Tia Regina** e meu irmão **Rafael**, minha família amada que sempre me deu apoio emocional e financeiro, que sempre me incentivou a correr atrás dos meus sonhos. Ao **Mario Joaquim**, pelo apoio e companheirismo incondicionais durante esse período. Acrescento tranquilamente que sem vocês quatro não existiria a possibilidade de um doutorado.

Ao meu orientador, **Dr. Thiago Rangel**, por ter topado me orientar, por permitir que eu fosse livre nas escolhas dos meus temas, por me mostrar caminhos sem me carregar por eles e por todas as conversas que já tivemos. Suas sugestões e comentários sempre melhoraram meu trabalho, e várias vezes iluminou uma porta que antes eu não enxergava.

Aos professores **Dr. José Alexandre Diniz-Filho** e **Dr. Luis Maurício Bini** por sempre estarem presentes nesses quatro anos, inclusive no auge da pandemia, sempre disponíveis, dando apoio e sempre com uma palavra amiga. Vocês são o coração do nosso PPG e para mim é uma grande honra ter sido sua aluna.

Aos meus colaboradores **Dr. Zander Vilaça**, **Guilherme Ferreira**, **Dr. Thiago Chacon**, **Dr. Michael Gavin**, **Dr. Marco Túlio**, e novamente os Drs. **José Alexandre Diniz-Filho** e **Thiago Rangel**, pois sem vocês, não teria concluído um capítulo sequer.

Aos amigos **Dr. Marco Túlio Coelho**, **Dra. Elisa Barreto** e **Dra. Rafaela Granzotti** por sempre se oferecem para um *friendly reading*, e sempre estarem disponíveis para discutir ideias, métodos e acalmar minhas dúvidas.

Aos amigos **Matheus Nunes**, **Igor Bione**, **Desirée Meireles, Luisa Latorre**, **Gisele Santos**, **Joedison Rocha**, **Erivelton Nascimento**, **Rejane Santos**, **Daisy Jorge** e **Carlos Libório** por incontáveis almoços, trocas acadêmicas, conversas e jogos de queimada. Aos amigos de Salvador, **Vinicius Santos**, **Carlos Calderón**, **Edson Junior**, **Thais Dória**, **Pietro Noga** e **Myrla Rocha** pela amizade e parceria que se manteve mesmo à distância.

Aos colegas do **LETS** pela troca de conhecimentos diária, especialmente aos pós-doutores **Dr. André Menegotto**, **Dr. Lucas Jardim**, **Dr. Renato Dala-Corte** e **Dra. Alessandra Bertassoni**, que sempre acolheram minhas dúvidas, me ajudaram com problemas metodológicos e aceitaram parcerias.

Aos meus queridos amigos não acadêmicos, **Guilherme Tai**, **Diego Neri**, **Artur Ferreira** e **Ritha Paixão**, por sempre segurar minha barra, minha onda, e minha conversinha acadêmica.

Aos meus professores de graduação, mestrado e doutorado, em especial aos Drs. **Blandina Viana**, **Daniel Brito**, **Francisco Barros**, **Joaquin Hortal**, **Marcus Cianciaruso**, **Mário Almeida-Neto**, **Natan Maciel**, **Paulo de Marco**, **Pedro Rocha**, **Priscilla Carvalho**,

**Rafael Loyola**, **Ricardo Dobrovolski**, **Rosane Collevatti** e **Viviane Ferro**, por terem me ensinado praticamente tudo que eu sei sobre ecologia e por sempre terem comprometimento com a ciência e a formação de pessoas. Serei eternamente grata por todas as lições.

À **Universidade Federal de Goiás**, pela estrutura física e incentivos como alimentação barata, acesso à periódicos científicos, posto médico e cursos de exercício físico gratuitos. Ao **Instituto Nacional de Ciência e Tecnologia em Ecologia, Evolução e Conservação da Biodiversidade** (INCT-EECBio), pela oficina de Macroecologia Humana e por disponibilizar acesso à cluster. Aos Drs. **Rhewter Nunes** e **Mariane Brom** por todo auxílio com a cluster. Finalmente, à **CAPES**, pois o presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

**SUMÁRIO**

# RESUMO

Tradicionalmente, *Homo sapiens* tem sido objeto de estudo exclusivo das humanidades. A resistência de cientistas da natureza em estudar humanos do ponto de vista evolutivo-ecológico é facilmente explicada pelos desdobramentos do movimento eugênico do século 20. A partir de avanços científicos sobre a baixa variabilidade genética entre populações humanas e os padrões espaciais da diversidade de línguas, criou-se a ideia de diversidade humana não-biológica, onde humanos formam inúmeros grupos culturais com padrões globais espaciais e demográficos complexos. Nesta tese aplicamos teorias e métodos evolutivos-ecológicos à dados humanos, focando em diferentes aspectos da diversidade linguística, seguindo uma abordagem macroecológica e também analisando as dinâmicas de propagação do coronavírus da síndrome respiratória aguda grave 2 (SARS-CoV-2). No Capítulo 1, revisamos como a evolução das línguas e das espécies estão muito além de meras analogias e acumularam uma crescente literatura que corroboram esses paralelismos em ambas as escalas micro e macroevolutivas. No Capítulo 2 criamos um teste dialetal para coletar dados linguísticos do português brasileiro (PB), e assim complementar dados existentes, preencher lacunas e posteriormente demarcar os diferentes dialetos do PB, reconstruir o histórico de imigração no Brasil, e pesquisar a evolução do PB. No Capítulo 3 investigamos os mecanismos responsáveis pela diversidade linguística na região Neotropical (México, Américas Central e Sul). Criamos um modelo mecanístico espacialmente explícito que incorpora altitude, recursos hídricos, precipitação e tamanho do grupo populacional como mecanismos capazes de predizer a diversidade linguística pré-colombiana observada no continente. No Capítulo 4, usamos um modelo epidemiológico SIR (Suscetíveis, Infectados, Removidos) para avaliar a efetividade das políticas públicas do Estado de Goiás em conter a propagação da COVID-19 em seu estágio inicial, entre março e maio de 2020. Em todos os capítulos aplicamos com sucesso teorias e métodos ecológicos à dados oriundos de humanos, seja a língua que falam ou o vírus que os infectam. Portanto, demonstramos como os métodos e teorias desenvolvidos nas áreas biológicas podem ser aplicados para avançar conhecimentos das áreas de humanidades, principalmente na linguística e na administração pública. Nesse sentido, demonstramos a importância e eficácia de estudos multidisciplinares, principalmente para um objeto de estudo tão complexo quanto o *Homo sapiens*.

*Palavras-chave:* COVID-19, dialetos brasileiros, diversidade biocultural, ecologia de populações, evolução de línguas, macroecologia humana, modelo mecanístico, saúde pública

## ABSTRACT

Traditionally, *Homo sapiens* have been an exclusive subject of study of the humanities. The resistance of natural scientists to study humans from an eco-evolutionary point of view is easily explained by the unfolding of the 20th century eugenics movement. Starting from the scientific advances on the low genetic variability between human populations and the spatial patterns of language diversity, the idea of a non-biological human diversity emerged, where humans form numerous cultural groups with complex global spatial and demographic patterns. In this thesis, we apply eco-evolutionary theories and methods to human data, focusing on different aspects of linguistic diversity, following a macroecological approach and also analyzing the propagation dynamics of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). In Chapter 1, we reviewed how the evolution of languages and species goes far beyond mere analogies and has accumulated a growing literature that supports these parallels at both micro and macroevolutionary scales. In Chapter 2 we created a dialect quiz to collect linguistic data from Brazilian Portuguese (BP), and thus complement existing data, fill in gaps and later demarcate the different BP dialects, reconstruct the immigration history in Brazil, and research the evolution of BP. In Chapter 3 we investigate the mechanisms responsible for linguistic diversity in the Neotropical region (Mexico, Central and South America). We created a spatially explicit mechanistic model that incorporates altitude, water resources, precipitation and population group size as mechanisms capable of predicting the pre-Columbian linguistic diversity observed on the continent. In Chapter 4, we used a SIR (Susceptible, Infected, Removed) epidemiological model to assess the effectiveness of public policies of the state of Goiás in containing the spread of COVID-19 in its initial stage, between March and May 2020. In all chapters we successfully apply ecological theories and methods to data originated from humans, whether it's the language they speak or the virus that infects them. Therefore, we demonstrate how the methods and theories developed in biological disciplines can be applied to advance knowledge in the humanities, especially in linguistics and public administration. In this sense, we demonstrate the importance and effectiveness of multidisciplinary studies, especially for an object of study as complex as *Homo sapiens*.

*Keywords:* COVID-19, biocultural diversity, brazilian dialects, human macroecology, language evolution, mechanistic model, population ecology, public health

# INTRODUÇÃO GERAL

Tradicionalmente, a espécie *Homo sapiens* é objeto de estudos ecológicos de forma indireta, através de como impactam outras espécies, ecossistemas, a atmosfera, o solo e o clima. Ironicamente, avaliamos como humanos impactam o ambiente, mas raramente analisamos como o ambiente impacta os humanos. Assim, os cientistas da natureza estudam o mundo focando em todas as espécies exceto naquela que é notoriamente a mais bem sucedida, que colonizou todos os continentes e modifica o ambiente de inúmeras formas.

É inegável que ainda hoje há resistência em estudar humanos do ponto de vista evolutivo-ecológico, fato facilmente explicado pelos desdobramentos do movimento eugênico do século 20. Estatísticos como Francis Galton (fundador da eugenia), Karl Pearson e Ronald Fisher eram engajados em identificar diferenças raciais, dando respaldo científico à delírios racistas e coloniais, e posteriormente à fascistas e nazistas (Clayton, 2020). Em 1975, E. O. Wilson publicaria seu controverso "Sociobiology", onde o comportamento humano é explicado exclusivamente por mecanismos genéticos. A ideia de determinismo biológico foi fortemente criticada por importantes cientistas de seu tempo (Gould & Lewontin, 1979), sendo inclusive agrupada com eugenia por Gould em seu livro "The Mismeasure of Man". O próprio Wilson foi retirado de seu armário racista após sua morte em 2022 (Diniz-Filho, 2022; McLemore, 2022).

O risco de respaldar conceitos de "pureza racial", "determinismo biológico" e "supremacia branca" afasta cientistas naturais em ter *Homo sapiens* como objeto de estudo, deixando a espécie para os holofotes das humanidades. A partir da década de 1990, observamos uma mudança nessa postura, principalmente porque técnicas moleculares permitiram demonstrar a baixa variabilidade genética entre populações humanas (Templeton, 1998). Por exemplo, toda humanidade varia menos geneticamente do que uma

população selvagem de chimpanzés (Kaessmann et al., 1999), enterrando de vez qualquer noção científica de raça ou de subespécies para humanos.

Ainda na mesma década, foi publicado pela primeira vez um atlas com o mapeamento de todas as línguas do mundo (Moseley & Asher, 1994). A congruência dos padrões linguísticos com os padrões já conhecidos para a biodiversidade, ambos apresentando maior diversidade nos trópicos, não passou despercebida. Logo, cientistas começaram a buscar explicações empíricas para a distribuição assimétrica das línguas, calculando correlações entre variáveis abióticas e diversidades linguísticas na América do Norte e África Ocidental (Mace & Pagel, 1995; Nettle, 1996).

Dessa forma, criou-se uma ideia de diversidade humana não-biológica, muito mais palatável. Se por um lado os humanos constituem um grupo genético praticamente homogêneo, por outro possuem inúmeros grupos culturais que formam padrões globais espaciais e demográficos complexos (Mirazón Lahr, 2016). Uma forma de quantificar esses diferentes grupos culturais é através da diversidade linguística (Pagel & Mace, 2004).

Existem hoje cerca de 7.000 línguas sobreviventes que se diferem em fonologia, morfologia e sintaxe e que estão aninhadas em padrões de descendência (Gavin et al., 2013). Essa hierarquia permite que se estude diversidade linguística em diferentes níveis taxonômicos, como número de línguas (riqueza linguística), número de famílias de línguas (diversidade filogenética linguística) (Gavin et al., 2013; Manne, 2003) e dialetos regionais (Gonçalves & Sánchez, 2014; Honkola et al., 2018), o que torna a diversidade linguística uma variável ideal para estudos macroecológicos sobre diversidade humana (Maffi, 2005).

De fato, na década de 2010 houve o rápido surgimento da área “Macroecologia Humana”, definida inicialmente como “o estudo das interações humano-ambiente em escalas espaciais e temporais, ligando interações de pequena escala com padrões emergentes de grande escala e seus processos subjacentes” (Burnside et al., 2012). Além

disso, esses autores defenderam uma abordagem macroecológica para interpretar padrões e sugerir possíveis mecanismos para, além da diversidade cultural e linguística, a ecologia de forrageamento humano, história de vida, estrutura populacional, ecologia de doenças, uso de espaço e sistemas industriais e urbanos (Burnside et al., 2012).

Nesta tese aplicamos teorias e métodos evolutivos-ecológicos à dados humanos, focando em diferentes aspectos da diversidade linguística, seguindo uma abordagem macroecológica e também analisando as dinâmicas de propagação do coronavírus da síndrome respiratória aguda grave 2 (SARS-CoV-2). Adeptos ao "anarquismo" metodológico, utilizamos diferentes ferramentas para responder nossas questões, desde a criação de um modelo mecanístico, ao uso de métricas de dissimilaridade, análises de cluster, inferências Bayesianas, regressões lineares, correlações, modelos epidemiológicos, e por fim, o uso indiscriminado das linguagens de programação Python e R. No quesito teórico, esta tese também é bastante multidisciplinar, tendo bebido de fontes das áreas de antropologia, ecologia, epidemiologia, evolução, genética, geografia, linguística, política pública, saúde pública e sociologia.

No **Capítulo 1**, revisamos como a evolução das línguas e das espécies estão muito além de meras analogias e acumularam uma crescente literatura que corroboram esses paralelismos em ambas as escalas micro e macroevolutivas. Espécies e línguas são formadas por uma população, formada por indivíduos que carregam genes ou palavras. As palavras precisam ser repetidas e os genes precisam ser replicados para que ambos sobrevivam por gerações. Tanto os genes quanto as palavras apresentam variabilidade, são herdáveis e possuem aptidão diferencial, atendendo assim às três condições para a evolução por meio da seleção natural (Godfrey-Smith, 2007).

A evolução de uma língua ocorre através da propagação de variantes linguísticos em uma população (Sneller & Roberts, 2018), por isso a divergência dialetal é considerada

um estágio inicial da divergência linguística (Honkola et al., 2018). Assim, um dialeto é uma variedade geográfica ou social de uma língua (Bagno, 2015), e um grupo dialético pode ser identificado através do uso comum de variáveis fonéticas, morfológicas e léxicas (Feagin, 2002). A partir da maneira como uma pessoa fala, sua pronuncia e as palavras que ela escolhe usar, é possível saber de onde ela é, sua idade, seu grau de escolaridade, e até seu gênero e orientação sexual.

Muitos trabalhos têm usado os dialetos de uma língua para estudar evolução linguística (Honkola et al., 2018; Katz & Andrews, 2013; Leemann et al., 2016, 2018; Syrjänen et al., 2016). Esses trabalhos geralmente possuem dados linguísticos históricos, coletados em entrevistas individuais em escala nacional, especialmente para países pequenos como a Finlândia, Suíça e Inglaterra. No entanto, para países com território extenso e economicamente em desenvolvimento, como o Brasil, tal tarefa demandaria tempo e recursos financeiros exorbitantes.

Portanto, no **Capítulo 2** criamos um teste dialetal para coletar dados linguísticos do português brasileiro (PB). Com o auxílio da internet e da ciência cidadã, testes de dialetos podem ser usados para complementar dados existentes, preencher lacunas e fornecer informações que os métodos tradicionais não conseguiriam em tempo hábil. Posteriormente, esses dados serão utilizados para demarcar os diferentes dialetos do PB, reconstruir o histórico de imigração no Brasil, e pesquisar a evolução do PB.

O Brasil, por mais que tenha apenas duas línguas oficiais: o Português e a Língua Brasileira de Sinais (Libras), possui em seu território outras ~274 línguas, em sua maioria oriundas dos povos indígenas originários (IBGE, 2010). Essa é uma “herança” de uma época pré-colonização europeia, onde estimam-se que apenas no território do Brasil de hoje, ocorriam 1.078 línguas distintas (Beilke, 2013). A origem dessa diversidade linguística nas Américas, e principalmente na América do Sul, sempre foi alvo de

diferentes hipóteses e especulações. Gregorio García (em 1607) e Antonio Vázquez de Espinosa (1630), ambos frades missioneiros e responsáveis por manuscritos importantes sobre os indígenas do "Novo Mundo", acreditavam que a diversidade linguística era culpa de uma intervenção demoníaca: uma maneira que o Diabo encontrou de impedir os esforços de conversão dos ameríndios ao cristianismo (Campbell, 1997).

Felizmente, a analogia entre espécies e línguas inspirou trabalhos que aplicam métodos ecológicos à dados culturais, com a expectativa de que os padrões de diversidade de línguas possam surgir dos mesmos processos que explicam os padrões de biodiversidade (Capitán et al., 2015). De tal modo, diferentes trabalhos têm encontrado fatores semelhantes responsáveis por determinar ambas diversidades, tais como variabilidade climática (Mace & Pagel, 1995; Nettle, 1998), barreiras geográficas (Axelsen & Manrubia, 2014; Huisman et al., 2019; Lee & Hasegawa, 2014), disponibilidade de recursos (Gorenflo et al., 2012), qualidade do solo e altitude (Michalopoulos, 2008), riqueza de parasitas (Fincher & Thornhill, 2008), e tamanho do grupo (Gavin et al., 2017; Pacheco Coelho et al., 2019).

Portanto, essa pesquisa visa também contribuir para explicações de como e porque existem tantas línguas no mundo, utilizando como modelo os padrões bioculturais encontrados nas Américas. No **Capítulo 3** investigamos os mecanismos responsáveis pela diversidade linguística na região Neotropical (México, Américas Central e Sul). Criamos um modelo mecanístico espacialmente explícito que incorpora altitude, recursos hídricos, precipitação e tamanho do grupo populacional como mecanismos capazes de predizer a diversidade linguística pré-colombiana observada no continente.

Línguas são frequentemente comparadas à um parasita, pois evoluem de forma adaptativa para se adequar aos seus "hospedeiros" humanos (Hung, 2019). Ambos línguas e vírus dependem de pessoas, vivendo em grupo, para se replicar e se propagar com

eficiência. Uma língua não sobrevive sem pessoas, enquanto pessoas existiriam sem uma língua. Da mesma forma, os vírus são parasitas obrigatórios, pois só sobrevivem se estiverem hospedados em uma pessoa ou outro organismo. Assim, doenças infecciosas aumentam conforme também aumentam o tamanho populacional, as conexões globais e a degradação ambiental (Stephens et al., 2016). De fato, tamanho populacional e conexões globais foram as principais causas para a rápida propagação inicial de COVID-19 no mundo (Pacheco Coelho et al., 2020), declarada como uma pandemia global em 2020 (WHO, 2020).

Sabendo que o vírus SARS CoV-2 é altamente transmissível e se propaga por contato entre pessoas, pela transmissão de gotículas respiratórias e por aerossóis suspensos no ar (WHO, 2021), a medida inicial mais eficiente de combatê-lo foi a instituição de políticas públicas de distanciamento social, uso de máscaras faciais e quarentenas em grande escala (Chu et al., 2020). No **Capítulo 4**, usamos um modelo epidemiológico SIR (Suscetíveis, Infectados, Removidos) para avaliar a efetividade das políticas públicas do Estado de Goiás em conter a propagação da COVID-19 em seu estágio inicial, entre março e maio de 2020.

## REFERÊNCIAS

Axelsen, J. B., & Manrubia, S. (2014). River density and landscape roughness are universal determinants of linguistic diversity. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1784), 20133029–20133029. https://doi.org/10.1098/rspb.2013.3029

Bagno, M. (2015). *Preconceito Linguístico* (56th ed.). Parábola Editorial.

Beilke, N. S. V. (2013). Do nativo ao pomerano: as línguas, os dialetos e falares vivos de um Brasil pouco conhecido. *Domínios de Lingu@gem*, 7(1), 263. Retrieved from https://doi.org/10.14393/DL13-v7n1a2013-14.

Burnside, W. R., Brown, J. H., Burger, O., Hamilton, M. J., Moses, M., & Bettencourt, L. M. A. (2012). Human macroecology: linking pattern and process in big-picture human ecology. *Biological Reviews*, *87*(1), 194–208. https://doi.org/10.1111/j.1469-185X.2011.00192.x

Campbell, L. (1997). *American Indian Languages: The Historical Linguistics of Native America*. New York: Oxford University Press, Inc.

Capitán, J. A., Bock Axelsen, J., & Manrubia, S. (2015). New patterns in human biogeography

revealed by networks of contacts between linguistic groups. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1802), 20142947–20142947. https://doi.org/10.1098/rspb.2014.2947.

Chu, D. K., Akl, E. A., Duda, S., Solo, K., Yaacoub, S., Schünemann, H. J., Chu, D. K., Akl, E. A., El-harakeh, A., Bognanni, A., Lotfi, T., Loeb, M., Hajizadeh, A., Bak, A., Izcovich, A., Cuello-Garcia, C. A., Chen, C., Harris, D. J., Borowiack, E., … Schünemann, H. J. (2020). Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis. *The Lancet*, *395*(10242), 1973–1987. https://doi.org/10.1016/S0140-6736(20)31142-9

Clayton, A. (2020). *How Eugenics Shaped Statistics*. Nautilus. https://nautil.us/how-eugenics-shaped-statistics-9365/

Diniz-Filho, J. A. F. (2022). *Darwin, Wilson e o "Racismo Científico."* Ciência, Universidade e Outras Ideias. https://www.blogalexdiniz.com/post/darwin-wilson-e-o-racismo-científico

Feagin, C. (2002). Entering the community: Fieldwork. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The Handbook of Language Variation and Change* (pp. 20–39). Blackwell Publishing Ltd.

Fincher, C. L., & Thornhill, R. (2008). A parasite-driven wedge: infectious diseases may explain language and other biodiversity. *Oikos*, *117*(April), 1289–1297. https://doi.org/10.1111/j.2008.0030-1299.16684.x

Gavin, M. C., Botero, C. A., Bowern, C., Colwell, R. K., Dunn, M., Dunn, R. R., Gray, R. D., Kirby, K. R., McCarter, J., Powell, A., Rangel, T. F., Stepp, J. R., Trautwein, M., Verdolin, J. L., & Yanega, G. (2013). Toward a Mechanistic Understanding of Linguistic Diversity. *BioScience*, *63*(7), 524–535. https://doi.org/10.1525/bio.2013.63.7.6

Gavin, M. C., Rangel, T. F., Bowern, C., Colwell, R. K., Kirby, K. R., Botero, C. A., Dunn, M., Dunn, R. R., McCarter, J., Pacheco Coelho, M. T., & Gray, R. D. (2017). Process-based modelling shows how climate and demography shape language diversity. *Global Ecology and Biogeography*, *26*(5), 584–591. https://doi.org/10.1111/geb.12563

Godfrey-Smith, P. (2007). Conditions for Evolution by Natural Selection. *The Journal of Philosophy*, *104*(10), 489–516. https://doi.org/0022-362X/07/0410/489-5

Gonçalves, B., & Sánchez, D. (2014). Crowdsourcing Dialect Characterization through Twitter. *PLoS ONE*, *9*(11), e112074. https://doi.org/10.1371/journal.pone.0112074

Gorenflo, L. J., Romaine, S., Mittermeier, R. A., & Walker-Painemilla, K. (2012). Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. *Proceedings of the National Academy of Sciences*, *109*(21), 8032–8037. https://doi.org/10.1073/pnas.1117511109

Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *205*(1161), 581–598. https://doi.org/10.1098/rspb.1979.0086

Honkola, T., Ruokolainen, K., Syrjänen, K. J. J., Leino, U. P., Tammi, I., Wahlberg, N., & Vesakoski, O. (2018). Evolution within a language: Environmental differences contribute to divergence of dialect groups. *BMC Evolutionary Biology*, *18*(1), 1–15. https://doi.org/10.1186/s12862-018-1238-6

Huisman, J. L. A., Majid, A., & van Hout, R. (2019). The geographical configuration of a language area influences linguistic diversity. *PLOS ONE*, *14*(6), e0217363. https://doi.org/10.1371/journal.pone.0217363

Hung, T. (2019). How Did Language Evolve? Some Reflections on the Language Parasite Debate. *Biological Theory*, *14*(4), 214–223. https://doi.org/10.1007/s13752-019-00321-x

IBGE. (2010). *Censo 2010: população indígena é de 896,9 mil, tem 305 etnias e fala 274*

*idiomas*. Instituto Brasileiro de Geografia e Estatística. https://censo2010.ibge.gov.br/noticias-censo.html

Kaessmann, H., Wiebe, V., & Pääbo, S. (1999). Extensive nuclear DNA sequence diversity among chimpanzees. *Science*, *286*(5442), 1159–1162. https://doi.org/10.1126/science.286.5442.1159

Katz, J., & Andrews, W. (2013). *How y'all, youse and you guys talk*. New York Times Online. https://www.nytimes.com/interactive/2014/upshot/dialect-quiz-map.html?r=0

Lee, S., & Hasegawa, T. (2014). Oceanic barriers promote language diversification in the Japanese Islands. *Journal of Evolutionary Biology*, *27*(9), 1905–1912. https://doi.org/10.1111/jeb.12442

Leemann, A., Kolly, M.-J., & Britain, D. (2018). The English Dialects App: The creation of a crowdsourced dialect corpus. *Ampersand*, *5*(August 2017), 1–17. https://doi.org/10.1016/j.amper.2017.11.001

Leemann, A., Kolly, M.-J., Purves, R., Britain, D., & Glaser, E. (2016). Crowdsourcing Language Change with Smartphone Applications. *PLOS ONE*, *11*(1), e0143060. https://doi.org/10.1371/journal.pone.0143060

Mace, R., & Pagel, M. (1995). A latitudinal gradient in the density of human languages in North America. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *261*, 117–121.

Maffi, L. (2005). Linguistic, Cultural, and Biological Diversity. *Annual Review of Anthropology*, *34*(1), 599–617. https://doi.org/10.1146/annurev.anthro.34.081804.120437

Manne, L. L. (2003). Nothing has yet lasted forever: Current and threatened levels of biological and cultural diversity. *Evolutionary Ecology Research*, *5*(4), 517–527.

McLemore, M. R. (2022). *The Complicated Legacy of E. O. Wilson*. Scientific American. https://www.scientificamerican.com/article/the-complicated-legacy-of-e-o-wilson/

Michalopoulos, S. (2008). The Origins of Ethnolinguistic Diversity: Theory and Evidence. *SSRN Electronic Journal*, 0–54. https://doi.org/10.2139/ssrn.1286893

Mirazón Lahr, M. (2016). The shaping of human diversity: filters, boundaries and transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1698), 20150241. https://doi.org/10.1098/rstb.2015.0241

Moseley, C., & Asher, R. E. (1994). *Atlas of the World's Languages*. Routledge.

Nettle, D. (1996). Language Diversity in West Africa: An Ecological Approach. *Journal of Anthropological Archaeology*, *15*(4), 403–438. https://doi.org/10.1006/jaar.1996.0015

Nettle, D. (1998). Explaining Global Patterns of Language Diversity. *Journal of Anthropological Archaeology*, *17*, 354–374.

Pacheco Coelho, M. T., Pereira, E. B., Haynie, H. J., Rangel, T. F., Kavanagh, P., Kirby, K. R., Greenhill, S. J., Bowern, C., Gray, R. D., Colwell, R. K., Evans, N., & Gavin, M. C. (2019). Drivers of geographical patterns of North American language diversity. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1899), 20190242. https://doi.org/10.1098/rspb.2019.0242

Pacheco Coelho, M. T., Rodrigues, J. F. M., Medina, A. M., Scalco, P., Terribile, L. C., Vilela, B., Diniz-Filho, J. A. F., & Dobrovolski, R. (2020). Global expansion of COVID-19 pandemic is driven by population size and airport connections. *PeerJ*, *8*, e9708. https://doi.org/10.7717/peerj.9708

Pagel, M., & Mace, R. (2004). The cultural wealth of nations. *Nature*, *428*(March), 275–278.

Sneller, B., & Roberts, G. (2018). Why some behaviors spread while others don't: A laboratory simulation of dialect contact. *Cognition*, *170*(October 2017), 298–311. https://doi.org/10.1016/j.cognition.2017.10.014

Stephens, P. R., Altizer, S., Smith, K. F., Alonso Aguirre, A., Brown, J. H., Budischak, S. A., Byers, J. E., Dallas, T. A., Jonathan Davies, T., Drake, J. M., Ezenwa, V. O., Farrell, M.

J., Gittleman, J. L., Han, B. A., Huang, S., Hutchinson, R. A., Johnson, P., Nunn, C. L., Onstad, D., … Poulin, R. (2016). The macroecology of infectious diseases: a new perspective on global-scale drivers of pathogen distributions and impacts. *Ecology Letters*, *19*(9), 1159–1171. https://doi.org/10.1111/ele.12644

Syrjänen, K., Honkola, T., Lehtinen, J., Leino, A., & Vesakoski, O. (2016). Applying Population Genetic Approaches within Languages. *Language Dynamics and Change*, *6*(2), 235–283. https://doi.org/10.1163/22105832-00602002

Templeton, A. R. (1998). Human Races: A Genetic and Evolutionary Perspective. *American Anthropologist*, *100*(3), 632–650. https://doi.org/10.1525/aa.1998.100.3.632

WHO. (2020). Timeline of WHO's response to COVID-19. *World Health Organization*.

WHO. (2021). Coronavirus disease (COVID-19): How is it transmitted? *World Health Organization*.

**CAPÍTULO 1**

---

# *Micro and macroevolutionary mechanisms underlying language evolution*

Christielly Borges[1]

[1] Departamento de Ecologia, Universidade Federal de Goiás, Goiânia, Brasil

*Status*: Em preparação para submissão

**ABSTRACT**

Languages and species are not fixed in space and time, they are continuously evolving and adapting, often by the same mechanisms. There is no consensus to how language first emerged, but the capacity for acquiring language is a human universal biological trait, not a cultural invention. Modern humans incontestably had language 70 kya, when we first dispersed out of Africa, and it is unarguable that language has not remained stable throughout this time. There are today approximately 7,000 extant languages, distributed unevenly across the world. Species and languages are made up of a population, made of individuals that carry genes or words. Both genes and words show variability, are inheritable and have differential fitness, thus fulfilling the three conditions for evolution by means of natural selection. Here we review how language and species evolution are beyond mere analogies and have accumulated a growing body of work corroborating theses parallelisms on both micro and macro evolutionary scales. The divergence observed within a language and between languages are often the reflection of different selective pressures, as languages evolve from a complex interplay between biological, cultural and social mechanisms. Theoretical and empirical approaches to language evolution and language diversity are highly multidisciplinary and the methodological tool-swapping has greatly advanced empirical support for language evolution.

**Keywords:** biocultural conservation, evolutionary biology, human macroecology, language phylogenetics, meme

## 1. Introduction

Before “On the Origin of Species” revolutionized science in 1859, linguists as far back as 1807 discussed how languages descended from earlier ancestral languages [1]. Many argue Darwin was greatly influenced by this knowledge [2] and it is highly cited how he notoriously claimed the process of descent with modification proposed for languages and species were “curiously parallel” [3–5]. This analogy was also used by Asa Gray, Charles Lyell and Thomas Huxley, as it made the theory of evolution more plausible and presentable at that time [6,7]. Nonetheless, we still ask “Is language evolution Darwinian?” [8,9], when the question should be “How are language and species evolution the same?”.

Since the 19th century, evolutionary theory has advanced considerably, especially in the field of evolutionary biology. Nonetheless, despite linguists agreeing straightway with Darwin’s analogy between language and species [1], language evolution remained a topic restricted to humanities and cultural studies until the 20th century, when Chomsky identified language as a biological object of study [10]. Later, in a paradigm-shifting paper, language evolution was successfully explained in Neo-Darwinian terms [11]. There has been numerous advances in language evolution research in the 21st century, even earning its own journal [12]. While the field may still be mined with multidisciplinary theorical disagreements [13–21], more and more, language evolution is comprehended as a multidimensional complex process driven by biological, cultural and social mechanisms.

The study of language evolution focuses on natural languages, which are spoken or signed, since they are innate to humans. Contrarily, written or artificial languages are considered strictly cultural inventions [22]. Nonetheless, to define what is a language, much like to define what is a species, has profound theoretical complications and multiple layering. According to textbook definitions, species can be determined on the basis of reproductive ability, morphology, ecological niche and DNA sequence [23]. Likewise, textbooks definitions

of languages are on the basis of mutual intelligibility, political territory, cultural differences and distinctive writing systems [24]. Ambiguous zones are to be expected due to the ongoing evolutionary processes in both. Though, curiously, out in the field neither biologists nor linguists have difficulty in determining the individuals of a species or the speakers of a language.

Here we review how language evolution and species evolution are beyond mere analogies and have accumulated a growing body of work corroborating theses parallelisms (Table 1). Species and languages are made up of a population, made of individuals that carry genes or words, respectively. Genes make up a genome and words make up a vocabulary, both are composed of sequence ordering that can be altered to create something new. Words need to be repeated and genes need to be replicated in order for both to survive over generations. We can then conclude genes and words show variability, are inheritable and have differential fitness, thus fulfilling the three conditions for evolution by means of natural selection [25]. Thus, languages and species are not fixed in space and time, they are continuously evolving and adapting, often by the same mechanisms. This process of evolving and adapting generates diversity, that is spatially and temporally explicit, and creates in-betweens such as dialects and subspecies.

## 2. Language, what is it used for? Absolutely everything!

Natural human language is learned through environmental interactions and is externalized to exchange information, to express emotion and identity, to bond socially, to share concepts, to perform religious and cultural rituals, and to explore the sonority of words, through rhymes, poetry and song [24]. Language can be private as well, as it serves as an instrument of thought and to create mental representations of entities (i.e., imagine a flying rainbow unicorn) [26].

Spoken language is a multimodal phenomenon, as it is consistently accompanied by the simultaneous use of face movements and hand gestures [27]. It is also a cognitive ability [28], demanding mimesis, imitation and enhanced memory. Human language has complex phonetic, phonological, morphological, syntactic and semantic structures, which enable a variety of vocabulary, grammar, pronunciation, phrases and ultimately, boundless discourse. Languages also vary in the exhibitions of these structures, which promotes greater language diversity [29].

Languages are transferred vertically between generations, such as parents teaching their children to speak, and horizontally within speakers of the same generation, such as teenagers sharing slang [30]. They are markers of ethnic, national, geographical, religious and social identities, and can facilitate inferences on class, status, occupation, age, gender and sexual orientation [31]. Hence, it is not frivolous to state that the way a person speaks answers questions about who they are and where they are from.

Because language is widely used in various contexts, researchers diverge on its main function: did language evolve for communication, for thinking, or is it merely a sophisticated social device where communication and conceptual representation are by-products? Going further, even when there is agreement communication is a primary function of language, there is disagreement on whether the information being exchanged is technological, for tool-making, or social, for gossiping. It was estimated humans spend from 65% to 78% of speaking time gossiping [32]. But the sharing of technological knowledge is essential for the dispersal to unknown habitats and niche construction, fundamental for human survival [33]. How and why, then, did language evolve?

To further discourse on the evolution of language, we have to distinct between the capacities for the emergence of language, as in the biological mechanisms necessary for the production and processing of speech and signals, and languages, as in the structured communication systems we call English, Portuguese, Mandarin, etc. [34]. As we will see, how

language came to be is not a settled question, there are no consensus and many pieces of the puzzle are missing (but even hominin evolution is not resolved in the literature). Nonetheless, we do have language, approximately 7,000 of them. They differ greatly in sound, conceptual and vocabulary systems between cultures and across time. The way language behaves might give a hint to how they first emerged.

## 3. On the origin of languages

### *3.1. Homo sapiens have a body predisposed to speak*

All animals have language, if we define it only as a communication system [35]. Nonhuman primates respond to sign language and other nonhuman animals have communication systems with complex learned vocalizations [36–38], distinct dialects [39–41], semantics [42], syntax [43] and even symbolic representation of space and time [44,45]. Nonetheless, nonhuman animals' communication systems are often limited to their fitness, such as announcing food or predators, even if their cognition allows other forms of mental representations that they cannot express [26].

Spoken language is dependent on the integration of auditory perception and vocal production [46]. Speech is a complex motor process that uses over 100 muscles and requires coordinated neural connections and movements of the vocal tract (i.e. the larynx, pharynx, oral and nasal cavities), the tongue, the lungs, the soft palate, the lower jaw, the lips [46,47], and even swallowing [48]. The expansion of the human brain, both in size and neural connectivity, evolved brain regions, such as the mirror system, that allows complex imitation and action recognition, which has an important role in language evolution [49].

The changes in the larynx, pharynx and mouth that allowed humans to developed a vocal tract and enable fast and efficient speech came at the cost of less efficient breathing and swallowing [50]. Thus, although humans have a body predisposed to speak, anecdotally, we

may easily choke on our food while a chimpanzee would not. Nonetheless, nonhuman primates share many cognitive and brain mechanisms with humans [51] and have the ability to produce contrasting vowels [52], both essential for the evolution of spoken language. Many argue that what separates humans from other animals is the capability of language and speech [53–55]. But while *Homo sapiens* do in fact share anatomy and cognitive abilities related to speech evolution with apes, cetaceans, bats, elephants and even birds and bees, it is our neural control over our speech that really separates us from other animals [56].

### *3.2. Not alone after all*

*Homo sapiens* are probably not the only species with the capacity of spoken language, as there is increasing evidence *Homo neanderthalensis* also had modern hyoids [57,58] and outer and middle ears consistent with speech and auditory capacities [59,60]. Further, Neanderthals also exhibited complex cultural adaptations, such as the usage of personal adornments and mortuary ceremonies [61], which is considered to be correlated with the emergence of language [62]. These indicate Neanderthals had at least some form of speech. Other species in the *Homo* genus show little indication to have had components of modern language and earlier hominins, such as the *Australopithecus afarensis* had a chimpanzee-like hyoid bone, as evidenced by a 3.3-million-year-old fossil [63], making the absence of speech clear.

This places the origin of spoken language at least around ~530 ka and probably deriving from the last common ancestor between modern humans and neandertals [64]. *Homo sapiens* appeared around 200,000 years ago [65], therefore the origin of modern humans is intrinsically linked to spoken language. All humans talk, clinical cases being the exception, and every population around the globe has a language. It is clear, as evidenced by all the cognitive and morphological changes listed above, the capacity for acquiring language is a human universal

biological trait, not a cultural invention that was passed down or taught between people, like agriculture or written language (but see [66] for an alternative perspective).

### *3.3. Hominins are highly social animals*

We have discussed the contemporary uses of language and the human morphological changes that allow them. Nonetheless, an essential question in any evolutionary model remains: why do we speak? Primary models for the emergence of language focused on rudimentary aspects of speech and were untestable and unscientific [11]. They can be grouped into five categories, where it was proposed human language emerged from (1) the imitation of animal calls, known as the "bow-wow" theory, (2) the noises made in moments of strong emotion (i.e., pain or anger), known as the "pooh-pooh" theory, (3) the imitation of physical sounds, known as the "ding-dong" theory, (4) the rhythmical grunts of physical efforts, known as the "yo-he-ho" theory, and finally, (5) the need to express love, poetry and music, known as the "la-la" theory [24]. As evidenced by the derogatory attributed names, these propositions have long been discarded.

More contemporary models concretely specify selection pressures for the emergence of language, especially its social function. To paraphrase [8], "only an early social function could have launched language on its evolutionary path". These models focus on the order of emergence, i.e., gesture came first and speech later, which also fails to proximate the mechanisms by which language evolved. We will discuss these models briefly, as they are beyond the scope of this review (but see [67–69]).

#### *3.3.1. Speech to gossip or to advance technology?*

The "grooming and gossip" hypothesis poses hominins went from manual grooming to vocal grooming as a cheaper form of social maintenance in large groups [70]. Nonhuman primates

spend approximately 20% of their time grooming friends and allies in groups of about 80 individuals [32]. But as survival pressure to explore unknown (often risky) habitats increases, so does group size. Hominins pushed group size boundaries and alternative mechanisms for bonding, such as low-cost vocal sounds, emerged [32]. Advantageously, vocal grooming, which was probably song-like, allows you to "groom" multiple friends and frees your hands for other tasks. Dunbar's hypothesis argues vocal grooming gradually evolved to language in the initial form of gossip. It is hard to keep tabs on everyone in a large social group, thus information about the social network when "your back is turned" is deeply valuable to avoid deceit and deepen alliances.

Behavioral studies show that gossip is fundamental in spreading reciprocity, trust, reputations [71], the rules and norms of a group, and for vicarious learning [72]. Having a good reputation guarantees more cooperative allies, as people tend to cooperate and bond less with individuals with a bad reputation [72,73]. Further, gossip drives altruistic behavior in individuals worried about their own reputation [74,75]. Thus, gossip is both a controlling tool for egotistic behavior and a protector against selfish, antisocial and exploitive individuals [76]. Nonetheless, gossip is not the only speculated essentiality for the emergence of language: tool-making rates fairly high on that list as well. Dating back to 2.5 million years ago, sharped-edged stone flakes are the oldest and the first evidence of hominins' capabilities of modifying materials into useful tools [77]. Researchers argue that since stone tool-making is an activity that requires both hands, gestural communication emerged first and eventually lead to vocal communication as a more efficient way to culturally transmit tool-making knowledge [78,79]. This hypothesis has received support from neurocognitive research, which shows tool-making emerges from perceptual-motor foundations [80] and that the brain's "language areas" are indeed shared with other behaviors, such as tool use [81]. This view puts language in a passive evolutionary role, evolving as a by-product of a cognitive ability when language is a cognitive

ability in itself. Controversially, recent experimental flintknapping studies have shown evidence both in favor [62,82] and against [83] speech evolving as a superior method of cultural transmission.

We focused on the gossip and tool-making hypotheses as they inevitably include both gestural and vocal communication as the core traits of language. What they have in common is that they converge in advocating for a coevolution between complex behavior and speech through human evolution [60]. Other models state hierarchical syntax is the ultimate step in language evolution, or state gestural and musical protolanguages gave way to a vocal speech [67]. Some researchers even rule out the possibility of language as an ancient feature of *Homo* and prefer a "language universals" hypothesis, where all languages have the same properties and appeared in a single evolutionary event in *H. sapiens* around 100 kya [66,84].

## 4. Language on its own evolutionary path

Despite conflicts on the chosen model for language emergence, there is consensus *H. sapiens*, the sole speaker in current times, diverged from other species and has since suffered different evolutionary pressures. Therefore, all lines of research agree modern humans incontestably had language 70 kya [85], when we first dispersed out of Africa, and it is unarguable that language has not remained stable throughout this time. Thus, any language evolution model needs to incorporate multiple hypotheses and mechanisms, including an adaptative perspective, such as environmental determinants of language [68], as to not create false dichotomies between language's essential traits.

The evolution of speech and emergence of language required a complex interchange between genetic mutations [86], neural development [10], cultural evolution [87,88], natural selection [11], adaptation [89], social pressure [90], phenotypic plasticity, exaptation and constraints [91]. Here we acknowledge the multidimensionality of the mechanisms underlying

the emergence of language and agree previous proposed models are complementary in explaining different parts of the same problem [67,92]. All these processes are not mutually exclusive, as some authors claim, and indicate language emerged from an interaction of anatomic, cognitive, environmental, social and cultural-historical mechanisms. These same forces underlie language evolution, the structured communication systems we use to speak.

Biological and linguistic evolution share many processes, as they both have discrete heritable units, similar transmission mechanisms, form separate populations due to equivalent evolutionary pressures, and even are congruent in spatial patterns (Table 1). Like species, languages are linked through hierarchical and nested patterns of descent. This is the reason we can illustrate these relationships in the form of a family tree, or phylogeny. The divergence observed between different linguistic groups are often the reflection of the action of different selective pressures and evolutionary processes [93].

**Table 1. Corroborated parallelisms between language and species evolution.**

| | **Biological Evolution** | **Linguistic Evolution** | **References** |
|---|---|---|---|
| Interactor | Species | Language | [26,27,36–45,28,46,47,29–35] |
| Replicator | Genes | Words | [48–50] |
| Discrete heritable units | Discrete characters | Lexicon, grammar and phonology | [43,46,51–53] |
| Transmission | Mainly Vertical | Mainly Horizontal | [54] |
| | Horizontal gene transfer | Loanwords and structural borrowing | [55–57] |
| | DNA replication | Teaching, learning and imitation | [58–60] |
| Microevolutionary mechanisms | Natural selection | Natural selection | [50,61–68] |
| | Drift | Drift | [69–72] |
| | Mutation | Mutation | [73–75] |
| | - | Cultural selection | [47,59,76,77] |
| | - | Social selection | [58,69,71,78–85] |
| Macroevolutionary mechanisms | Cladogenesis | Cladogenesis | [86,87] |
| | Anagenesis | Anagenesis | [88,89] |
| | Allopatric speciation | Geographic isolation | [90–92] |

| | | | |
|---|---|---|---|
| | Sympatric speciation | Social and contextual niche | [93,94] |
| | Hybridization | Language contact | [95–98] |
| | Hybrid species | Creole | ∗ |
| | Sterile species | Pidgins | ∗ |
| | Dispersal and migration | Dispersal and migration | [26,27,85,91,92,99–102,28,29,33,34,41,44,83,84] |
| | Extinction | Language death | [103–106] |
| | Homology | Cognate | [27,34,107] |
| Rate of evolution | Slow | Fast or slow | [76,108,109] |
| Spatial patterns | Geographic clines | Dialect continuum | [110–112] |
| | Latitudinal Diversity Gradient | Latitudinal Diversity Gradient | [113,114,123,115–122] |
| | Rapoport's rule | Rapoport's rule | [113,124] |
| | Biodiversity hotspots | Language hotspots | [125,126] |
| Past information retainment | Chronospecies | Proto-language | [53,127] |
| | Fossil | Ancient text | [88] |
| | Genetic admixture | Loanwords or structural borrowings | [107,128] |

First versions of this table are in [5,33,129]. Our adaptation updates the parallelisms and includes the empirical advancements. ∗ These are theorical comparisons.

Species inherit genes and their discrete characters through vertical transmission, from parent to offspring. Nonetheless, genes can be transferred horizontally between nonrelated individuals, in a process called horizontal gene transfer (HGT), where alien genes are incorporated into a genome by DNA recombination or insertion [94]. HGT is common in eukaryotes, bacteria, plants and even between species such as plant parasitic-to-host HGT and between organisms such as bacteria-to-fungus HGT [95].

Languages inherit words and its properties: phonology, lexicon and grammar through vertical transmission as well, when parents teach their children to speak. But language transmission is mainly horizontal, within a generation of speakers. Peer influence is much greater than parent influence in language acquisition [96], as children build language identity in a process of group distinction [97]. For instance, first generation children of immigrant parents generally do not inherit their parent's second language (L2) accent. Thus, due to

horizontal transmission, language also evolves through cultural and social evolutionary mechanisms, which allows knowledge accumulation and long-term transmission of language.

The heritability in both cases is continuously transferred between individuals in a group, which allows the discrete heritable units to persist across generations and their frequencies to be changed over time [98]. This is exactly what makes language evolution evolutionary. Evolutionary theory, or evolution by natural selection, is applicable whenever changes in a system are due to variation, differential fitness and inheritance [25]. Additionally, since the Modern Synthesis, its understood that not every evolution happens by natural selection, though every adaptive evolution does, and changes may occur by random drift and mutation as well [99].

### *4.1. Language evolution, never revolution*

Some researchers argue languages do not evolve; they merely change [53,100]. Language change, nonetheless, is comparable to the microevolution of species. Microevolution is any change in gene frequency within a population over a short period of time [101]. Due to the social learning required by languages, language change often occurs by learning biases at the individual level. Overall, microevolutionary processes within language happen by borrowing, grammaticalization, semantic or sound change, innovation or errors [102], which means the replacement of a word by a new pronunciation, by a non-cognate, or by a new meaning. Linguistic microevolution creates variability and takes off pressure from macroevolutionary processes from having to explain complex and intrinsic aspects of language [67]. On the other hand, macroevolution discusses processes that are above the species level and take a long period of time to happen [103].

More importantly, these disciplinary distinctions occur because evolutionary processes are scale dependent and have mechanisms of feedback, such as gene-culture coevolution, eco-

evolutionary feedback and evolutionary development (evo-devo). Natural selection acts at the individual level, while adaptive and non-adaptive evolutionary forces, such as selection and drift, act at the population-level, on a microevolutionary scale (Figure 1a). The forces at the population-level will determine the scale of any change, whether phenotypic or linguistic, and any ecological consequences [104]. On a larger temporal and taxonomic (macroevolutionary) scale happens the emergence of a new species or language, by the accumulation of populational changes or by distinct macroevolutionary processes (Figure 1b). Dispersal, speciation and extinction are fundamental macroevolutionary processes shaping patterns of diversity. This means macroevolution is more than the accumulation of microevolutionary processes, and that natural selection operates on different scales [105]. Just as change within a population eventually leads to differences between populations, so too language change within a language may lead to dialects. On another scale, just as species interaction may lead to niche partitioning and speciation, so too language contact and contextual niche adaptation may lead to new languages.

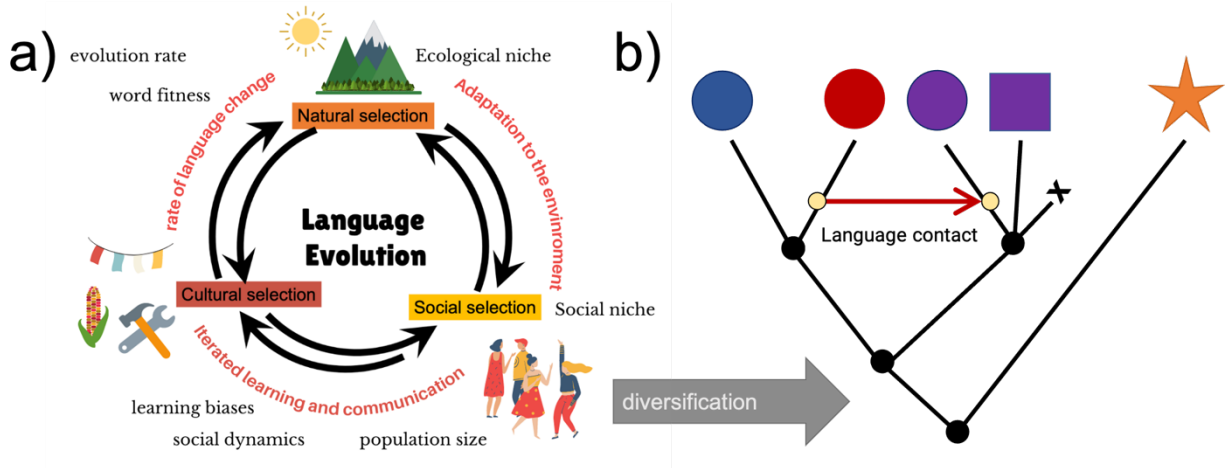


**Figure 1.** Language is a bio-cultural hybrid. a) On a microevolutionary scale, language evolves by natural, social, and cultural selection mechanisms that feedback into each other. b) On a macroevolutionary scale, language diversifies by diversification processes, such as language contact and geographic isolation, dispersal or migration, and extinction.

Language can further be studied in a population biology perspective, by bridging evolution and ecology together and providing a framework to account for the effects of demography on phonetical, lexical and grammar variants within a language [98,106], much in the same way demography effects phenotype and gene frequency [104]. Integrating ecological and evolutionary dynamics allows us to understand eco-evolutionary feedbacks [107], which happens when a population alters their environment and that alteration feedbacks the evolution of the population [108]. The most important aspect of a species environment tends to be its own population's demographics, since a population's size and density drives adaptation demands [107]. So too a population's demography has an important role in language evolution, as language change is greatly influenced by the environment, whether in an ecological nature or in socioeconomic dynamics [109].

It is important to understand the role of population dynamics in shaping selection pressures, since micro and macroevolutionary mechanisms are linked. We will discuss language evolution on both micro and macroevolutionary perspectives, since their methodological approaches were applied to languages with great success [98,110–114]. We will further discuss methodological approaches from macroecology and conservation biology, since biodiversity and linguistic diversity overlap in the geographical space.

## 5. Microevolutionary mechanisms of language evolution

### *5.1. The reproduction of words*

Before we discuss the specific microevolutionary mechanisms of language evolution, it is important to highlight that the limit to the analogy between genes and words lies in their reproductive systems. Survival of the fittest means organisms that are able to pass their genes to their offspring through sexual reproduction. Words, of course, do not sexually reproduce and are not transmitted genetically. Asexual organisms, such as viruses, have no mating success

either and are also not genetically passed down. Nonetheless, they still have reproductive fitness. Thus, both viruses and words depend on a type of reliable copy machinery to replicate and perpetuate in time.

Viruses are obligate parasites, meaning viruses cannot reproduce on their own and replicate at the expense of their host [115]. In a nutshell, the virus first enters the host cell membrane and releases viral genome into the cytoplasm. Viral genome will then use cellular machinery to replicate and assemble new progeny virions [116]. Finally, virions are released and infect other host cells, while part of the viral genome may integrate into the host's chromosome to maintain viral latency [116]. This life-cycle promotes viral gene expression and viral replication, after all, a viral attack is much more efficient to infect and perpetuate in multiple hosts than a single virus.

Analogously, if we think of a child learning to speak, she will learn through a bombardment of a single word being repeated over and over again by her parents, family members, and others around her [117]. Each time the word is pronounced it creates a replicate, meaning a "word-offspring" is generated. The child will eventually become infected by the word-offspring, and it will "take up residence" in the child's brain, where it will remain dormant until it is replicated by the child [117]. In the direct words of [117] "An informational thing doesn't have a mind, any more than a virus does, but, like a virus, it is designed (by evolution) to provoke and enhance its own replication and every token it generates is one of its offspring".

Dennett (2017) further argues words want to be said, just like viruses want to be replicated, and in this sense, words are selfish in the same way genes are selfish. If a word is not spoken, it will go extinct, thus the "repetition" of words is beneficial for the descendants of words themselves and not for the descendants of the speaker. This means words are the perfect memes, with their own fitness and replicative ability [117]. Memes are replicators, spreading from brain to brain through imitation, such as genes are replicators, spreading from body to

body through sexual reproduction [118]. In the virus analogy, memes parasitize brains as a vessel to propagate themselves much in the same way viruses parasitize the genetic mechanisms of their host cells [118].

Dawkins had already proposed a cultural trait may have evolved in a specific way because it is advantageous to itself [118]. A meme perspective on language evolution also offers the idea of information occupying brains and being spread without it being understood by the hosts, indicating human comprehension is unnecessary for the spread or fixation of a meme in a population [117]. The evolution of words, however, are beyond the meme, since the meme is only how they spread, and they evolve by copying errors or environment pressure. This further means the *design* of a cultural trait, of a language, has no author or architect, its aptness being the product of natural selection [117].

### *5.2. Language is shaped by natural selection*

There are abundant evidences language adapts to different environments and abiotic factors, remarked especially in the way they are used to communicate about the world. Phonology, lexicon and grammar are often assessed separately as different components of language, but all three show variation, heredity and fitness, and have also been shown to be influenced by different environmental factors [119]. Just as specific morphological structures, such as beak size and shape, can be seen as an adaptation to the environment, so too can specific speech sounds and vocabulary be interpreted as adaptations [89].

The Acoustic Adaptation Hypothesis (AAH) states acoustic communication is adapted to habitat structure to optimize signal transmission and to reduce sound degradation [120]. For example, birds habiting areas of dense vegetation vocalize in lower frequencies and simpler temporal structures [121]. Similarly, a large-scale study of 663 languages, found languages spoken in areas of higher precipitation and tree coverage show lower frequency of consonant

use [122]. A follow-up study also found languages spoken in higher temperatures favor more sonorous sounds [123]. In both cases, a harsher environment shapes the linguistic structure to be simpler and more easily transmitted to a far-away listener.

In tune with AAH, whistled languages, the whistled speech of a natural language, are used as a form of long-distance communication and are spoken mainly by populations in mountainous and dense forested regions where crossing is difficult [124]. Whistled speech has been described in over 60 languages [124], emerging mostly independently, which shows adaptation to environment roughness, social isolation and even human hearing [125]. Though whistled speech is a clear adaptation to overcome pressures of the acoustic environment, it has not been tested in an AAH context.

Atmospheric conditions have also been at the center of environmental impact on sound production studies. Languages with ejective sounds tend to occur in regions of high altitude, such as the Southern African plateau, the East African rift, the North American cordillera and the Andes [126]. Ejectives are glottalized sounds that do not use vocal cords in its production, and are made instead by compressing the air in the pharyngeal cavity [126]. It was proposed ejectives reduce water vapor lost through exhalation, decreasing high elevation dehydration. Nonetheless, this work has been criticized for its statistics and the lack of clarity in the proposed physiological mechanism [127]. Another large-scale study showed air dryness creates languages with absence of tones and more consonants [128]. That is not to say a speaker of a tonal language would not be able to communicate in an arid region, but that such an adaptation would not be acquired by a language spoken in such a region over a long period of time. The authors back their hypothesis with a survey of laryngology studies showing ambient air with reduced humidity affects human phonation [129].

Word competition is analogous to biological competition; thus, words show fitness in the frequency of their usage over another. Words may form a synonym set, called synset, where

many words mean the same thing. Synsets evolve by natural selection since they show variation in words entering and leaving the synset, word formation is hereditable and differential fitness happens as some words become more frequent than others [130]. In fact, frequently used words evolve at slower rates while infrequent words are replaced quicker [131]. Shorter and more distinctive words show greater fitness, as they take less social and individual effort to be remembered, pronounced and understood [132]. For instance, the words for the numbers “two”, “three” and “five” remained stable for 10,000 to 100,000 years, which is 3.5-20 times longer than other lexical replacement [133].

In the same way many words can have the same meaning, (i.e., a person of great mental capacity may be referred to as “smart”, “intelligent”, “clever” or “sharp”), different meanings may use the same word (i.e., “bat” may refer to a flying mammal or a baseball equipment). Preferences for one word over another evolve over time [130]. When tested against random drift, word frequency for over 91% of studied meanings was driven by a selection process, whether directional selection, positive frequency-dependent selection or a combination of directional and positive selection [134]. Selection was also the driver of irregular forms of past-tense verbs, which followed rhyming patterns [135].

Lexical adaptation to local physical environment is less controversial and better documented given populations often use and name endemic flora and fauna or have specific behaviors in accordance to their surroundings. But the environment drives lexicon variability in an adaptative way as well. For instance, although every human has the same anatomical body with the same limbs, clinical cases being the exception, universal cross-linguistical categorization is not the norm [136]. Many languages in the tropics do not distinct limbs and use a single term for both “hand” and “arm”, for “foot” and “leg” [137], and for “finger” and “hand” [138]. It was hypothesized individual limb parts are more noticeable in colder climates,

due to the inevitable use of winter clothing, such as gloves, socks and shoes, singling out the hands and foot [137].

Languages in the tropics also lack words for the color "blue" more often than languages in temperate zones [139], something the authors related to the increase of UV-B insolation (also correlated to latitude). A more recent study found warm colors (red/yellow) are better communicated, thus show better fitness, than cool colors (green/blue), because natural objects, i.e., fruits, tend to be warm colored while backgrounds, i.e., a forest, are cool colored [140]. Industrialization in Western cultures promotes artificial colored objects, increasing color-naming efficiency in those languages [140], further proving languages adapt to the environment when specific color ranges are presented [141].

Perhaps most famously, the case of innumerous words for "snow" and "sea ice" acquired by circumpolar populations speaking Inuit and Yupik languages [142], illustrates language adaption to the surrounding of its speakers. A study found languages that use the same term for "ice" and "snow" are spoken exclusively in the tropics [143]. Though some natural languages spoken in warmer climate do distinct lexically between ice and snow, these appear to be borrowings from colonial languages [143]. In parallel, speakers of the Western Desert dialect of the Pintupi language have up to 18 words for "hole", including clear distinction of different animal's burrows [144]. Both examples highlight speakers' need for informative terms about their environment and lifestyle.

### *5.3. Language is shaped by random drift and mutation*

The neutral theory of evolution states changes at the molecular level are caused by the stochastic fixation of random drift or neutral mutation, rather than by selection [99]. Linguists are no strangers to the concept of drift, where it was proposed initially as a process of dialect development by individual variation [145]. Clear analogies between biology's understanding

of drift and cultural drift [146] and linguistic drift have since been propositioned. Drift is now understood as a random copying process [147], where a speaker copies the speech of another speaker [134].

Random drift and mutation are greater in small populations, due to the greater chance of fixation of deleterious variants and greater vulnerability to stochastic events [148]. A case has been found for Polynesian languages, where small populations had greater rates of word loss without turnover [149], comparable to the loss of heterozygosity and diversity arising from random sampling [150]. Smaller populations are more vulnerable to drift as they seem to vary more in their linguistic behaviors [151,152]. Nonetheless, drift alone was insufficient to explain the emergence of a New Zealand English dialect [153]. Phonetic drift also decreases as L2 learners gain language experience and exposure [154].

Selection is a stronger force than drift in word frequency [134], but random drift seems to be the driver of rarer words forms and replacement [135]. Drift further accounts for the replacement of old linguistic variants, such as verb-object for object-verb word order from old to modern English, the power-law distribution of word frequencies and the replacement of new (also rare) variants [155].

Mutation can be lexical or phonetic, as words just like gene sequences can show insertions, deletions and reversals. Mutations happen by laziness of the tongue, through sound assimilation, grammaticalization, consonant and vowel weaking or alternation [24]. For instance, Celtic languages show initial consonant alternations [156,157], Nuer has vowel-lowering mutation [158] and Welsh shows vowel alternation [159]. Mutations tend to be neutral and determined by random drift, with small or zero effect in a population (especially large populations). However, mutations may also be deleterious or advantageous, determined by natural selection, or slightly deleterious or advantageous, in which case it is determined by both random drift and natural selection [148]. In other words, mutation needs selection pressure to

favor them in order to guarantee their persistence in next generations. Word mutation, much like genetic mutation, may be irrelevant from an evolutionary perspective, but may also accumulate over time and eventually give rise to a new language.

### *5.4. Language is shaped by cultural selection*

Culture refers to any nongenetic information transmitted through social learning. Once more humans are not alone in having culture, as animals have social learning skills and often innovate social or foraging behavior, which produces stable cultures and even effects gene distribution [160–162]. Nonetheless, human culture stands out for being cumulative, generating intricate technology that an individual would otherwise not create alone [163]. In other words, our cumulative culture allows us to stand "on the shoulder of giants".

Culture is increasingly seen as a driver of positive selection throughout human evolution, as up to 10% of our genome was recently affected by selective sweeps [164] probably favored by our cultural practices [165]. Language is at the center of cultural practices, such as religion and trade, which are dependent on language for its transmission and accumulation [166]. Thus, to isolate language from cultural practices is rather tricky, especially since most studies treat language as strictly cultural, and not as a bio-cultural hybrid.

As discussed previously, humans evolved biological underpinnings that allows and constrains language, but socially learns the specific linguistic structure used in their environment [29]. Cultural selection drives language change through mutations and the selection of beneficial variants from learning biases [167]. Social learning generates bias from a speaker copying another speaker, including content biases, the fitness of informative communication, model-based biases, when a speaker copies a person of prestige, and frequency-dependent biases, when a speaker copies frequent or infrequent words [168]. These changes can create variation within a population overtime, acting at the population-level.

Cultural change may also happen at the individual-level in a process called biased transformation [167], when individuals direct the transformation and interpretation of information creating guided variation [168].

Language change can be generated by either process, as they act on different aspects of culture, or by a combination of both [167]. Nonetheless, biased transformation alone is not cumulative across generations, leaving cultural selection as the main driver of cultural evolutionary mechanisms in language evolution. Linguistic structure is under cultural selection from both iterated learning and communication, pressuring language to be compressible and distinctive [169]. Iterated learning over generations will favor simplicity and construct behavior in short descriptions, such as a concise grammar, and communication will shape expressiveness and clarity of language use [170].

A model created with a gene-culture coevolutionary framework showed language and biological evolution do coevolve, as in their experiments biological and cultural evolution alternated in their coevolutionary dynamics [171]. They also showed that the rate of cultural evolution is faster than the biological rate of evolution, in tune with other work affirming cultural evolution is faster [33,163]. Biological evolution is indeed unable to keep up with cultural evolution over short time scales, but on a longer time scale, culture rates are slower, which uncovers mutual directional effects [171].

The distribution of word-order features are strongly correlated across language families, indicating cultural evolution (in lineage specific histories) is the main driver of this linguistic structure [172]. This is another result that once again debunks the notion of language universals. Moreover, transition to agriculture drove language change as well, as the subsistence modification brought dietary changes that modified human bite, which resulted in the “f” and “v” sound innovations [173].

### *5.5. Language is shaped by social selection*

Population and demography are important aspects of evolution. It is at the population-level that adaptative forces are at play. Thus, social dynamics, such as social status and ties, and demographic history, such as population growth, decline, dispersal and bottleneck are all mechanisms driving evolution (Figure 1). The rate of language change depends on the frequency of word use, by natural or cultural selection, and sociodemographic factors, by social selection. Rates of language change can be affected by speaker population size, social network (i.e., social status and social ties) and language contact or isolation [131].

The Linguistic Niche Hypothesis (LNH) states languages adapt to the social niches in which they are used [174], in the same way species adapt to their ecological niche. LNH further separates languages into exoteric and esoteric, where exoteric languages spoken over large territories are pressured to be learnable by non-native L2 adults (inter-group) and esoteric languages are spoken by small populations with a shared sense of cultural identity (intra-group) [175]. In both concepts, the niche is an attribute of the population in relation to its environment and a constrainer of dispersion [176].

How population size effects the rate of language evolution has had mixed results, with plausible hypotheses for rapid spread of change in both small and large speaker populations [149,177–179]. There is a global correlation between population size and phoneme inventory size [180], as larger populations in Africa have languages with more phonemes than smaller language populations in South America and Oceania [181]. Atkinson proposed that just as genetic diversity decreases with distance from Africa, due to the serial founder effect, so does phonetic diversity in languages [181]. Subsequent work also found that low density populations lose phonemes with distance from Africa [182] and high density populations gain phonemes [183].

We've discussed how drift affected small populations with greater rates of cognate word loss from basic vocabulary, and in that same study, larger populations had greater rates of word gain [149]. Nonetheless, Bromham only analyzed Polynesian languages, where populations are geographically restricted and culturally similar. Another study tested the same relationship between population size and rates of word loss and gain in Austronesian, Indo-European and Bantu, which are three of the largest language families spread worldwide [177]. They found greater rates of word loss did happen in small populations speaking Indo-European languages, but found no relation for the Austronesian and Bantu languages or for word gain rate in large populations [177].

Languages in large speaking populations tend to be less morphologically complex, with a simpler grammar, than languages in small speaking populations [174]. Acquiring language from fewer speakers facilitates the learning of morphological complexity, supporting that an increase in population size, decreases languages' complexity [179]. Languages with a large population of speakers have a simpler structure, which also facilitates infant acquisition and increases L2 learners, consequently, increasing language fitness [174]. Larger populations also seem to develop more constant and systematic languages, which facilitates accuracy and comprehension [152].

Languages are subjected to a variety of situational contexts, which are social dependent, and adapts to them to avoid ambiguity in communication [141]. For instance, usage and interpretation of the word "mole", whether it refers to an animal or birthmark, is governed by the context in which it is used. Technological advances also change the contextual situation in which languages are used [89]. The invention of writing allowed long-distance communication and pressured language to be comprehensible in a lasting way, thus written language is more complex and harder to process to avoid ambiguity. The advent of the internet also pressured for a linguistic adaptation [89], but here, language had to adapt to long-distance communication in

real time, which is facilitated by internet slang, acronyms, and stand-ins for nonverbal communication, such as emoticons, emojis and gifs [184]. Nonetheless, their usage also depends on the online social context [185]. These adaptations do tend to infiltrate spoken language, as literacy instruction can improve speakers' language skills [186] and internet slang is often used in face-to-face conversations.

## 6. Macroevolutionary mechanisms of language evolution

### *6.1. A phylogenetic revolution*

Evolutionary trees, are at the core of parallelisms between language and species evolution [1]. They represent graphically the evolutionary history and relationship of a group through patterns of ancestry, descent and divergence [187]. They are inferred through shared traits or features of organisms. A tree has patterns of branching, indicating the persistence of a lineage, which bifurcate at nodes, indicating the last common ancestor, and ending in tips, specifying unique organisms (diversity) [188]. In the last decade, we saw a surge of studies focused on large-scale temporal, spatial and taxonomic patterns of language. They tried to answer questions on language classification, emergence, divergence, dispersal, diversity, distribution and conservation. These studies were greatly facilitated by phylogenetic statistical methods, which also revolutionized evolutionary biology since first introduced.

Evolutionary trees were traditionally constructed by the comparative method, widely used from anthropology to zoology [189,190]. In the comparative method, a researcher calculates the relationship of certain traits from multiple groups and later infers the evolutionary processes underlying the pattern found [191]. In biology this usually involved the comparison of two species occurring in different environments or the comparison of a trait across an environmental range [192]. In linguistics, trees were built mainly by establishing sound

correspondences, which were later used to infer sequences of sound changes and to identify any borrowing or shared innovations [193].

Biologists were first called out on this method in 1985, when Felsenstein revolutionized evolutionary biology by introducing phylogenetic thinking [191,192]. He addressed the issue of nonindependence in comparative methods and proposed a bootstrap approach to assess sampling error in phylogenies (because phylogenies are estimated). At that time, phylogenies were rare, used mainly in systematics or zoology, but some evolutionists did try to incorporate phylogenetic corrections in multispecies analyses [191]. Nonetheless, with the advancement of DNA sequencing technologies, phylogenies exploded in the literature accompanied by new statistical modelling methods and inference techniques [194]. Soon phylogenies penetrated every branch of biology, opening a new era for evolutionary, ecological and conservation studies.

Contemporarily, phylogenies are quantitatively inferred using four main methods: distance-based, maximum parsimony, maximum likelihood, and Bayesian inference [195], all with an optimality criterion to search for the best tree that fits the input data. They are easily implemented via software or programming packages [196–198], and can deal with different types of data, different evolutionary models and phylogenetic uncertainty [199]. These methods can model evolutionary processes with a constant rate (molecular clock), a varying rate (relaxed clock), a setting of branch lengths (Brownian motion, accelerating or decelerating rates), a choice of tree and parameter priors (if Bayesian), and the trees can be rooted or unrooted, weighted or unweighted, have binary or multiple branching (see [195,199] for an overview). Thus, phylogenies do more than just classification, they can also infer past diversification events, estimate speciation and extinction rates, infer the species or the clade age and distinguish between evolutionary processes.

The tips of phylogenetic trees can be species, populations, genes, languages, dialects, words (whether lexical, phonetic or grammar variants), and even other cultural artifacts, such as folktales [200] and skateboards [201]. Uncontroversial phylogenies, were the relationships and history are fully resolved, known or largely accepted are extremely rare. And that is fine, as phylogenies are hypothesizes of ancestry and relationships, and thus, are prone to errors and uncertainties [199].

The first quantitatively reconstructed linguistic phylogenetic tree was published in 2000 for 77 Austronesian languages [202]. Though many authors subsequently called for a phylogenetic approach to languages [5,193,203–206], linguists have generally shown skepticism towards computational phylogenetic methods. Critics of computational phylogenetic use reject the overall evolutionary approach to language [207,208], disagree with the models' assumptions [209,210], criticize cognate identification and the common use of lexical data to build the trees (instead of sound change) [205,211], and finally, point to horizontal transmission and word borrowing [212,213].

Tool-swapping and borrowing between fields are pretty common, even when the systems are different (i.e., game theory from economics to animal behavior to cultural dynamics) [7]. Models of evolution can and should be adaptable to the system being investigated, as to not violate assumptions. Most times, the differences between language and species evolution are simply not as important as expected by theory [7]. Any reconstructed phylogeny is a hypothesis about a relationship; thus, they are models. In fact, a famous aphorism in biostatistics is that "All models are wrong, but some are useful" [214]. The choice of model and methodological approach needs to depend only on whether that model performs the way the researcher intends: is it to mimic reality? to explore patterns? to make predictions? or to go back in time?

Further, there is no reason to assume lexical data is worse or unfit to build phylogenies compared to other aspects of language, as they too show variation, inheritance and fitness [215]. There are even advantages to using lexical data, such as non-binarity (decreases chance coincidences), universalities (i.e., a word for "mother"), produce areal isoglosses and borrowings are easier to identify [216]. Phylogenies built from basic vocabulary for three major language families are similar to the classifications inferred by linguists using phonology or grammar data [206]. Methodological comparisons between different phylogenetic methods also produce similar trees to the ones proposed by linguistic experts [216,217].

Wrong cognate detection, the distinction of shared innovation from shared inheritance, can lead to false relationship inferences. This is a shared problem, known as "phylogenetic non-independence" to biologists and "Galton's problem" to linguists [7]. To counter this, linguists sample across languages, a method that does not really solve the nonindependence problem [218]. Contrarily, Bayesian methods can control shared inheritance and will not compensate when there is no signal of relatedness [215]. Further, suitable data, with previous cognancy judgement, for 6,892 language and dialects were compiled into a dataset using machine learning methods [219] and other algorithms for automatic cognate classification are available in the LingPy python package [220].

Though biological representation is mainly tree-like, due to vertical transmission, even biological organisms do not follow a tree-like way. HGT and hybridization in species have pushed for a network model of connections, called phylogenetic networks [221], which has seen recent methodological advancements [222,223]. Phylogenetic networks can and have been applied to linguistic trees due to historical reticulation, such as language contact, creolization and word borrowing [224,225]. In fact, it was argued the phylogenetic network approach is more appropriate to model language evolution than strictly bifurcating the trees [225]. Nonetheless, phylogenetic inference has been shown to be robust even against high levels of

borrowing on a global and local scale [226]. Average borrowing rates in basic vocabulary are also low in most languages [227], and phylogenetic trees are constructed mainly from basic vocabulary. Overall, a small number of loanwords will not bias a phylogeny made with "big data", though they may sometimes bias the classification of a single language [215].

Even as the above methodological concerns are addressed [216,217,225,226,228,229], resistance to the phylogenetic methodology amongst linguists still exist, principally because most researchers using them are not trained in linguistics [230]. Nonetheless, there was a recent new surge of articles flaunting and praising the advantages (and disadvantages) of a computational phylogenetic approach to language and cultural evolution [7,215,231–233]. We expect any ensuing skepticisms to eventually subdue, especially given the facilities of large-scale linguistic data available in databases such as D-Place [234], ASJP [235], Glottolog [236], WALS [237] and the availability of trees with branch length in Newick format [238].

*6.1.1. Can't stop phylogenetic language trees*

Phylogenetic reconstructions were shown to produce accurate trees when compared to the classifications made by linguist experts [239]. Computational phylogenies have been used successfully to reconstruct relationships for many language families (Table 2). Far from being exhaustive, these represent only 15 out of the 424 estimated language families and isolates [240], and are all for a single language family or subgroups. Additionally, and often in the same study, language phylogenies have become fundamental to explore historical dynamics, including to infer the ancestral geographic location of a language family, to date the origin of a lineage, to investigate migration routes and language expansions [114,193,202,241–250], to date language divergences [251,252], to estimate evolution rates of linguistic variants [113,172,177,253–256], to study cultural evolution [232,257,258], to uncover kinship structures [259,260], and to understand how cultural traits evolved and co-evolved [261].

**Table 2.** Language families reconstructed through phylogenetic methods. This tables includes only phylogenetic reconstructions of the family's time depth or the ancestral homeland.

| Language family | Number of Languages [a] | Time Depth [b] | Ancestral Homeland | References |
|---|---|---|---|---|
| Arawak | 60 | - | Western Amazonia | [26] |
| Austronesian | 1268 | 5230 YBP | Taiwan | [27] |
| Bantu | 668 | 4000-5000 YBP | West central Africa (present-day Cameroon/Nigeria) | [28–30,101] |
| Dene-Yeniseian | 39 | - | Beringia | [31] |
| Dravidian | 80 | 4500 YBP | - | [32] |
| Indo-European | 449 | 8700 YBP [33,34] or 5000-6000 YBP [35] | Anatolia [33,34] or Steppe [35] | [33–36] |
| Japonic | 57 | 2400 YBP | Korean Peninsular | [37] |
| Pama-Nyungan | 306 | 4455-6996 YBP | Gulf of Carpentaria | [38,39] |
| Semitic | 25 | 5750 YBP | Levant | [40] |
| Sino-Tibetan | 500 | 7400 YBP | North Chinese Plateau | [41] |
| Turkic | 26 | 2408 YBP | - | [43] |
| Tupi | 48 | 3500 YBP | West-central Brazil (present-day Rondônia) | [44] |
| Tupí-Guaraní (Monophyletic subgroup of Tupi) | 30 | 2000-3000 YBP | Amazon | [45] |
| Uralic | 47 | 5300 YBP | - | [300] |
| Uto-Aztecan | 61 | 5000 YBP | Southern Mexico | [102,307] |

[a]Number of languages are according to their respective articles. [b]Time depths are estimations.

Perhaps one of the greatest contributions of the phylogenetic statistical approach to language evolution is the ability to *really* push back ancestry to a deeper time depth. Linguists seem to agree that the comparative method is limited to a time depth of 8,000 to 10,000 years [240,250,256,262], as it's estimated after that time it is impossible to distinguish cognates, chance resemblances, loanwords or borrowings in a language's vocabulary [204]. While most words do evolve quickly (on a macroevolutionary scale), with a 50% chance of being replaced every 2,000-4,000 years, frequently used words (numbers, pronouns, special adverbs) have greater fitness and evolve more slowly, being replaced every 10,000 or 20,000 years [131,263].

The discovery of these ultraconserved words allowed phylogenetic inference to push time depths further and uncover deep linguistic ancestry. For instance, the time depth for the Eurasiatic macrofamily (composed of seven language families) was estimated at ~15,000 years [256]. Phylogenetic inference above the family level has found support for seven macrofamilies, with greater clade confidence for the Eurasiatic, Austro-Tai and Mongolic + Tungusic macrofamilies [262]. Though no time depth was estimated, this evidence does support some macrofamilies previously argued for in the literature.

### *6.2. Language diversification: creating mutually unintelligible populations*

Phylogenies also paved the way to study evolutionary models, as they can be used to distinguish between gradualism and punctualism [264]. Phyletic gradualism, postulated first by Darwin, proposed species evolve gradually and slowly [265]. This gradualist view was dominant until the introduction of punctuated equilibrium, which proposed species underwent periods of long stasis with rapid and rare speciation events [266]. The punctuated equilibrium hypothesis was highly controversial amongst biologists, though it has amassed empirical support in the literature [267–270]. Nonetheless, both processes might trigger macroevolutionary change, as gradualism underlies constant-rate models of evolution, such as Brownian motion or random walk [271], and punctualism underlies models of stochastic processes with jumps, such as the Lévy process [269].

The dichotomy between the gradualist and punctuated models did reach language evolution. Newly formed languages were shown to diverge quickly in their vocabulary, in a punctuated burst manner, followed by a long period of stasis [272]. The percentages of total lexical divergence attributed to punctualism in three language families studied (31% in Bantu, 21% in Indo-European, and 33% in the Polynesian subclade, the Austronesian family being the exception with only 9.5%) were in range with the percentages found for genetic changes in

species, of approximately 22% [270,272]. This work was received with skepticism [273], and a study using branch length and a larger data set did not find support for punctuated equilibrium in languages [274]. Nonetheless, punctuational language evolution did eventually find empirical support in the literature [219] and was evidenced in cultural evolution as well [275].

Gradualist and punctuated models play directly into the question if species divergence occurs by gradual accumulated differentiation within a species or population (anagenesis) or by rapid burst of speciation events (cladogenesis) [271]. The evolution of a new species, or speciation, happens when a population is divided into two reproductively isolated populations. This usually occurs by allopatric speciation, sympatric speciation, parapatric speciation or hybridization [23]. Unsurprisingly, the processes a language becomes a new language are highly parallel to species speciation. In summary, the evolution of a new language happens when a speaking population is divided into two mutually unintelligible populations.

Anagenesis happens as a language may change without splitting. It was estimated a language becomes unintelligible within a time interval of 1,000 years (on average), at which point it would be incomprehensible between speakers at each temporal ends [240]. This explains why it is so hard for speakers of modern English to understand the original writings of Shakespeare. The English language has gone from “Old English” spoken from the 5$^{th}$ to 12$^{th}$ centuries, “Middle English” spoken until the end of 15$^{th}$ century, “Early Modern English” until the end of the 17$^{th}$ century, to finally the current “Modern English” [276]. A text from the mid 16$^{th}$ century, such as Shakespeare’s plays, is semantically difficult because the common words used in that period changed meaning and spelling, or completely disappeared from the modern language [277]. Contrarily, reading “Beowulf”, which was originally written in Old English, is impossible without a translation because it fundamentally is another language [276,278]. Though, of course, a lot of changes in English are due especially to historical language contact and loanwords.

Contrarily, cladogenesis requires lineage splitting. Geographically isolated populations will lead to allopatric speciation [279,280]. Two populations of the same species, or speakers of the same language, inaccessible to each other will lead to the emergence of two distinctive species or to two distinctive languages. The high language diversity found in the Austronesian family, which has nearly 1200 languages, is regularly explained by the island and archipelago peopling history, which increased isolation [202,281]. The diversity of the Japonic languages is also explained by island isolation, since the ocean acts as a geographic barrier [110]. Further, geographic isolation also increases language diversity by impeding language contact and creating isolates such as the Basque and Mirandese languages, whose populational isolation is evidenced genetically as well [282,283].

Sympatric speciation happens by ecological or reproductive separation in the same geographic area [284]. Populations of the same species in the same geographic area may partition their niche, where one forages by diving in shallow water and the other dives in deep water [285], a behavior which eventually leads to a lineage split. Similarly, speakers of the same language may occupy different social niches or be influenced by other group-enforcing factor [286], also leading to a linguistic split. Dialects emerge from social variables, as evidenced by sociolinguistics, and are transmitted and maintained by social niches. Dialects are in-group makers, through which an individual may express to be from a neighborhood or a working class. For example, a dialect can be easily associated with an elite, such as “the Queen’s English”, which is transmitted and enforced through boarding schools and elite universities [287].

Parapatric speciation happens in contiguous populations by geographic distance, offering a partial barrier [288]. A population occurring over a continuous range will diverge as distance inevitably separates subpopulations that occupy distinct niches [289]. Consequently, populations near each other will be more similar compared to the ones far away. This generates

a spatial pattern called a geographic cline, for species, and dialect continuum, for languages. The Western Romance continuum is a textbook example of a dialect continuum [24]. The continuum chains Portuguese dialects in the far west, to Spanish dialects, to French dialects and finally Italian dialects (Figure 2). Speakers of Portuguese will probably not easily understand speakers of Italian, but will have less trouble understanding speakers of Spanish.

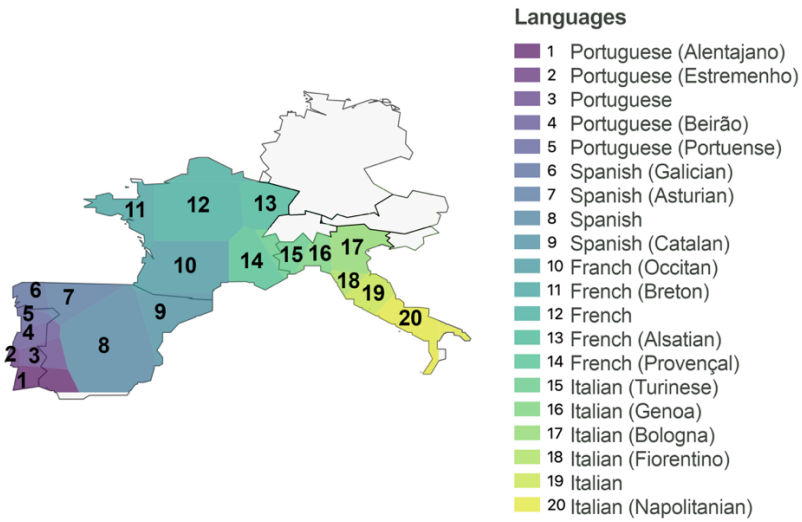


**Figure 2.** Dialect continuum for the Western Romance languages, chaining Portuguese, to Spanish, to French, to Italian. The continuum also generates hybrids, such as Galician (6), a Spanish dialect closely related to Portuguese, or Turinese, an Italian dialect closely related to French. Data obtained from WALS [237].

*6.2.1. Diversification by language contact*

Finally, and perhaps most interestingly, there are parallelisms between species hybridization and language contact in diversification processes. Thought to be rare (like HGT), molecular methods proved species hybridization to be quite common in both plants and animal species [290–292]. Of course, this statement may be unfitting depending on the definition of species you prefer. Different species meet, either in hybrid zones or by introduction of invasive species, and breed [293]. This results in a hybrid species, usually sterile or inviable beyond a single generation. For it to be able to stand on its own as a new distinct species, it needs to find other hybrids, to produce a progeny, and colonize an unoccupied niche, to be isolated from the

parental species [294,295]. Thus, hybridization is both a homogenizing and diversity force, capable of contributing to adaptation, speciation and extinction [292,294,295].

In contrast, language contact was never thought to be rare and is well documented throughout human history. In fact, multilingual communities have been argued to be necessary mechanisms for the emergence of language and are the primal outcome of language evolution [296]. Language contact can lead to language symbiosis, language shift, or Creole languages. Most languages are small in number of speakers and territory occupied [240]. These populations coexist, they engage in socioeconomic activities such as trading goods, social and ecological information sharing, and most importantly, spousal exchanges. Their languages also coexist without forming a mixed language [297,298].

Even distinct ethnolinguistic groups in the Amazon or the Sumba Island, commonly thought of as small and isolated, have recently been shown to maintain high contact as evidenced by their high genetic diversity [299,300]. For instance, indigenous populations in the Amazon are known to practice linguistic exogamy, marrying spouses from different language groups [298,301]. Spousal exchanges strengthen ties amongst populations and helps maintain group distinction, language symbiosis and multilingualism. In these communities, often a language will be deemed appropriate for one social context, but unsuitable for another [302]. Speakers then code-switch between languages, reflecting a political competition between the use of languages but also a high multilingual proficiency [303]. Code-switching is a process well documented for dialects as well [304,305]. Thus, bilingualism and multilingualism seem to be the norm, while monolingualism is a modern exception of big states and nations [240].

There can be, however, language competition for political and economic domination, and populations may be pressured to adapt linguistically to a changing socioeconomic environment [306]. As populations migrate, they can have different settlement styles, which sets the tone for how they interact with the native populations of the place they come to occupy

or colonize [306]. Thus, this interaction may lead to the displacement of the native languages in favor of the new dominant language, may lead to the emergence of a new language, and may lead to code-switching and pidgins.

Colonizers establishing settlement colonies intend to settle new homes [306]. They move with the clear purpose of occupying land and replacing the original population, creating a new world (which is the recreation of their old world in a new land) where their languages are dominant and indigenous languages are marginalized. Thus, indigenous languages are forced to shift, not because of lack of pride, but as an adaptative response to the new socioeconomic niche [306]. This is the case with the displacement of Gaelic, a language once spoken across Scotland that saw the number of its speaker decline since the nineteenth century, due to English socioeconomic dominance [307]. This historical trajectory parallels the displacement of indigenous languages in North America, also by the socioeconomic dominance of the English language.

Creole languages are natural languages, emerging from the contact between two or more languages [308]. A Creole language, like a hybrid species, is distinct from its parental languages and stands on its own, thus it has a complete vocabulary and grammar and is passed down in the population as a native language to children [309]. The emergence of Creoles is in no way exclusive to the European colonization period, as Old English is argued to be a creole originally, formed from the contact of Celts and Germanic colonizers [306]. The Roman provinces are also understood to have been creolized, rather than "Romanized" [310]. Nonetheless, plantation colonies in the 16$^{th}$ and 17$^{th}$ centuries provided the ultimate environment for the genesis of modern Creole languages [306,309]. For instance, a French plantation colony in the Caribbean put in contact African slaves, speaking many West African languages, but primarily Ewe and Fongbe, and European slave owners, speaking French, Spanish, Portuguese and English [311]. In that social context of anarchic multilingualism, Haitian Creole emerged. Haitian Creole is

today one of the official languages of Haiti, the other being French, and is spoken by over 10 million native speakers [311]. Overall, there are 92 described Creole languages [312], all highly diverse in structure and linguistic features [313].

Exploitation colonies, contrarily, implement an economic system where the colonizer's language is not necessarily a threat to the indigenous language. In these colonies, each sector of the economy (a social niche) uses a language, i.e., colonial languages in administrative economic activities and the indigenous or creoles languages in the other activities [306]. Thus, populations do not shift their languages either, but learn or create a new one. Similarly, pidgins arise specially in trade colonies, where these communities also have limited contact with their traders [308]. Pidgins are grammatically simplified languages used when a multilingual community has no language in common [308]. A pidgin is not considered a natural language, as contrarily to creoles, it is not taught as a native language to children nor does it have a fully developed vocabulary or grammar. In our analogy with hybridization, pidgins are comparable to sterile species. Pidgins are often considered as an initial phase for the development of Creoles [314,315], a statement that is greatly disputed [306,308,316,317].

Long-distance dispersal and founding are responsible for many speciation events, some of which include hybridization [290,292]. This speciation process proved equal for languages, though the social context may force different linguistic adaptations. For instance, 15-17$^{th}$ century settlement colonies (i.e., the United States) led to more language shift and replacement than exploitation colonies (i.e., Congo), while Creole languages emerged and persisted in the Caribbean but not in South America. Colonization, nonetheless, is a recurring theme in human history, the first form probably being of *H. sapiens* dominating and eventually eliminating *H. neanderthalensis* [318–321], and later of agriculturalists dominating and replacing hunter-gatherers. But before humans were able to colonize, they had to disperse.

### *6.3. Migration and dispersal: occupying new territory*

Dispersal and migration are fundamental macroevolutionary processes, since they influence population dynamics and are major contributors to diversification and speciation [322]. While dispersal refers to the movement of a group from a population to a new location, migration refers to the moving of the whole population [323]. Phylogenetic methods (such as dating with relaxed molecular clocks) has allowed the testing of different dispersal hypothesis, but specially lineage and geographic-focused inquiries on the mode, rate, timing and directionality of dispersals [322]. These methods are often applied to language evolution questions, especially for reconstructions of historical processes, such as past human dispersal, migration routes, the origin and the expansion of languages. The expansion of languages is analogous to the concept of dispersal, meaning as a population grows, it moves to occupy new territory [324].

Language phylogenies are even claimed to be preferrable to gene trees in studies where variables are culturally transmitted [93]. This happens because genes can be easily absorbed between populations without the culture, while migrants do adapt to the local culture, meaning language tracks culture transmission [93,244]. Nonetheless, there is correlation between genetic trees and language families of major populations, as evidenced by the first paper to triangulate genetic, archaeological, and linguistic data to study human history [325]. Both genes and words retain information from past demographic events, which can be accessed via genetic admixture or linguistic borrowings [326], thus language histories can be compared with other historical metrics.

Language-gene coevolution proposes a branching model, were language expansions and migration results in population splits and geographic and social isolation. Further studies, using computational phylogenetic methods, found genetic groupings of people often conforms to language grouping at a fine spatial scale, implying a language-gene coevolution where languages act as a barrier to gene flow [327–332]. However, other studies found no barriers to

gene flow or language contact, an exchange that has erased any patterns of populational branching [245,333,334]. This means the choice of tree type for peopling studies is dependable on whether the data is better to reconstruct the movement of people based on genes or based on language [93].

Serial founder-effect models have shown both genetic [335–337] and phonemic [181–183] diversities decline with distance from Africa, forming a clinal pattern. Both genes and words are susceptible to small population size, which tends to be a product of founder effect with a series of succeeding population bottlenecks from human migration. To be clear, phonemes are not lost due to bottleneck effect (like genes are), but undergo small speaker population pressures [181]. Phonemic diversity was further used to date the origin of modern language, an estimation between 244 kya, at the maximum, and 75 kya, at the minimum (here the highest and lowest estimates from many models) [338]. Together these results support an African origin of modern languages, congruent with the origin of modern humans and the out-of-Africa expansion [339].

Evidently, these works sparked great discussions across the multidisciplinary literature. The main criticism of the founder effect is that human migration of the past 50,000 years would overshadow the out-of-Africa migration, and point out particularly to the fact that genetic admixture is much more probable than genetic replacement [340,341]. In this same line of reasoning, it was argued historical events, high-rate phonemic change and horizontal transfers in language would mask any evidence of deep founder effects [342–344]. It was further claimed the linguistic data was inappropriate for the model [343,345], the same signal is not detectable in lexical and grammar patterns [345], a stronger correlation appears when the origin is central Asia [346] and that the current distribution of languages reflects Holocene rather than the Pleistocene expansions [347].

Phonemic and genetic data were shown to be indeed spatially autocorrelated on a global scale [348]. Despite this relationship between dispersal and linguistic variation and identifying Eurasia as an origin dispersion point, the authors rejected a founder effect explanation, and pointed to how phoneme inventory size is a "coarse summary statistic" [348]. Nonetheless, they acknowledge joining genetic, geographic and linguistic data is essential for insights into language evolution, a feat that language data alone would not otherwise provide [348]. Accordingly, even if there still is uncertainty on the origin and date of the "first" language, founder effect models can be applied to other language families. For instance, a founder effect has been implied for the Tupi family, since southern Tupian languages show less vowel variability then languages in the Central Amazon homeland [349].

Human migrations and dispersal are often triggered by technological advancements, such as fire-making, the invention of agriculture, and improvements in shipbuilding and navigation driving the movements of modern humans in the Pleistocene, of farmers in the Holocene and maritime expansions in the fifteenth century, respectively [258,350–352]. Nonetheless, abiotic factors, such as climate, topography and ecology, also play a hand in how and when routes are chosen. It is increasingly clear that human dispersion or migration is nonrandom, as humans prefer to move to and through known habitats and are constrained by barriers such as oceans, mountains and deserts [110,324].

### *6.3.1. Pleistocene movements*

Archaeological record has been able to recover a migration ordering where humans dispersed out of Africa in the Upper Pleistocene (128-12 kya) to the Arabian Peninsula, to Southwest Asia (the Levant) [353], to Southeast Asia [354], reaching Australia by 65,000 kya [355], and lastly occupying the Americas. These migrations are proposed to have happened in multiple waves and following multiple routes, but without a clear historical timeline. Linguistic

evidence, through phylogenies and language expansion assessments, has been able to provide pieces to the puzzle of both the Pleistocene and Holocene migrations.

There is a lot of uncertainty regarding the peopling of the Americas. Before the 1990s, it was believed people crossed the Bering Land Bridge to North America around 13,500 YBP and dispersed to South America through the interior of the continent [356]. Nonetheless, South America has much older artefacts and archeological sites than North America [357], and a higher number of language isolates [358]. Using language spread rates, Nichols showed the entry date in South America would have to be ~26,896 BP [250,358], in accordance with genomic evidence which placed the entry at ~23,000 BP [359].

Further, genetic and climatic data suggest at least three migration waves into North America [360,361]. The earliest being around 16-40,000 YBP for the generic Paleoindian group, believed to be the wave with the greatest language groups. The second migration wave is of the Na-Dene language family around 14,000 YBP and the third wave is of the Eskimo-Aleut family around 9,000 YBP [360]. Linguistic phylogenies, evidencing a Na-Dene and Yeniseian linguistic connection, even showed these migrations are not unidirectional, supporting Na-Dene dispersals out-of-Beringia both back into Siberia and into North America [362].

### *6.3.2. Holocene movements*

The Farming/Language Dispersal Hypothesis (FLDH) proposes migrations and dispersals of farming techniques at the beginning of the Holocene shaped the distribution of modern languages [258]. In these expansion scenarios, farmers eventually incorporated and replaced the hunter-gatherers that were already there since the Pleistocene expansions. Phylogenetic methods, alongside archaeological evidence, has contributed to great advancements in uncovering these migration histories and testing specifically for FLDH.

Branching order of Bantu languages overlap with the archaeological record for the spreading of farming between 3,000-5,000 YBP in sub-Saharan Africa [242]. Bantu genetic and linguistic distances are correlated, favoring a southern migration route [245]. It was later confirmed the Bantu speakers preferred to disperse to known habitats, since they moved exclusively along a savannah corridor and considered the rainforest as a barrier [244]. Populations did disperse from the savannahs to the rainforest, but this process was slow, delaying their expansion by 300 years [244]. When Bantu populations emerged on the south side of the rainforest, they moved south and west, branching into the East and West Bantu distinct lineages [248] around ~2,000 YBP. Though the dispersal routes through the rainforest in both studies are somewhat divergent.

Dated phylogenies provided evidence for the Sino-Tibetan farming origin in northern China around 7,200 YBP [363]. They identified six cognate domesticate names (i.e., millet, rice, pig, cattle) and crossed them with the archaeological record, which showed agriculture probably played a boosting role in the family language expansion. Farmers dispersing out of China kept on the move, settling in Taiwan and originating Austronesian languages, which later expanded in pulses and pauses throughout the Pacific around 5,230 YBP [114,202]. The Japonic origin is also correlated with the arrival of the first farmers in Japan coming from China through the Korean Peninsula around 2,400 YBP [364].

There are two competing hypotheses for the Indo-European language expansion, both depend on agriculture (be it farming spread or horse domestication), but one proposes the family origin was in Anatolia (in present-day Turkey) and another proposes it was in the steppes north of the Black Sea [258]. Statistical support of the Anatolian hypothesis, where a Indo-European expansion accompanied agriculture spread out of Anatolia around 8,000-9,500 YBP, was first provided by [365] and later by [366]. Statistical support for the steppe hypothesis, where Kurgan horse-riding pastoralists dispersed into Europe around 5,000-6,000 YBP, was

also found [241]. Nonetheless, these two hypotheses are not mutually exclusive, as the Celtic, Germanic, Italic, Balto-Slavic and Indo-Iranian subfamilies all emerged as distinct lineages around 4,000-6,000 YBP, in congruency with the Kurgan expansion [366].

However, investigation of Uto-Aztecan speakers in Mesoamerica introducing maize agriculture (and their language) in the North American Southwest around 4,000 YBP did not support the FLDH [367]. Given their results, this scenario would be possible only if the dispersing Uto-Aztecan speakers were all males, thus not a demographic expansion. It was speculated an in-situ population expansion ~2,105 YPB may have blurred the genetic relationship between these populations [367].

Finally, FLDH has not been formally tested for family languages in South America, but the phylogeography of Arawak expansion did hint at it. Arawak homeland was placed in Western Amazonia, which suggests manioc cultivation to be a probable driver of their expansion [368]. Arawak dispersed in many directions, first to the south following the Madeira and Purus rivers, then north to the Circum-Caribbean, then moving to Central Brazil, Central Amazon and finally up the Rio Negro to Northwest Amazonia. The Northwest Amazonia was the last clade to diverge and it is presently the region with greatest Arawak language diversity [368].

Dispersals routes for other language families, unrelated to agriculture, have also been uncovered. The Tupi origin is placed in west-central Brazil [261], with the expansion estimated to be at 2,000-6,000 YBP and following a radial pattern, with subsequent isolation [369]. These expansions showed probable bottleneck effects, with cultural complexity declining through time [261]. The Je family is also originally from Central Amazon, the linguistic time-depth between Tupi and Je is of 7,000 – 5,000 years BP [369]. The Je seemed to have dispersed randomly, in a process that did not leave compatible genetic, geographic and historical scenarios [369].

### *6.4. Macroecology: uncovering large-scale geographic patterns*

Macroecology, a subfield in ecology, also answers questions about diversity, but through the assessment of large-scale spatial, temporal and taxonomic patterns [370,371]. Macroecology seeks explanations or mechanisms underlying large spatial patterns, often by working with polyphyletic groups (i.e., all birds vs. only passerines), by estimating richness and abundance metrics, functional and phylogenetic diversity indices, range sizes, habitat preferences and climate variation in space [372]. The high diversity of the ~7,000 different languages in the world, most of which are small in both population and territory size, and are distributed non-randomly in the geographic space caught the eye of macroecologists, who are successfully applying statical macroecological approaches to language diversity.

The first time all of the world's languages were completely mapped was with the publication of the "Atlas of the World's Language" in 1994 [373]. The congruency of the linguistic patterns with biodiversity patterns, both showing higher diversity towards the tropics, did not go unnoticed. Soon researches started seeking empirical explanations for the languages' asymmetrical distribution, calculating correlations between abiotic variables and language diversities in North America and West Africa [374,375]. Both approaches propose distinct hypothesis, something that set the tone for future work.

Nettle proposed the Ecological Risk Hypothesis (ERH), where areas of greater productivity could provide for more distinct groups contrary to areas of low productivity which pose a greater thread to subsistence and requires larger populations to ensure survival [375]. He inferred ecological risk through the mean growing season, calculated as "the number of months in which the monthly rainfall is greater than twice the monthly temperature" [375]. He later tested the ERH for global language patterns, also finding a positive correlation [376]. Many authors have since tested the ERH, including productivity variables in their analysis be

it through the mean growing season [377–382], land suitability for agricultural production [378,383–385] or soil fertility [381].

Noticing the Latitudinal Diversity Gradient (LDG), Mace and Pagel proposed language diversity could be explained by ecological diversity, through habitat heterogeneity [374]. The LDG found for mammals in North America were at the time correlated to habitat diversity, with greater species diversity and smaller species range in the south than in the north. They also found a positive correlation for language and habitat type, as language diversity increases with greater habitat diversity [374]. Thus, subsequent work also tested for correlations with habitat heterogeneity, through variables such as animal [379,386–388] and plant species richness [377,387], plant productivity [389], altitude [377–381,383,385,386,390], river density and distance to the nearest body of water [378,383,390] and number of ecoregions [381].

You will notice many authors tests for both ERH and habitat heterogeneity in the same study, often finding no support for neither ERH [377–381,383,385,390] or habitat heterogeneity [377–379,381–383,387,388,390,391]. These mixed results occur specifically because many of these variables are themselves correlated with latitude or with each other, rather than being causal of language diversity. Nonetheless, these studies have helped narrow the main mechanisms for language diversification to be climate, topography, biodiversity and sociocultural factors [111,382,392].

Given that language diversity is shaped by multiple processes, a mechanistic approach has been called for [392]. Correlative methods only statistically associate diversity with the predictor variables, thus it is difficult to infer causality, control spatial and phylogenetic autocorrelation or explore nonlinear relationships [392]. Mechanistic approaches, such as simulation modeling, can provide explicit tests of diversification mechanisms and have been used to uncover biodiversity patterns [393–396].

Mechanistic models were developed for languages as well, providing evidence that environmental carrying capacity (limiting group size) is an important process driving patterns of language diversity in Australia [112] and North America [111]. The Australian model also identified precipitation as a driver of diversity, and overall, their model explained 56% of the continent's language diversity. The indirect effects of topographic, climatic and demographic variables were further assessed through a path analysis model [111]. The strongest direct effects were for population density, carrying capacity and ecoregion richness. Nonetheless, their model explains from 86% to less than 40% of the North American language diversity, showing the drivers of language diversity are not universal in the continent or even directly associated to linguistic factors [111].

Despite being the most studied spatial pattern, there is no consensus on the underlying causes of the LDG. The literature accumulates at least 26 hypotheses, all compatible with the observed data [397]. Unsurprisingly, a mechanistic modelling approach was also called for studies investigating the LDG [397], six years after this necessity was recognized for language patterns. Perhaps the progress of language studies lies indeed with their multidisciplinary characteristic, addressing challenges and pushing the field forward in a pioneering way.

### *6.5. Extinction: when the last speaker dies*

Most of the species that once existed on Earth are now extinct, with only one out a thousand surviving into present-day [398]. The Earth is approximately 4.567 billion years old, experiencing 5 major mass extinction events in the last 439 million years, brought about by sudden changes in geological conditions [399]. A mass extinction is an exceptional event, defined by a large eradication of species in a short time span (Jablonski 1986). However mass extinctions account for only 5% of all extinctions [400]. Background extinction is the natural extinction of a species over a gradual period of time [401], though background extinctions are

also an infrequent event on the human timescale. Background extinction (henceforth extinction) is a process, involving both geological and biological time, of a long-term generational loss of reproductive fitness [401].

The “official” extinction of a species happens when its last individual dies. Much in the same way, a language death happens when the last native speaker dies. Linguists use the term “language death”, though there is some mention of “language extinction” in the literature, a probable influence of studies comparing modern species and language extinctions. But before species and languages disappear, species populations have to stop transmitting their genes, and a population of speakers have to stop teaching their language. For both organisms, this usually happens when populations are small in size and territory occupied, facilitating their slip into an extinction vortex [402]. Extinction vortex postulates that as population declines, population dynamics also deteriorate. They are thus probable to keep declining due to their small size, smaller range size and vulnerability to stochastic events [402]. For instance, genocides, when a whole population is exterminated all at once, are genuinely rare for both humans [306] and species. Nonetheless, when populations are small, a genocide becomes more probable and could result in a linguicide or ecocide.

Although species extinction is inevitable and often a driver of speciation, mechanisms of extinction have been neglected by biologists [401]. In this regard, the mechanisms behind language death are much better understood. Language is often thought of as a parasite, because it could not exist without people, their “hosts”, while people could exist without language [87,403]. Thus, languages die when their speaking population dies. For small populations, because of extinction vortex, this could happen by natural disasters, such as hurricanes, floods, volcanic eruptions and drought, by famine, diseases, war and by outside exploitation [404]. Language may also die by socioeconomical factors brought by the movement of other

populations, as discussed previously. In this situation, even as the native speakers live, their language dies.

Modern speaking populations range from approximately 40 speakers to over 1 billion speakers [240]. Languages with less than 40 speakers are considered unstable, perhaps already without intergenerational transmission [240], and languages with less than 1,000 speakers are generally endangered [312]. But speaker populations do tend to be small and spoken in small areas, for instance, half of the world's population speak only 23 out of ~7,139 languages [312]. Ecology and social niche are often factors that determine language death or maintenance in relation to small populations. A rural population of 200 individuals speaking a language is not the same as 200 individuals speaking a language inside a metropolis, where another dominant language operates [404]. Thus, there is never just one reason for a language death, it is a gradual process that effects populations in different ways [404].

### *6.5.1. Unnaturally fast extinctions*

Background extinction rates are used to compare current extinctions rates with the past. This is why authors allude to a sixth mass extinction event, since current vertebrate and invertebrate extinction rates are much higher than expected by the background extinction rates [405,406]. To put it in numbers, approximately 200 vertebrate species have gone extinct in the last 100 years, roughly two species every year. Compared to the background extinction rates, the expected was for these 200 species disappear in 10,000 years [407]. These modern rates are much higher exclusively because of anthropogenic action. No ecosystem is free of human influence, as human activity alters the soil surface, disrupts the biogeochemical cycles, introduces exotic species, removes native species, causes habitat loss and fragmentation, pollutes habitats, overexploit resources, and ultimately, changes the climate [408,409].

In our final parallelism, current language extinction risk is also driven by anthropogenic action. Historically, these languages were greatly decimated during colonial times. In the 21st century, remaining Indigenous and traditional communities are deeply affected by gold-mining and deforestation in the Amazon [410], by lack of land rights, by forced resettlement, by little or no access to justice, by discrimination, marginalization and violations of cultural rights [411]. More recently, indigenous and traditional communities in countries such as Brazil were greatly impacted by COVID-19, without any government guidelines to fight the disease [412].

Calculations of how languages deviate from a hypothetical stability situation have been made, showing language diversity declined 20% globally from 1970 to 2005 [413]. Regionally, that percentage gets much higher as linguistic diversity declined over 60%, 30% and 20% in the Americas, the Pacific, and Africa, respectively [413]. For illustrative purposes, the number of languages found in present-day Brazil at the time of Portuguese arrival in 1500 was estimated to be around 1,175 languages, a number that has since been reduced to ~180 languages [414]. This means 1,000 languages disappeared in only 500 years, also coming down to two languages every year. Thus, it is safe to conclude both species and languages extinction are happening unnaturally fast.

The IUCN Red List assesses species extinction threat using a three-risk component: small geographic range (<20 $km^2$), small population size (<1,000 individuals) and rapid population decline [415]. The same criteria were applied to languages, and they identified 291 languages in areas smaller than 20 $km^2$, 1,492 languages with less than 1,000 speakers and 193 (out of 649 languages with this information) languages declining in native speaker numbers [416]. Geographically, small range size and small speaker populations were found in the tropics and the Arctic regions, while declining speaker growth rates tended to be higher in temperate zones and desert areas in Africa and west Asia [416].

Other geographic assessments have placed languages in regions of biodiversity hotspots, which are regions of both high endemic species and loss of over 70% of the natural habitat [417]. Their analyses showed 3,202 languages are currently distributed in 35 hotspots and 1,622 languages are in regions of high biodiversity wilderness across the world [418]. This further means that 70% of the world's languages are distributed in only 24% of the Earth's surface. Nonetheless, a spatial congruence between threatened species and languages was not found in the island of New Guinea, the place with the highest language diversity in the world [391]. But the importance of islands to conservation was highlighted in these comparisons, showing they hold 37% of all critically endangered species and 25% of the critically endangered languages [419].

Species conservation is often costly and deemed as frivolous by governments. Thus, biologists are trained to think in cost-effective conservation measures, such as concentrating efforts in creating protected areas in hotspots to protect the greatest number of threatened species with one action. The confirmation that areas of high biological diversity are also areas of high linguistic diversity emphasizes the clear benefits these populations have on maintaining biodiversity [420]. Traditional populations do develop a bond with the surrounding natural environments, and population's values on nature are often linked to their language identity [421]. Biocultural approaches to conservation focus on both biological and cultural diversity, through a co-management, community-based conservation and integrated conservation and development [422]. These approaches focus on the conservation of biocultural aspects, traditional ecological knowledge, recommendations for actions and relational and intrinsic values [423].

But merely acknowledging a role of indigenous and traditional populations in biodiversity conservation, environment management and ecological restoration is not enough. These communities are regularly treated as passive actors, often included in the conversation

as an afterthought. For instance, studies proposing biocultural approaches to conservation rarely address governance, collaboration, empowerment, power and gender issues [423]. Actively including indigenous and traditional communities in conservation planning avoids disastrous conservation management strategies such as the distortion of cultural practices, prohibition or limiting of resource use, or population removal and resettlement [422,424]. All these strategies reduce indigenous and traditional populations engagement with the environment, restraining their culture and ecological knowledge, which results in both cultural and diversity loss [424].

Lastly, government policies and resettlement programs are the biggest thread to language diversity. Thus, securing the rights of indigenous and traditional communities should be a major priority of conservation organizations [425]. This guarantees access and control of their traditional territories, as it will settle land disputes with farmers and the overexploitation of miners and loggers. Currently, indigenous territories represent 37% of remaining natural lands and intersects ~40% of all protected areas in the world [426]. Thus, even without a biocultural approach to conservation, ensuring Indigenous' right to their land is essential to meeting local and global biodiversity conservation goals [426].

## 7. Concluding remarks

The evolution of language and species have been compared infinite times, by multiple authors in different fields, by different analogies and commonly as mere illustrations. Here we have exhaustively shown that language evolution is in fact evolutionary. Thus, any method from evolutionary biology is applicable to language evolution and parallelisms are more than a useful analogy, since language and species often evolve by the same mechanisms. Language is a bio-cultural hybrid, therefore, understanding how biological, cultural and social evolutionary processes interplay and reinforce each other is key to understand language evolution.

These processes have distinct selective pressures and operate on a microevolutionary scale, influencing language change on the individual and populational-level, and on a macroevolutionary scale, influencing the emergence of new languages. But microevolution alone does not explain macroevolution [105]. Patterns of species and language diversity are the result of dispersal, interaction, speciation and extinction. Diversification processes are clearly evidenced by dialects and subspecies, considered an intermediate level where the mechanisms of change occur [286]. Thus, diversity, whether biological or linguistic, is a reflection of evolution.

Language evolution on a macroevolutionary scale is well resolved in the literature due to three decades of intensive multidisciplinary research. There is value to the analogies of genes and words and virus and words on a microevolutionary scale, though we recognize there are limits to both analogies. Perhaps language evolution is Darwinian only at the populational level, but more research is surely needed in this regard. Nonetheless, heritage needs not to be only organic: memetic theory (though frowned upon) is Darwinian.

The combination of vocal and auditory learning, speech production, the expression of complex concepts and context-dependent semantic interpretation of signals into a language allows us to convey infinite different messages [67]. To answer questions on how languages evolve we can investigate the past through inquiries on the rate and tempo of language evolution, migration routes, bio-cultural adaptations, environmental correlations, uncover ancestral languages, and test drivers and mechanisms of diversity.

Reverse engineering is surely a tool moving forward: if we are able to recreate the evolutionary steps of an existing pattern or relation, then we will start to collect and accumulate pieces of the puzzle. In this regard, phylogenetic and macroecological approaches provide powerful methods to achieve a comprehension of human cultural diversity and to address many of these knowledge gaps. But while this emerging field is strongly empirical, it still needs

theory from historical linguists and anthropologists to guide their models and hypothesis. Possibly, language evolution is a success case of multidisciplinary challenges.

## References

1. van Wyhe J. The descent of words: evolutionary thinking 1780–1880. Endeavour. 2005;29:94–100.
2. Warner B. Charles Darwin and John Herschel. S Afr J Sci. 2010;105:432–9.
3. Pagel M. Darwinian perspectives on the evolution of human languages. Psychon Bull Rev. 2017;24:151–7.
4. Whitfield J. Across the curious parallel of language and species evolution. PLoS Biol. 2008; 6:1370–2.
5. Atkinson QD, Gray RD. Curious parallels and curious connections - Phylogenetic thinking in biology and historical linguistics. Syst Biol. 2005;54:513–26.
6. Mendívil-Giró J-L. Languages and Species: Limits and Scope of a Venerable Comparison. In: Rosselló J, Martin J, editors. Biolinguistic Turn Issues Lang Biol. Universitat de Barcelona; 2006. p. 1–38.
7. Bromham L. Curiously the same: swapping tools between linguistics and evolutionary biology. Biol Philos. Springer Netherlands; 2017;32:1–32.
8. Knight C, Studdert-Kennedy M, Hurford JR. Language: A Darwinian Adaptation? In: Knight C, Studdert-Kennedy M, Hurford JR, editors. Evol Emerg Lang Soc Funct Orig Linguist Form. Cambridge: Cambridge University Press; 2000. p. 1–15.
9. Heijdra M. Darwinian Explanations of the Origin of Language. 2009;
10. Chomsky N. Syntatic structures. The Hague, Netherlands: Mouton; 1957.
11. Pinker S, Bloom P. Natural language and natural selection. Behav Brain Sci. 1990;13:707–27.
12. Dediu D, De Boer B. Language evolution needs its own journal. J Lang Evol. 2016;1:1–6.
13. Steels L. Modeling the cultural evolution of language. Phys Life Rev. Elsevier B.V.; 2011;8:339–56.
14. Gong T. Where could biolinguists and evolutionary linguists meet?. Comment on “Modeling the cultural evolution of language” by Luc Steels. Phys Life Rev. Elsevier B.V.; 2011;8:373–4.
15. Mufwene SS. An ecological account of language evolution! Way to go!. Commentary on “Modeling the cultural evolution of language” by Luc Steels. Phys Life Rev. Elsevier B.V.; 2011;8:367–8.
16. Fitch WT. Biological versus cultural evolution: Beyond a false dichotomy. Comment on “Modeling the cultural evolution of language” by Luc Steels. Phys Life Rev. Elsevier B.V.; 2011;8:357–8.
17. de Boer B. Modeling evolution of speech. Comment on “Modeling the cultural evolution of language” by Luc Steels. Phys Life Rev. Elsevier B.V.; 2011;8:361–2.
18. Hashimoto T, Konno T. Language origin from simulation of language evolution. Comment on “Modeling the cultural evolution of language” by Luc Steels. Phys Life Rev. Elsevier B.V.; 2011;8:365–6.
19. Croft W. Social factors in the cultural evolution of language. Comment on “Modeling the cultural evolution of language” by Luc Steels. Phys Life Rev. Elsevier B.V.; 2011;8:359–60.
20. Baronchelli A. The maturity of modeling. A comment on “Modeling the Cultural Evolution of Language” by Luc Steels. Phys Life Rev. Elsevier B.V.; 2011;8:377–8.
21. Perlovsky L. Abstract concepts in language and cognition. Commentary on “Modeling the cultural evolution of language” by Luc Steels. Phys Life Rev. Elsevier B.V.; 2011;8:375–6.
22. Sproat R. A computational model of the discovery of writing. Writ Lang Lit. 2017;20:194–226.
23. Ridley M. Evolution. 3rd ed. Wiley-Blackwell; 2003.
24. Crystal D. The Cambridge Encyclopedia of Language. 3rd ed. Cambridge: Cambridge University Press; 2010.
25. Godfrey-Smith P. Conditions for Evolution by Natural Selection. J Philos. 2007;104:489–516.
26. Fitch WT. Animal cognition and the evolution of human language: why we cannot focus solely on

communication. Philos Trans R Soc B Biol Sci. 2020;375:20190046.
27. Vigliocco G, Perniss P, Vinson D. Language as a multimodal phenomenon: implications for language learning, processing and evolution. Philos Trans R Soc B Biol Sci. 2014;369:20130292.
28. Berwick RC, Friederici AD, Chomsky N, Bolhuis JJ. Evolution, brain, and the nature of language. Trends Cogn Sci. Elsevier Ltd; 2013;17:89–98.
29. Evans N, Levinson SC. The myth of language universals: Language diversity and its importance for cognitive science. Behav Brain Sci. 2009;32:429–48.
30. Kirby S. Culture and biology in the origins of linguistic structure. Psychon Bull Rev. 2017;24:118–37.
31. Bagno M. Preconceito Linguístico. 56th ed. São Paulo: Parábola Editorial; 2015.
32. Dunbar RIM. Gossip in evolutionary perspective. Rev Gen Psychol. 2004;8:100–10.
33. Laland KN, Odling-Smee J, Feldman MW. Niche construction, biological evolution, and cultural change. Behav Brain Sci. 2000;23:131–75.
34. Steels L. Do languages evolve or merely change? J Neurolinguistics. Elsevier Ltd; 2017;43:199–203.
35. Prat Y. Animals Have No Language, and Humans Are Animals Too. Perspect Psychol Sci. 2019;14:885–93.
36. ten Cate C, Okanoya K. Revisiting the syntactic abilities of non-human animals: natural vocalizations and artificial grammar learning. Philos Trans R Soc B Biol Sci. 2012;367:1984–94.
37. Poole JH, Tyack PL, Stoeger-Horwath AS, Watwood S. Elephants are capable of vocal learning. Nature. 2005;434:455–6.
38. Prat Y, Azoulay L, Dor R, Yovel Y. Crowd vocal learning induces vocal dialects in bats: Playback of conspecifics shapes fundamental frequency usage by pups. PLoS Biol. 2017;15:1–14.
39. Sharpe DL, Castellote M, Wade PR, Cornick LA. Call types of Bigg's killer whales (Orcinus orca) in western Alaska: using vocal dialects to assess population structure. Bioacoustics. Taylor & Francis; 2019;28:74–99.
40. Martins BA, Rodrigues GSR, de Araújo CB. Vocal dialects and their implications for bird reintroductions. Perspect Ecol Conserv. Associação Brasileira de Ciência Ecológica e Conservação; 2018;16:83–9.
41. Keighley M V., Heinsohn R, Langmore NE, Murphy SA, Peñalba J V. Genomic population structure aligns with vocal dialects in Palm Cockatoos (Probosciger aterrimus); evidence for refugial late-Quaternary distribution? Emu. Taylor & Francis; 2019;119:24–37.
42. Seyfarth RM, Cheney DL, Marler P. Vervet monkey alarm calls: Semantic communication in a free-ranging primate. Anim Behav. 1980;28:1070–94.
43. Berwick RC, Okanoya K, Beckers GJL, Bolhuis JJ. Songs to syntax: the linguistics of birdsong. Trends Cogn Sci. Elsevier Ltd; 2011;15:113–21.
44. Riley JR, Greggers U, Smith AD, Reynolds DR, Menzel R. The flight paths of honeybees recruited by the waggle dance. Nature. 2005;435:205–7.
45. von Frisch K. The Dance Language and Orientation of Bees. Cambridge, Massachusetts: Harvard University Press; 1967.
46. Gick B, Wilson I, Derrick D. Articulatory Phonetics. Malden, MA: Wiley-Blackwell; 2013.
47. Conant D, Bouchard KE, Chang EF. Speech map in the human ventral sensory-motor cortex. Curr Opin Neurobiol. 2014;24:63–7.
48. Ogden R. Swallowing in Conversation. Front Commun. 2021;6:1–19.
49. Arbib MA. Toward the Language-Ready Brain: Biological Evolution and Primate Comparisons. Psychon Bull Rev. Psychonomic Bulletin & Review; 2017;24:142–50.
50. Lieberman P. Vocal tract anatomy and the neural bases of talking. J Phon. Elsevier; 2012;40:608–22.
51. Fagot J, Boë L-J, Berthomier F, Claidière N, Malassis R, Meguerditchian A, et al. The baboon: A model for the study of language evolution. J Hum Evol. 2019;126:39–50.
52. Boë L-J, Sawallis TR, Fagot J, Badin P, Barbier G, Captier G, et al. Which way to the dawn of speech?: Reanalyzing half a century of debates and data in light of speech science. Sci Adv. 2019;5:eaaw3916.
53. Hauser MD, Chomsky N, Fitch WT. The faculty of language: what is it, who has it, and how did it

evolve? Science (80- ). 2002;298:1569–79.
54. Fedurek P, Slocombe KE. Primate vocal communication: A useful tool for understanding human speech and language evolution? Hum Biol. 2011;83:153–73.
55. Chomsky N. The language capacity: architecture and evolution. Psychon Bull Rev. Psychonomic Bulletin & Review; 2017;24:200–3.
56. Jarvis ED. Evolution of vocal learning and spoken language. Science (80- ). 2019;366:50–4.
57. Rodríguez L, Cabo L, Egocheaga JE. Breve nota sobre el hioides neandertalense de Sidrón (Piloña, Asturias). In: Aluja MP, Malgosa A, Nogués RM, editors. Antropol y Divers. 1st ed. Barcelona: Edicions Bellaterra; 2003. p. 484–93.
58. Arensburg B, Tillier AM, Vandermeersch B, Duday H, Schepartz LA, Rak Y. A Middle Palaeolithic human hyoid bone. Nature. 1989;338:758–60.
59. D'Anastasio R, Wroe S, Tuniz C, Mancini L, Cesana DT, Dreossi D, et al. Micro-biomechanics of the Kebara 2 hyoid and its implications for speech in Neanderthals. PLoS One. 2013;8:6–12.
60. Conde-Valverde M, Martínez I, Quam RM, Rosa M, Velez AD, Lorenzo C, et al. Neanderthals and Homo sapiens had similar auditory and speech capacities. Nat Ecol Evol. 2021;5:609–15.
61. Dediu D, Levinson SC. Neanderthal language revisited: not only us. Curr Opin Behav Sci. Elsevier Ltd; 2018;21:49–55.
62. Morgan TJH, Uomini NT, Rendell LE, Chouinard-Thuly L, Street SE, Lewis HM, et al. Experimental evidence for the co-evolution of hominin tool-making teaching and language. Nat Commun. Nature Publishing Group; 2015;6:4–11.
63. Alemseged Z, Spoor F, Kimbel WH, Bobe R, Geraads D, Reed D, et al. A juvenile early hominin skeleton from Dikika, Ethiopia. Nature. 2006;443:296–301.
64. Martínez I, Arsuaga JL, Quam R, Carretero JM, Gracia A, Rodríguez L. Human hyoid bones from the middle Pleistocene site of the Sima de los Huesos (Sierra de Atapuerca, Spain). J Hum Evol. 2008;54:118–24.
65. White TD, Asfaw B, DeGusta D, Gilbert H, Richards GD, Suwa G, et al. Pleistocene Homo sapiens from Middle Awash, Ethiopia. Nature. 2003;423:742–7.
66. Tattersall I. The material record and the antiquity of language. Neurosci Biobehav Rev. Elsevier Ltd; 2017;81:247–54.
67. Fitch WT. Empirical approaches to the study of language evolution. Psychon Bull Rev. Psychonomic Bulletin & Review; 2017;24:3–33.
68. Żywiczyński P. How research on language evolution contributes to linguistics. Yearb Pozn Linguist Meet. 2019;5:1–34.
69. Kolodny O, Edelman S. The evolution of the capacity for language: the ecological context and adaptive value of a process of cognitive hijacking. Philos Trans R Soc B Biol Sci. 2018;373:20170052.
70. Dunbar R. Grooming, Gossip and the Evolution of Language. Cambridge, MA: Harvard University Press; 1996.
71. Sommerfeld RD, Krambeck H-J, Milinski M. Multiple gossip statements and their effect on reputation and trustworthiness. Proc R Soc B Biol Sci. 2008;275:2529–36.
72. Jolly E, Chang LJ. Gossip drives vicarious learning and facilitates social connection. Curr Biol. Elsevier Ltd.; 2021;31:2539-2549.e6.
73. Cuesta JA, Gracia-Lázaro C, Ferrer A, Moreno Y, Sánchez A. Reputation drives cooperative behaviour and network formation in human groups. Sci Rep. 2015;5:7843.
74. Piazza J, Bering JM. Concerns about reputation via gossip promote generous allocations in an economic game. Evol Hum Behav. 2008;29:172–8.
75. Beersma B, Van Kleef GA. How the Grapevine Keeps You in Line. Soc Psychol Personal Sci. 2011;2:642–9.
76. Feinberg M, Willer R, Stellar J, Keltner D. The virtues of gossip: Reputational information sharing as prosocial behavior. J Pers Soc Psychol. 2012;102:1015–30.
77. Semaw S, Renne P, Harris JWK, Feibel CS, Bernor RL, Fesseha N, et al. 2.5-million-year-old stone tools from Gona, Ethiopia. Nature. 1997;385:333–6.
78. Montagu A. Toolmaking, hunting, and the origin of language. Ann N Y Acad Sci. 1976;280:266–74.
79. Isaac GL. Stages of cultural elaboration in the pleistocene: possible archaeological indicators of the

development of language capabilities. Ann N Y Acad Sci. 1976;280:275–88.
80. Stout D, Chaminade T. The evolutionary neuroscience of tool making. Neuropsychologia. 2007;45:1091–100.
81. Higuchi S, Chaminade T, Imamizu H, Kawato M. Shared neural correlates for language and tool use in Broca's area. Neuroreport. 2009;20:1376–81.
82. Lombao D, Guardiola M, Mosquera M. Teaching to make stone tools: New experimental evidence supporting a technological hypothesis for the origins of language. Sci Rep. 2017;7:1–14.
83. Cataldo DM, Migliano AB, Vinicius L. Speech, stone tool-making and the evolution of language. Berwick RC, editor. PLoS One. 2018;13:e0191071.
84. Bolhuis JJ, Tattersall I, Chomsky N, Berwick RC. How Could Language Have Evolved? PLoS Biol. 2014;12:1–6.
85. Oppenheimer S. Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map. Philos Trans R Soc B Biol Sci. 2012;367:770–84.
86. Fisher SE, Vernes SC. Genetics and the Language Sciences. Annu Rev Linguist. 2015;1:289–310.
87. Christiansen MH, Chater N. Language as shaped by the brain. Behav Brain Sci. 2008;31:489–558.
88. Bender A. What Early Sapiens Cognition Can Teach Us: Untangling Cultural Influences on Human Cognition Across Time. Front Psychol. 2020;11:1–6.
89. Lupyan G, Dale R. Why Are There Different Languages? The Role of Adaptation in Linguistic Diversity. Trends Cogn Sci. Elsevier Ltd; 2016;20:649–60.
90. Dunbar RIM, Shultz S. Evolution in the Social Brain. Science (80- ). 2007;317:1344–7.
91. Fitch WT. Evolutionary Developmental Biology and Human Language Evolution: Constraints on Adaptation. Evol Biol. 2012;39:613–37.
92. Pleyer M, Hartmann S. Constructing a Consensus on Language Evolution? Convergences and Differences Between Biolinguistic and Usage-Based Approaches. Front Psychol. 2019;10.
93. Pagel M. Human language as a culturally transmitted replicator. Nat Rev Genet. 2009;10:405–15.
94. Kurland CG. What tangled web: barriers to rampant horizontal gene transfer. BioEssays. 2005;27:741–7.
95. Keeling PJ, Palmer JD. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet. 2008;9:605–18.
96. Labov W. The Social Stratification of English in New York City. 2nd ed. Cambridge: Cambridge University Press; 1966.
97. Stanford JN. Child dialect acquisition: New perspectives on parent/peer influence. J Socioling. 2008;12:567–96.
98. Syrjänen K, Honkola T, Lehtinen J, Leino A, Vesakoski O. Applying Population Genetic Approaches within Languages. Lang Dyn Chang. 2016;6:235–83.
99. Kimura M. The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press; 1983.
100. Mendívil-Giró J-L. Why Don't Languages Adapt to Their Environment? Front Commun. 2018;3:1–10.
101. Hendry AP, Kinnison MT. An introduction to microevolution: Rate, pattern, process. Genetica. 2001;112–113:1–8.
102. Blythe RA, Croft W. How individuals change language. PLoS One. 2021;16:1–23.
103. Stanley SM. A theory of evolution above the species level. Proc Natl Acad Sci. 1975;72:646–50.
104. Lowe WH, Kovach RP, Allendorf FW. Population Genetics and Demography Unite Ecology and Evolution. Trends Ecol Evol. Elsevier Ltd; 2017;32:141–52.
105. Reznick DN, Ricklefs RE. Darwin's bridge between microevolution and macroevolution. Nature. 2009;457:837–42.
106. Honkola T, Ruokolainen K, Syrjänen KJJ, Leino UP, Tammi I, Wahlberg N, et al. Evolution within a language: Environmental differences contribute to divergence of dialect groups. BMC Evol Biol. BMC Evolutionary Biology; 2018;18:1–15.
107. Kokko H, Chaturvedi A, Croll D, Fischer MC, Guillaume F, Karrenberg S, et al. Can Evolution Supply What Ecology Demands? Trends Ecol Evol. Elsevier Ltd; 2017;32:187–97.
108. Post DM, Palkovacs EP. Eco-evolutionary feedbacks in community and ecosystem ecology: interactions between the ecological theatre and the evolutionary play. Philos Trans R Soc B Biol Sci.

2009;364:1629–40.
109. Solé R V., Corominas-Murtra B, Fortuny J. Diversity, competition, extinction: the ecophysics of language change. J R Soc Interface. 2010;7:1647–64.
110. Lee S, Hasegawa T. Oceanic barriers promote language diversification in the Japanese Islands. J Evol Biol. 2014;27:1905–12.
111. Pacheco Coelho MT, Pereira EB, Haynie HJ, Rangel TF, Kavanagh P, Kirby KR, et al. Drivers of geographical patterns of North American language diversity. Proc R Soc B Biol Sci. 2019;286:20190242.
112. Gavin MC, Rangel TF, Bowern C, Colwell RK, Kirby KR, Botero CA, et al. Process-based modelling shows how climate and demography shape language diversity. Glob Ecol Biogeogr. 2017;26:584–91.
113. Greenhill SJ, Atkinson QD, Meade A, Gray RD. The shape and tempo of language evolution. Proc R Soc B Biol Sci. 2010;277:2443–50.
114. Gray RD, Drummond AJ, Greenhill SJ. Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. Science (80- ). 2009;323:479–83.
115. Moelling K, Broecker F. Viruses and evolution - Viruses first? A personal perspective. Front Microbiol. 2019;10:1–13.
116. Wang P. The Opening of Pandora's Box: An Emerging Role of Long Noncoding RNA in Viral Infections. Front Immunol. 2019;9:1–16.
117. Dennett DC. From bacteria to Bach and back: the evolution of minds. New York: W. W. Norton & Company; 2017.
118. Dawkins R. The Selfish Gene. 40th anniv. Oxford: Oxford University Press; 2016.
119. Perfors A, Navarro DJ. Language Evolution Can Be Shaped by the Structure of the World. Cogn Sci. 2014;38:775–93.
120. Morton ES. Ecological Sources of Selection on Avian Sounds. Am Nat. 1975;109:17–34.
121. Boncoraglio G, Saino N. Habitat structure and the evolution of bird song: a meta-analysis of the evidence for the acoustic adaptation hypothesis. Funct Ecol. 2007;21:134–42.
122. Maddieson I, Coupé C. Human spoken language diversity and the acoustic adaptation hypothesis. J Acoust Soc Am. 2015;138:1838–1838.
123. Maddieson I. Language Adapts to Environment: Sonority and Temperature. Front Commun. 2018;3:1–8.
124. Meyer J. Bioacoustics of human whistled languages: an alternative approach to the cognitive processes of language. An Acad Bras Cienc. 2004;76:406–12.
125. Meyer J. Typology and acoustic strategies of whistled languages: Phonetic comparison and perceptual cues of whistled vowels. J Int Phon Assoc. 2008;38:69–94.
126. Everett C. Evidence for Direct Geographic Influences on Linguistic Sounds: The Case of Ejectives. PLoS One. 2013;8.
127. Dediu D, Janssen R, Moisik SR. Language is not isolated from its wider environment: Vocal tract influences on the evolution of speech and language. Lang Commun. Elsevier Ltd; 2017;54:9–20.
128. Everett C, Blasi DE, Roberts SG. Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. Proc Natl Acad Sci. 2015;112:1322–7.
129. Everett C, Blasí DE, Roberts SG. Language evolution and climate: the case of desiccation and tone. J Lang Evol. 2016;1:33–46.
130. Turney PD, Mohammad SM. The natural selection of words: Finding the features of fitness. PLoS One. 2019;14:1–21.
131. Pagel M, Atkinson QD, Meade A. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. Nature. 2007;449:717–20.
132. Bolinger DL. The Life and Death of Words. Am Scholar. 1953;22:323–35.
133. Pagel M, Meade A. The deep history of the number words. Philos Trans R Soc B Biol Sci. 2017;373:20160517.
134. Pagel M, Beaumont M, Meade A, Verkerk A, Calude A. Dominant words rise to the top by positive frequency-dependent selection. Proc Natl Acad Sci. 2019;116:7397–402.
135. Newberry MG, Ahern CA, Clark R, Plotkin JB. Detecting evolutionary forces in language change. Nature. Nature Publishing Group; 2017;551:223–6.

136. Enfield NJ, Majid A, van Staden M. Cross-linguistic categorisation of the body: Introduction. Lang Sci. 2006;28:137–47.
137. Witkowski SR, Brown CH. Climate, Clothing, and Body-Part Nomenclature. Ethnology. 1985;24:197–214.
138. Brown CH. Hand and Arm. In: Dryer MS, Haspelmath M, editors. World Atlas Lang Struct Online. Leipzig: Max Planck Institute for Evolutionary Anthropology; 2013. p. 1–8.
139. Lindsey DT, Brown AM. Color Naming and the Phototoxic Effects of Sunlight on the Eye. Psychol Sci. 2002;13:506–12.
140. Gibson E, Futrell R, Jara-Ettinger J, Mahowald K, Bergen L, Ratnasingam S, et al. Color naming across languages reflects color use. Proc Natl Acad Sci. 2017;114:10785–90.
141. Winters J, Kirby S, Smith K. Languages adapt to their contextual niche. Lang Cogn. 2015;7:415–49.
142. Krupnik I, Müller-Wille L. Franz Boas and Inuktitut Terminology for Ice and Snow: From the Emergence of the Field to the "Great Eskimo Vocabulary Hoax." SIKU Knowing Our Ice. Dordrecht: Springer Netherlands; 2010. p. 377–400.
143. Regier T, Carstensen A, Kemp C. Languages Support Efficient Communication about the Environment: Words for Snow Revisited. Wennekers T, editor. PLoS One. 2016;11:e0151138.
144. DOEE D of E& E. Uluru-Kata Tjuta National Park: Knowledge Handbook. Canberra: Australian Government - Director of National Parks; 2012. p. 147.
145. Sapir E. Language: An Introduction to the Study of Speech. New York: Harcourt, Brace and Company; 1921.
146. Koerper HC, Stickel EG. Cultural Drift: A Primary Process of Culture Change. J Anthropol Res. 1980;36:463–9.
147. Bentley RA, Hahn MW, Shennan SJ. Random drift and culture change. Proc R Soc London Ser B Biol Sci. 2004;271:1443–50.
148. Lanfear R, Kokko H, Eyre-Walker A. Population size and the rate of evolution. Trends Ecol Evol. Elsevier Ltd; 2014;29:33–41.
149. Bromham L, Hua X, Fitzpatrick TG, Greenhill SJ. Rate of language evolution is affected by population size. Proc Natl Acad Sci. 2015;112:2097–102.
150. Allendorf FW. Genetic drift and the loss of alleles versus heterozygosity. Zoo Biol. 1986;5:181–90.
151. Raviv L, Meyer A, Lev-Ari S. The Role of Social Network Structure in the Emergence of Linguistic Structure. Cogn Sci. 2020;44.
152. Raviv L, Meyer A, Lev-Ari S. Larger communities create more systematic languages. Proc R Soc B Biol Sci. 2019;286:1–9.
153. Baxter GJ, Blythe RA, Croft W, McKane AJ. Modeling language change: An evaluation of Trudgill's theory of the emergence of New Zealand English. Lang Var Change. 2009;21:257–96.
154. Chang CB. A novelty effect in phonetic drift of the native language. J Phon. Elsevier; 2013;41:520–33.
155. Reali F, Griffiths TL. Words as alleles: Connecting language evolution with Bayesian learners to models of genetic drift. Proc R Soc B Biol Sci. 2010;277:429–36.
156. Borsley RD. Mutation and constituent structure in Welsh. Lingua. 1999;109:267–300.
157. Kennard HJ, Lahiri A. Mutation in Breton verbs: Pertinacity across generations. J Linguist. 2017;53:113–45.
158. Faust N. How low can you go? A note on vowel mutation in Nuer. J African Lang Linguist. 2017;38:51–64.
159. Hannahs SJ. Constraining Welsh vowel mutation. J Linguist. 2007;43:341–63.
160. Aplin LM, Farine DR, Morand-Ferron J, Cockburn A, Thornton A, Sheldon BC. Experimentally induced innovations lead to persistent culture via conformity in wild birds. Nature. Nature Publishing Group; 2015;518:538–41.
161. Perry SE, Barrett BJ, Godoy I. Older, sociable capuchins (Cebus capucinus) invent more social behaviors, but younger monkeys innovate more in other contexts. Proc Natl Acad Sci U S A. 2017;114:7806–13.
162. Whitehead H. Gene–culture coevolution in whales and dolphins. Proc Natl Acad Sci U S A.

2017;114:7814–21.
163. Birch J, Heyes C. The cultural evolution of cultural evolution. Philos Trans R Soc B Biol Sci. 2021;376.
164. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. Localizing Recent Adaptive Evolution in the Human Genome. McVean G, editor. PLoS Genet. 2007;3:e90.
165. Laland KN, Odling-Smee J, Myles S. How culture shaped the human genome: Bringing genetics and the human sciences together. Nat Rev Genet. Nature Publishing Group; 2010;11:137–48.
166. Creanza N, Feldman MW. Worldwide genetic and cultural change in human evolution. Curr Opin Genet Dev. The Author(s); 2016;41:85–92.
167. Mesoudi A. Cultural selection and biased transformation: Two dynamics of cultural evolution. Philos Trans R Soc B Biol Sci. 2021;376.
168. Mesoudi A. Cultural evolution: integrating psychology, evolution and culture. Curr Opin Psychol. Elsevier Ltd; 2016;7:17–22.
169. Tamariz M, Kirby S. The cultural evolution of language. Curr Opin Psychol. Elsevier Ltd; 2016;8:37–43.
170. Kirby S, Tamariz M, Cornish H, Smith K. Compression and communication in the cultural evolution of linguistic structure. Cognition. Elsevier B.V.; 2015;141:87–102.
171. Azumagakito T, Suzuki R, Arita T. An integrated model of gene-culture coevolution of language mediated by phenotypic plasticity. Sci Rep. Springer US; 2018;8:8025.
172. Dunn M, Greenhill SJ, Levinson SC, Gray RD. Evolved structure of language shows lineage-specific trends in word-order universals. Nature. Nature Publishing Group; 2011;473:79–82.
173. Blasi DE, Moran S, Moisik SR, Widmer P, Dediu D, Bickel B. Human sound systems are shaped by post-Neolithic changes in bite configuration. Science (80- ). 2019;363:eaav3218.
174. Lupyan G, Dale R. Language Structure Is Partly Determined by Social Structure. O'Rourke D, editor. PLoS One. 2010;5:e8559.
175. Wray A, Grace GW. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. Lingua. 2007;117:543–78.
176. Colwell RK, Rangel TF. Hutchinson's duality: The once and future niche. Proc Natl Acad Sci. 2009;106:19651–8.
177. Greenhill SJ, Hua X, Welsh CF, Schneemann H, Bromham L. Population Size and the Rate of Language Evolution: A Test Across Indo-European, Austronesian, and Bantu Languages. Front Psychol. 2018;9:1–18.
178. Bowern C. Correlates of Language Change in Hunter-Gatherer and Other "Small" Languages. Linguist Lang Compass. 2010;4:665–79.
179. Atkinson M, Smith K, Kirby S. Sociocultural determiners of linguistic complexity. Evol Lang. WORLD SCIENTIFIC; 2014. p. 379–80.
180. Hay J, Bauer L. Phoneme inventory size and population size. Language (Baltim). 2007;83:388–400.
181. Atkinson QD. Phonemic diversity supports a serial founder effect model of language expansion from Africa. Science (80- ). 2011;332:346–9.
182. Pérez-Losada J, Fort J. A serial founder effect model of phonemic diversity based on phonemic loss in low-density populations. PLoS One. 2018;13:1–22.
183. Fort J, Pérez-Losada J. Can a linguistic serial founder effect originating in Africa explain the worldwide phonemic cline? J R Soc Interface. 2016;13.
184. Tolins J, Samermit P. GIFs as Embodied Enactments in Text-Mediated Conversation. Res Lang Soc Interact. Routledge; 2016;49:75–91.
185. Derks D, Bos AER, Grumbkow J von. Emoticons and social interaction on the Internet: the importance of social context. Comput Human Behav. 2007;23:842–9.
186. Ramachandra V, Karanth P. The role of literacy in the conceptualization of words: Data from Kannada-speaking children and non-literate adults. Read Writ. 2006;20:173–99.
187. Baum DA, Smith SD, Donovan SSS. The Tree-Thinking Challenge. Science (80- ). 2005;310:979–80.
188. Gregory TR. Understanding Evolutionary Trees. Evol Educ Outreach. 2008;1:121–37.
189. Mace R, Pagel M, Bowen JR, Otterbein KF, Ridley M, Schweizer T, et al. The Comparative

Method in Anthropology. Curr Anthropol. 1994;35:549–64.
190. Moller AP, Birkhead TR. A pairwise comparative method as illustrated by copulation frequency in birds. Am Nat. 1992;139:644–56.
191. Huey RB, Jr TG, Turelli M. Revisiting a Key Innovation in Evolutionary Biology: Felsenstein's "Phylogenies and the Comparative Method." Am Nat. 2019;755–72.
192. Felsenstein J. Phylogenies and the comparative method. Am Nat. 1985;125:1–15.
193. Dunn M, Terrill A, Reesink G, Foley RA, Levinson SC. Structural phylogenetics and the reconstruction of ancient language history. Science (80- ). 2005;309:2072–5.
194. Pagel M. Inferring the historical patterns of biological evolution. Nature. 1999;401:877–84.
195. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat Publ Gr. Nature Publishing Group; 2012;13.
196. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17:754–5.
197. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. Pertea M, editor. PLOS Comput Biol. 2019;15:e1006650.
198. Bilderbeek RJC, Laudanno G, Etienne RS. Quantifying the impact of an inference model in Bayesian phylogenetics. Methods Ecol Evol. 2021;12:351–8.
199. Rezende EL, Diniz-Filho JAF. Phylogenetic Analyses: Comparing Species to Infer Adaptations and Physiological Mechanisms. Compr Physiol. Wiley; 2012. p. 639–74.
200. Graça da Silva S, Tehrani JJ. Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. R Soc Open Sci. 2016;3.
201. Prentiss AM, Walsh MJ, Skelton RR, Mattes M. Mosaic Evolution in Cultural Frameworks: Skateboard Decks and Projectile Points. In: Mendoza Straffon L, editor. Cult Phylogenetics Interdiscip Evol Res vol 4. Springer, Cham; 2016. p. 113–30.
202. Gray RD, Jordan FM. Language trees support the express-train sequence of Austronesian expansion. Nature. 2000;405:1052–5.
203. Mace R, Holden CJ. A phylogenetic approach to cultural evolution. Trends Ecol Evol. 2005;20:116–21.
204. Gray R. Pushing the time barrier in the quest for language roots. Science (80- ). 2005;309:2007–8.
205. Nichols J, Warnow T. Tutorial on computational linguistic phylogeny. Linguist Lang Compass. 2008;2:760–820.
206. Gray RD, Greenhill SJ, Ross RM. The Pleasures and Perils of Darwinizing Culture (with Phylogenies). Biol Theory. 2007;2:360–75.
207. Pereltsvaig A, Lewis MW. The Indo-European Controversy: Facts and Fallacies in Historical Linguistics. Cambridge: Cambridge University Press; 2015.
208. Blevins J. Evolutionary Phonology: The Emergence of Sound Patterns. Cambridge, UK: Cambridge University Press; 2004.
209. Blench R. 'New mathematical methods' in linguistics constitute the greatest intellectual fraud in the discipline since Chomsky. Nijmegen, Neth: Presented at Max Planck Inst. Psycholinguist.; 2015.
210. Andersen H. Synchrony, diachrony, and evolution. In: Thomsen ON, editor. Curr Issues Linguist Theory vol 279 Compet Model Linguist Chang Evol beyond. Amsterdam: Benjamins; 2006. p. 59–90.
211. Eska JF, Ringe DA. Recent Work in Computational Linguistic Phylogeny. Language (Baltim). 2004;80:569–82.
212. Steele J, Kandler A. Language trees ≠ gene trees. Theory Biosci. 2010;129:223–33.
213. Donohue M, Denham T. Languages and Genes Attest Different Histories in Island Southeast Asia. Ocean Linguist. 2011;50:536–42.
214. Box GEP. Science and Statistics. J Am Stat Assoc. 1976;71:791–9.
215. Bowern C. Computational Phylogenetics. Annu Rev Linguist. 2018;4:281–96.
216. Kassian A. Towards a Formal Genealogical Classification of the Lezgian Languages (North Caucasus): Testing Various Phylogenetic Methods on Lexical Data. Anisimova M, editor. PLoS One. 2015;10:e0116950.
217. Barbançon F, Evans SN, Nakhleh L, Ringe D, Warnow T. An experimental study comparing linguistic phylogenetic reconstruction methods. Diachronica. 2013;30:143–70.

218. Levinson SC, Gray RD. Tools from evolutionary biology shed new light on the diversification of languages. Trends Cogn Sci. Elsevier Ltd; 2012;16:167–73.
219. Jäger G. Data descriptor: Global-scale phylogenetic linguistic inference from lexical resources. Sci Data. The Author(s); 2018;5:1–16.
220. List JM, Walworth M, Greenhill SJ, Tresoldi T, Forkel R. Sequence comparison in computational historical linguistics. J Lang Evol. 2018;3:130–44.
221. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 2006;23:254–67.
222. Francis AR, Steel M. Which phylogenetic networks are merely trees with additional arcs? Syst Biol. 2015;64:768–77.
223. Schliep K, Potts AJ, Morrison DA, Grimm GW. Intertwining phylogenetic trees and networks. Methods Ecol Evol. 2017;8:1212–20.
224. Nakhleh L, Ringe DA, Warnow T. Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages. Language (Baltim). 2005;81:382–420.
225. List JM, Nelson-Sathi S, Geisler H, Martin W. Networks of lexical borrowing and lateral gene transfer in language and genome evolution. BioEssays. 2014;36:141–50.
226. Greenhill SJ, Currie TE, Gray RD. Does horizontal transmission invalidate cultural phylogenies? Proc R Soc B Biol Sci. 2009;276:2299–306.
227. Bowern C, Epps P, Gray R, Hill J, Hunley K, McConvell P, et al. Does lateral transmission obscure inheritance in hunter-gatherer languages? PLoS One. 2011;6.
228. Dediu D. Genetic biasing through cultural transmission: Do simple Bayesian models of language evolution generalise? J Theor Biol. 2009;259:552–61.
229. Chacon TC, List J-M. Improved computational models of sound change shed light on the history of the Tukanoan languages. J Lang Relatsh. Gorgias Press; 2010;3:177–204.
230. Robert Ladd D, Roberts SG, Dediu D. Correlational Studies in Typological and Historical Linguistics. Annu Rev Linguist. 2015;1:221–41.
231. Bhattacharya T, Retzlaff N, Blasi DE, Croft W, Cysouw M, Hruschka D, et al. Studying language evolution in the age of big data. J Lang Evol. 2018;3:94–129.
232. Lukas D, Towner M, Borgerhoff Mulder M. The potential to infer the historical pattern of cultural macroevolution. Philos Trans R Soc B Biol Sci. 2021;376:rstb.2020.0057.
233. Evans CL, Greenhill SJ, Watts J, List J-M, Botero CA, Gray RD, et al. The uses and abuses of tree thinking in cultural evolution. Philos Trans R Soc B Biol Sci. 2021;376:rstb.2020.0056.
234. Kirby KR, Gray RD, Greenhill SJ, Jordan FM, Gomes-Ng S, Bibiko H-J, et al. D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity. Mesoudi A, editor. PLoS One. 2016;11:e0158391.
235. Wichmann S, Holman EW, Brown CH, editors. The ASJP Database (version 19). 2020.
236. Hammarström H, Forkel R, Haspelmath M, Bank S. Glottolog 4.3. [Internet]. Jena Max Planck Inst. Sci. Hum. Hist. 2020. Available from: http://glottolog.org
237. Dryer MS, Haspelmath M. The World Atlas of Language Structures Online. Dryer MS, Haspelmath M, editors. Leipzig Max Planck Inst. Evol. Anthropol. 2013.
238. Dediu D. Making genealogical language classifications available for phylogenetic analysis. Lang Dyn Chang. 2018;8:1–21.
239. Pompei S, Loreto V, Tria F. On the accuracy of language trees. PLoS One. 2011;6.
240. Hammarström H. Linguistic diversity and language evolution. J Lang Evol. 2016;1:19–29.
241. Chang W, Cathcart C, Hall D, Garrett A. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. Language (Baltim). 2015;91:194–244.
242. Holden CJ. Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis. Proc R Soc B Biol Sci. 2002;269:793–9.
243. Dunn M. Contact and phylogeny in Island Melanesia. Lingua. 2009;119:1664–78.
244. Grollemund R, Branford S, Bostoen K, Meade A, Venditti C, Pagel M. Bantu expansion shows that habitat alters the route and pace of human dispersals. Proc Natl Acad Sci U S A. 2015;112:13296–301.
245. de Filippo C, Bostoen K, Stoneking M, Pakendorf B. Bringing together linguistic and genetic evidence to test the Bantu expansion. Proc R Soc B Biol Sci. 2012;279:3256–63.

246. De Filippo C, Barbieri C, Whitten M, Mpoloka SW, Gunnarsdóttir ED, Bostoen K, et al. Y-chromosomal variation in sub-Saharan Africa: Insights into the history of Niger-Congo groups. Mol Biol Evol. 2011;28:1255–69.
247. Bouckaert RR, Bowern C, Atkinson QD. The origin and expansion of Pama-Nyungan languages across Australia. Nat Ecol Evol. Springer US; 2018;2:741–9.
248. Currie TE, Meade A, Guillon M, Mace R. Cultural phylogeography of the Bantu Languages of sub-Saharan Africa. Proc R Soc B Biol Sci. 2013;280.
249. Gomes V, Pala M, Salas A, Álvarez-Iglesias V, Amorim A, Gómez-Carballa A, et al. Mosaic maternal ancestry in the Great Lakes region of East Africa. Hum Genet. Springer Berlin Heidelberg; 2015;134:1013–27.
250. Nichols J. Language Spread Rates and Prehistoric American Migration Rates. Curr Anthropol. 2008;49:1109–17.
251. Honkola T, Vesakoski O, Korhonen K, Lehtinen J, Syrjänen K, Wahlberg N. Cultural and climatic changes shape the evolutionary history of the Uralic languages. J Evol Biol. 2013;26:1244–53.
252. Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, et al. History of Click-Speaking Populations of Africa Inferred from mtDNA and Y Chromosome Genetic Variation. Mol Biol Evol. 2007;24:2180–95.
253. Greenhill SJ, Wu C-H, Hua X, Dunn M, Levinson SC, Gray RD. Evolutionary dynamics of language systems. Proc Natl Acad Sci. 2017;114:E8822–9.
254. Maurits L, Griffiths TL. Tracing the roots of syntax with Bayesian phylogenetics. Proc Natl Acad Sci. 2014;111:13576–81.
255. Gell-Mann M, Ruhlen M. The origin and evolution of word order. Proc Natl Acad Sci. 2011;108:17290–5.
256. Pagel M, Atkinson QD, Calude AS, Meade A. Ultraconserved words point to deep language ancestry across Eurasia. Proc Natl Acad Sci. 2013;110:8471–6.
257. Teixidor-Toneu I, Kool A, Greenhill SJ, Kjesrud K, Sandstedt JJ, Manzanilla V, et al. Historical, archaeological and linguistic evidence test the phylogenetic inference of Viking-Age plant use. Philos Trans R Soc B Biol Sci. 2021;376.
258. Diamond J, Bellwood P. Farmers and Their Languages: The First Expansions. Science (80- ). 2003;300:597–603.
259. Opie C, Shultz S, Atkinson QD, Currie T, Mace R. Phylogenetic reconstruction of Bantu kinship challenges Main Sequence Theory of human social evolution. Proc Natl Acad Sci. 2014;111:17414–9.
260. Lansing JS, Abundo C, Jacobs GS, Guillot EG, Thurner S, Downey SS, et al. Kinship structures create persistent channels for language transmission. Proc Natl Acad Sci. 2017;114:12910–5.
261. Walker RS, Wichmann S, Mailund T, Atkisson CJ. Cultural phylogenetics of the tupi language family in lowland south america. PLoS One. 2012;7.
262. Jäger G. Support for linguistic macrofamilies from weighted sequence alignment. Proc Natl Acad Sci. 2015;112:12752–7.
263. Pagel M, Meade A. Estimating rates of lexical replacement on phylogenetic trees of languages. In: Forster P, Renfrew C, editors. Phylogenetic methods prehistory Lang. Cambridge: McDonald Institute Monographs; 2006. p. 173–82.
264. Bokma F. Detection of punctuated equilibrium from molecular phylogenies. J Evol Biol. 2002;15:1048–56.
265. Mayr E. Speciation and Macroevolution. Evolution (N Y). 1982;36:1119–32.
266. Eldredge N, Gould SJ. Punctuated Equilibria: An Alternative to Phyletic Gradualism. In: Schopf TJM, Thomas JM, editors. Model Paleobiol. San Francisco, CA: Freeman, Cooper & Company; 1972. p. 82–115.
267. Mattila TM, Bokma F. Extant mammal body masses suggest punctuated equilibrium. Proc R Soc B Biol Sci. 2008;275:2195–9.
268. Gemmell MR, Trewick SA, Hills SFK, Morgan-Richards M. Phylogenetic topology and timing of New Zealand olive shells are consistent with punctuated equilibrium. J Zool Syst Evol Res. 2020;58:209–20.
269. Landis MJ, Schraiber JG, Liang M. Phylogenetic Analysis Using Lévy Processes: Finding Jumps in the Evolution of Continuous Traits. Syst Biol. 2013;62:193–204.

270. Pagel M, Venditti C, Meade A. Large Punctuational Contribution of Speciation to Evolutionary Divergence at the Molecular Level. Science (80- ). 2006;314:119–21.
271. Pennell MW, Harmon LJ, Uyeda JC. Is there room for punctuated equilibrium in macroevolution? Trends Ecol Evol. Elsevier Ltd; 2014;29:23–32.
272. Atkinson QD, Meade A, Venditti C, Greenhill SJ, Pagel M. Languages Evolve in Punctuational Bursts. Science (80- ). 2008;319:588–588.
273. Joseph BD, Mufwene SS. Parsing the Evolution of Language Change. Science (80- ). 2008;320:446.
274. Holman EW, Wichmann S. New Evidence from Linguistic Phylogenetics Identifies Limits to Punctuational Change. Syst Biol. 2016;66:syw106.
275. Valverde S, Sole R V. Punctuated equilibrium in the large-scale evolution of programming languages. J R Soc Interface. 2015;12.
276. Deskis SE. English Past and English Present: The Phrase “Old English” in Middle English Texts. Neophilologus. Springer Netherlands; 2018;102:141–53.
277. Alam S, Yao N. Big Data Analytics, Text Mining and Modern English Language. J Grid Comput. Journal of Grid Computing; 2019;17:357–66.
278. Matkovska M V. Transformations in the translation of Beowulf from Old English to present-day English. Philol Sci Transl Stud Eur Potential. Baltija Publishing; 2021. p. 142–4.
279. Mayr E. Systematics and the Origin of Species. New York: Columbia University Press; 1942.
280. Dobzhansky T. Genetics and the Origin of Species. New York: Columbia University Press; 1937.
281. Padilla-Iglesias C, Gjesfjeld E, Vinicius L. Geographical and social isolation drive the evolution of Austronesian languages. PLoS One. 2020;15:1–16.
282. Mairal Q, Santos C, Silva M, Marques SL, Ramos A, Aluja MP, et al. Linguistic isolates in Portugal: Insights from the mitochondrial DNA pattern. Forensic Sci Int Genet. 2013;7:618–23.
283. Bauduer F, Feingold J, Lacombe D. The Basques: Review of population genetics and Mendelian disorders. Hum Biol. 2005;77:619–37.
284. Bolnick DI, Fitzpatrick BM. Sympatric speciation: Models and empirical evidence. Annu Rev Ecol Evol Syst. 2007;38:459–87.
285. Petalas C, Lazarus T, Lavoie RA, Elliott KH, Guigueno MF. Foraging niche partitioning in sympatric seabird populations. Sci Rep. Nature Publishing Group UK; 2021;11:1–12.
286. Foley RA. The Evolutionary Ecology of Linguistic Diversity in Human Populations. In: Jones M, editor. Traces Ancestry Stud honour Colin Renfrew. Cambridge: McDonald Institute Monographs; 2004. p. 61–71.
287. Britain D. Beyond the “gentry aesthetic”: elites, Received Pronunciation and the dialectological gaze. Soc Semiot. Taylor & Francis; 2017;27:288–98.
288. Butlin RK, Galindo J, Grahame JW. Sympatric, parapatric or allopatric: the most important way to classify speciation? Philos Trans R Soc B Biol Sci. 2008;363:2997–3007.
289. Gavrilets S, Li H, Vose MD. Patterns of parapatric speciation. Evolution (N Y). 2000;54:1126–34.
290. Levin BA, Gandlin AA, Simonov ES, Levina MA, Barmintseva AE, Japoshvili B, et al. Phylogeny, phylogeography and hybridization of Caucasian barbels of the genus Barbus (Actinopterygii, Cyprinidae). Mol Phylogenet Evol. Elsevier; 2019;135:31–44.
291. Willis PM, Symula RE, Lovette IJ. Ecology, song similarity and phylogeny predict natural hybridization in an avian family. Evol Ecol. 2014;28:299–322.
292. Soltis PS, Soltis DE. The role of hybridization in plant speciation. Annu Rev Plant Biol. 2009;60:561–88.
293. Mallet J. Hybridization as an invasion of the genome. Trends Ecol Evol. 2005;20:229–37.
294. Mallet J. Hybrid speciation. Nature. 2007;446:279–83.
295. Counterman BA. Hybrid Speciation. Encycl Evol Biol. 2016;2:242–8.
296. Evans N. Did language evolve in multilingual settings? Biol Philos. 2017;32:905–33.
297. Nichols J. Non-linguistic Conditions for Causativization as a Linguistic Attractor. Front Psychol. 2018;8.
298. Chacon T. Arawakan and Tukanoan contacts in Northwest Amazonia prehistory. Papia. 2017;27:237–65.
299. Arias L, Barbieri C, Barreto G, Stoneking M, Pakendorf B. High-resolution mitochondrial DNA

analysis sheds light on human diversity, cultural interactions, and population mobility in Northwestern Amazonia. Am J Phys Anthropol. 2018;165:238–55.
300. Cox MP, Hudjashov G, Sim A, Savina O, Karafet TM, Sudoyo H, et al. Small Traditional Human Communities Sustain Genomic Diversity over Microgeographic Scales despite Linguistic Isolation. Mol Biol Evol. 2016;33:2273–84.
301. Epps P. Amazonian linguistic diversity and its sociocultural correlates. In: Crevels M, Muysken P, editors. Lang Dispersal, Divers Contact A Glob Perspect. Oxford University Press; 2020. p. 275–90.
302. Garrett PB. What a language is good for: Language socialization, language shift, and the persistence of code-specific genres in St. Lucia. Lang Soc. 2005;34:327–61.
303. Muysken P. Language contact outcomes as the result of bilingual optimization strategies. Biling Lang Cogn. 2013;16:709–30.
304. Weston D. The lesser of two evils: Atypical trajectories in English dialect evolution. J Socioling. 2015;19:671–87.
305. Terry NP, Connor CM, Johnson L, Stuckey A, Tani N. Dialect variation, dialect-shifting, and reading comprehension in second grade. Read Writ. 2016;29:267–95.
306. Mufwene SS. Language Birth and Death. Annu Rev Anthropol. 2004;33:201–22.
307. McLeod W, O'Rourke B. "New speakers" of Gaelic: perceptions of linguistic authenticity and appropriateness. Appl Linguist Rev. 2015;6:151–72.
308. Mufwene SS. The Ecology of Language Evolution. Cambridge University Press; 2001.
309. Kouwenberg S, Singler JV. Creolization in Context: Historical and Typological Perspectives. Annu Rev Linguist. 2018;4:213–32.
310. Webster J. Creolizing the Roman Provinces. Am J Archaeol. 2001;105:209.
311. Spears AK. Haitian Creole Language. In: Paolo M Di, Spears AK, editors. Lang Dialects US Focus Divers Linguist. Routledge; 2014.
312. Eberhard DM, Simons GF, Fennig CD, editors. Ethnologue: Languages of the World. 24th ed. Dallas, Texas: SIL International; 2021.
313. Daval-Markussen A, Bakker P. Creole typology II: Typological features of creoles: From early proposals to phylogenetic approaches and comparisons with non-creoles. Creole Stud – Phylogenetic Approaches. Amsterdam: John Benjamins Publishing Company; 2017. p. 103–40.
314. Siegel J. The Emergence of Pidgin and Creole Languages. Oxford, New York: Oxford University Press; 2008.
315. Baker P. No Creolisation without prior pidginisation. Te Reo Journal, Linguist Soc New Zeal. 2001;44:31–50.
316. DeGraff M. Against Creole Exceptionalism. Language (Baltim). 2003;79:391–410.
317. Alleyne MC. Acculturation and the cultural matrix of creolization. In: Hymes D, editor. Pidginization Creolization Lang. Cambridge: Cambridge University Press; 1971. p. 169–86.
318. Benazzi S, Douka K, Fornai C, Bauer CC, Kullmer O, Svoboda J, et al. Early dispersal of modern humans in Europe and implications for Neanderthal behaviour. Nature. 2011;479:525–8.
319. Flores JC. Diffusion coefficient of Modern Humans outcompeting Neanderthals. J Theor Biol. Elsevier; 2011;280:189–90.
320. Gilpin W, Feldman MW, Aoki K. An ecocultural model predicts Neanderthal extinction through competition with modern humans. Proc Natl Acad Sci U S A. 2016;113:2134–9.
321. Banks WE, d'Errico F, Peterson AT, Kageyama M, Sima A, Sánchez-Goñi MF. Neanderthal extinction by competitive exclusion. PLoS One. 2008;3:1–8.
322. Hackel J, Sanmartín I. Modelling the tempo and mode of lineage dispersal. Trends Ecol Evol. Elsevier Ltd; 2021;1–11.
323. Schneider F. Dispersal And Migration. Annu Rev Entomol. 1962;7:223–42.
324. Neureiter N, Ranacher P, Van Gijn R, Bickel B, Weibel R. Can Bayesian phylogeography reconstruct migrations and expansions in linguistic evolution? R Soc Open Sci. 2021;8.
325. Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J. Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. Proc Natl Acad Sci U S A. 1988;85:6002–6.
326. Pakendorf B. Coevolution of languages and genes. Curr Opin Genet Dev. Elsevier Ltd; 2014;29:39–44.
327. Belle EMS, Barbujani G. Worldwide Analysis of Multiple Microsatellites: Language Diversity has

a Detectable Influence on DNA Diversity. Am J Phys Anthropol. 2007;133:1137–46.
328. Balanovsky O, Dibirova K, Dybo A, Mudrak O, Frolova S, Pocheshkhova E, et al. Parallel evolution of genes and languages in the Caucasus region. Mol Biol Evol. 2011;28:2905–20.
329. Lansing JS, Cox MP, Downey SS, Gabler BM, Hallmark B, Karafet TM, et al. Coevolution of languages and genes on the island of Sumba, eastern Indonesia. Proc Natl Acad Sci. 2007;104:16022–6.
330. Srithawong S, Srikummool M, Pittayaporn P, Ghirotto S, Chantawannakul P, Sun J, et al. Genetic and linguistic correlation of the Kra-Dai-speaking groups in Thailand. J Hum Genet. Nature Publishing Group; 2015;60:371–80.
331. Karafet TM, Bulayeva KB, Nichols J, Bulayev OA, Gurgenova F, Omarova J, et al. Coevolution of genes and languages and high levels of population structure among the highland populations of Daghestan. J Hum Genet. Nature Publishing Group; 2016;61:181–91.
332. Sun H, Zhou C, Huang X, Liu S, Lin K, Yu L, et al. Correlation between the linguistic affinity and genetic diversity of Chinese ethnic groups. J Hum Genet. 2013;58:686–93.
333. Hunley K, Dunn M, Lindström E, Reesink G, Terrill A, Healy ME, et al. Genetic and linguistic coevolution in Northern Island Melanesia. PLoS Genet. 2008;4.
334. Hunley K, Long JC. Gene flow across linguistic boundaries in Native North American populations. Proc Natl Acad Sci U S A. 2005;102:1312–7.
335. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci. 2005;102:15942–7.
336. Prugnolle F, Manica A, Balloux F. Geography predicts neutral genetic diversity of human populations. Curr Biol. 2005;15:R159–60.
337. Deshpande O, Batzoglou S, Feldman MW, Cavalli-Sforza LL. A serial founder effect model for human settlement out of Africa. Proc R Soc B Biol Sci. 2009;276:291–300.
338. Perreault C, Mathew S. Dating the origin of language using phonemic diversity. PLoS One. 2012;7.
339. Stringer C. Human evolution: Out of Ethiopia. Nature. 2003;423:692–4.
340. Pickrell JK, Reich D. Toward a new history and geography of human genes informed by ancient DNA. Trends Genet. Elsevier Ltd; 2014;30:377–89.
341. DeGiorgio M, Jakobsson M, Rosenberg NA. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. Proc Natl Acad Sci U S A. 2009;106:16057–62.
342. Hunley K, Bowern C, Healy M. Rejection of a serial founder effects model of genetic and linguistic coevolution. Proc R Soc B Biol Sci. 2012;279:2281–8.
343. Cysouw M, Dediu D, Moran S. Comment on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa." Science (80- ). 2012;335:657–657.
344. Dahl Ö. Are small languages more or less complex than big ones? Linguist Typology. 2011;15:171–5.
345. Maddieson I, Bhattacharya T, Smith DE, Croft W. Geographical distribution of phonological complexity. Linguist Typology. 2011;15:267–79.
346. Wang C-C, Ding Q-L, Tao H, Li H. Comment on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa." Science (80- ). 2012;335:657–657.
347. Bowern C. Out of Africa? The logic of phoneme inventories and founder effects. Linguist Typology. 2011;15:207–16.
348. Creanza N, Ruhlen M, Pemberton TJ, Rosenberg NA, Feldman MW, Ramachandrand S. A comparison of worldwide phonemic and genetic variation in human populations. Proc Natl Acad Sci U S A. 2015;112:1265–72.
349. Rodrigues C. Founder effect in tupian languages. Rev Diadorim. 2020;22:65–97.
350. Law J. On the Social Explanation of Technical Change: The Case of the Portuguese Maritime Expansion. Technol Cult. 1987;28:227.
351. Alperson-Afil N. Continual fire-making by Hominins at Gesher Benot Ya'aqov, Israel. Quat Sci Rev. 2008;27:1733–9.
352. MacDonald K. The use of fire and human distribution. Temperature. Taylor & Francis; 2017;4:153–65.

353. Beyin A. Upper Pleistocene Human Dispersals out of Africa: A Review of the Current State of the Debate. Int J Evol Biol. 2011;2011:1–17.
354. Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, et al. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Genet. 2011;89:516–28.
355. Clarkson C, Jacobs Z, Marwick B, Fullagar R, Wallis L, Smith M, et al. Human occupation of northern Australia by 65,000 years ago. Nature. Macmillan Publishers Limited, part of Springer Nature. All rights reserved.; 2017;547:306.
356. Waguespack NM. Why we're still arguing about the Pleistocene occupation of the Americas. Evol Anthropol Issues, News, Rev. 2007;16:63–74.
357. Dillehay TD, Ocampo C, Saavedra J, Sawakuchi AO, Vega RM, Pino M, et al. New archaeological evidence for an early human presence at Monte Verde, Chile. PLoS One. 2015;10:1–27.
358. Nichols J. How America Was Colonized: Linguistic Evidence. In: Frachetti MD, Spengler III RN, editors. Mobil Anc Soc Asia Am. Cham: Springer International Publishing; 2015. p. 117–26.
359. Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, et al. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science (80- ). 2015;349.
360. Williams RC, Steinberg AG, Gershowitz H, Bennett PH, Knowler WC, Pettitt DJ, et al. GM allotypes in Native Americans: Evidence for three distinct migrations across the Bering land bridge. Am J Phys Anthropol. 1985;66:1–19.
361. Rogers RA, Rogers LA, Hoffmann RS, Martin LD. Native American Biological Diversity and the Biogeographic Influence of Ice Age Refugia. J Biogeogr. 1991;18:623.
362. Sicoli MA, Holton G. Linguistic Phylogenies Support Back-Migration from Beringia to Asia. Caramelli D, editor. PLoS One. 2014;9:e91722.
363. Sagart L, Jacques G, Lai Y, Ryder RJ, Thouzeau V, Greenhill SJ, et al. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. Proc Natl Acad Sci. 2019;116:10317–22.
364. Lee S, Hasegawa T. Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. Proc R Soc B Biol Sci. 2011;278:3662–9.
365. Gray RD, Atkinson QD. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature. 2003;426:435–9.
366. Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko A V., Drummond AJ, et al. Mapping the Origins and Expansion of the Indo-European Language Family. Science (80- ). 2012;337:957–60.
367. Kemp BM, González-Oliver A, Malhi RS, Monroe C, Schroeder KB, McDonough J, et al. Evaluating the farming/language dispersal hypothesis with genetic variation exhibited by populations in the Southwest and Mesoamerica. Proc Natl Acad Sci U S A. 2010;107:6759–64.
368. Walker RS, Ribeiro LA. Bayesian phylogeography of the Arawak expansion in lowland South America. Proc R Soc B Biol Sci. 2011;278:2562–7.
369. Ramallo V, Bisso-Machado R, Bravi C, Coble MD, Salzano FM, Hünemeier T, et al. Demographic expansions in South America: Enlightening a complex scenario with genetic and linguistic data. Am J Phys Anthropol. 2013;150:453–63.
370. McGill BJ. The what, how and why of doing macroecology. Glob Ecol Biogeogr. 2019;28:6–17.
371. Brown JH, Maurer BA. Macroecology: The Division of Food and Space Among Species on Continents. Science (80- ). 1989;243:1145–50.
372. McGill BJ, Chase JM, Hortal J, Overcast I, Rominger AJ, Rosindell J, et al. Unifying macroecology and macroevolution to answer fundamental questions about biodiversity. Glob Ecol Biogeogr. 2019;28:1925–36.
373. Moseley C, Asher RE. Atlas of the World's Languages. London: Routledge; 1994.
374. Mace R, Pagel M. A latitudinal gradient in the density of human languages in North America. Proc R Soc London Ser B Biol Sci. 1995;261:117–21.
375. Nettle D. Language Diversity in West Africa: An Ecological Approach. J Anthropol Archaeol. 1996;15:403–38.
376. Nettle D. Explaining Global Patterns of Language Diversity. J Anthropol Archaeol. 1998;17:354–74.
377. Currie TE, Mace R. Political complexity predicts the spread of ethnolinguistic groups. Proc Natl

Acad Sci. 2009;106:7339–44.
378. Derungs C, Köhl M, Weibel R, Bickel B. Environmental factors drive language density more in food-producing than in hunter–gatherer populations. Proc R Soc B Biol Sci. 2018;285:20172851.
379. Moore JL, Manne L, Brooks T, Burgess ND, Davies R, Rahbek C, et al. The distribution of cultural and biological diversity in Africa. Proc R Soc B Biol Sci. 2002;269:1645–53.
380. Sutherland WJ. Parallel extinction risk and global distribution of languages and species. Nature. 2003;423:276–9.
381. Gavin MC, Sibanda N. The island biogeography of languages. Glob Ecol Biogeogr. 2012;21:958–67.
382. Hua X, Greenhill SJ, Cardillo M, Schneemann H, Bromham L. The ecological drivers of variation in global language diversity. Nat Commun. Springer US; 2019;10:2047.
383. Michalopoulos S. The Origins of Ethnolinguistic Diversity: Theory and Evidence. SSRN Electron J. 2008;0–54.
384. Bailey DH, Hamilton MJ, Walker RS. Latitude, population size, and the language-farming dispersal hypothesis. Evol Ecol Res. 2012;14:1057–67.
385. Laitin DD, Moortgat J, Robinson AL. Geographic axes and the persistence of cultural diversity. Proc Natl Acad Sci. 2012;109:10263–8.
386. Manne LL. Nothing has yet lasted forever: Current and threatened levels of biological and cultural diversity. Evol Ecol Res. 2003;5:517–27.
387. Smith EA. On the Coevolution of Cultural, Linguistic, and Biological Diversity. In: Maffi L, editor. Biocultural Divers Link Lang Knowledge, Environ. Washington and London: Smithsonian Institution Press; 2001. p. 95–116.
388. Cardillo M, Bromham L, Greenhill SJ. Links between language diversity and species richness can be confounded by spatial autocorrelation. Proc R Soc B Biol Sci. 2015;282:20142986.
389. Codding BF, Jones TL. Environmental productivity predicts migration, demographic, and linguistic patterns in prehistoric California. Proc Natl Acad Sci. 2013;110:14569–73.
390. Axelsen JB, Manrubia S. River density and landscape roughness are universal determinants of linguistic diversity. Proc R Soc B Biol Sci. 2014;281:20133029–20133029.
391. Turvey ST, Pettorelli N. Spatial congruence in language and species richness but not threat in the world's top linguistic hotspot. Proc R Soc B Biol Sci. 2014;281:20141644.
392. Gavin MC, Botero CA, Bowern C, Colwell RK, Dunn M, Dunn RR, et al. Toward a Mechanistic Understanding of Linguistic Diversity. Bioscience. 2013;63:524–35.
393. Rangel TF, Edwards NR, Holden PB, Diniz-Filho JAF, Gosling WD, Coelho MTP, et al. Modeling the ecology and evolution of biodiversity: Biogeographical cradles, museums, and graves. Science (80-). 2018;361:eaar5452.
394. Gotelli NJ, Anderson MJ, Arita HT, Chao A, Colwell RK, Connolly SR, et al. Patterns and causes of species richness: a general simulation model for macroecology. Ecol Lett. 2009;12:873–86.
395. Colwell RK, Rangel TF. A stochastic, evolutionary model for range shifts and richness on tropical elevational gradients under Quaternary glacial cycles. Philos Trans R Soc B Biol Sci. 2010;365:3695–707.
396. Rangel TF, Diniz-Filho JAF, Colwell RK. Species Richness and Evolutionary Niche Dynamics: A Spatial Pattern–Oriented Simulation Experiment. Am Nat. 2007;170:602–16.
397. Pontarp M, Bunnefeld L, Cabral JS, Etienne RS, Fritz SA, Gillespie R, et al. The Latitudinal Diversity Gradient: Novel Understanding through Mechanistic Eco-evolutionary Models. Trends Ecol Evol. Elsevier Ltd; 2019;34:211–23.
398. Newman MEJ. A Model of Mass Extinction. J Theor Biol. 1997;189:235–52.
399. Raup D. Biological extinction in earth history. Science (80- ). 1986;231:1528–33.
400. Raup DM. The role of extinction in evolution. Proc Natl Acad Sci U S A. 1994;91:6758–63.
401. Wiens D, Slaton MR. The mechanism of background extinction. Biol J Linn Soc. 2012;105:255–68.
402. Gilpin ME, Soulé ME. Minimum Viable Populations: Processes of Species Extinction. Conserv Biol Sci Scarcity Divers. Sunderland: Sinauer Associates; 1986. p. 19–34.
403. Hung T. How Did Language Evolve? Some Reflections on the Language Parasite Debate. Biol Theory. Springer Netherlands; 2019;14:214–23.

404. Crystal D. Language Death. Cambridge: Cambridge University Press; 2000.
405. Barnosky AD, Matzke N, Tomiya S, Wogan GOU, Swartz B, Quental TB, et al. Has the Earth's sixth mass extinction already arrived? Nature. 2011;471:51–7.
406. Ceballos G, Ehrlich PR, Barnosky AD, García A, Pringle RM, Palmer TM. Accelerated modern human – induced species losses: entering the sixth mass extinction. Sci Adv. 2015;1:1–5.
407. Ceballos G, Ehrlich PR, Dirzo R. Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. Proc Natl Acad Sci. 2017;201704949.
408. Elhacham E, Ben-Uri L, Grozovski J, Bar-On YM, Milo R. Global human-made mass exceeds all living biomass. Nature. Springer US; 2020;588:442–4.
409. Díaz S, Settele J, Brondizio ES, Ngo HT, Agard J, Arneth A, et al. Pervasive human-driven decline of life on Earth points to the need for transformative change. Science (80- ). 2019;1327.
410. Vallejos PQ, Veit PG, Tipula P, Reytar K. Undermining rights: Indigenous lands and mining in the Amazon. 2020.
411. United Nations. Indigenous Peoples [Internet]. Dep. Econ. Soc. Aff. 2021 [cited 2021 Sep 17]. Available from: https://www.un.org/development/desa/indigenouspeoples/mandated-areas1/environment.html
412. Polidoro M, de Assis Mendonça F, Meneghel SN, Alves-Brito A, Gonçalves M, Bairros F, et al. Territories Under Siege: Risks of the Decimation of Indigenous and Quilombolas Peoples in the Context of COVID-19 in South Brazil. J Racial Ethn Heal Disparities. Journal of Racial and Ethnic Health Disparities; 2020;
413. Harmon D, Loh J. The Index of Linguistic Diversity : A New Quantitative Measure of Trends in the Status of the World ' s Languages. Lang Doc Conserv. 2010;4:97–151.
414. Rodrigues AD. Línguas indígenas: 500 anos de descobertas e perdas. DELTA Doc. e Estud. em Linguística Teórica e Apl. 1993. p. 83–103.
415. IUCN. The IUCN Red List of Threatened Species. Version 2020-2. [Internet]. 2020. Available from: https://www.iucnredlist.org
416. Amano T, Sandel B, Eager H, Bulteau E, Svenning J-C, Dalsgaard B, et al. Global distribution and drivers of language extinction risk. Proc R Soc B Biol Sci. 2014;281:20141574–20141574.
417. Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J. Biodiversity hotspots for conservation priorities. Nature. 2000;403:853–8.
418. Gorenflo LJ, Romaine S, Mittermeier RA, Walker-Painemilla K. Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. Proc Natl Acad Sci. 2012;109:8032–7.
419. Tershy BR, Shen KW, Newton KM, Holmes ND, Croll DA. The importance of islands for the protection of biological and linguistic diversity. Bioscience. 2015;65:592–7.
420. Frainer A, Mustonen T, Hugu S, Andreeva T, Arttijeff E-M, Arttijeff I-S, et al. Opinion: Cultural and linguistic diversities are underappreciated pillars of biodiversity. Proc Natl Acad Sci. 2020;117:26539–43.
421. Inglis D, Pascual U. On the links between nature's values and language. People Nat. 2021;pan3.10205.
422. Gavin MC, McCarter J, Mead A, Berkes F, Stepp JR, Peterson D, et al. Defining biocultural approaches to conservation. Trends Ecol Evol. Elsevier Ltd; 2015;30:140–5.
423. Hanspach J, Jamila Haider L, Oteros-Rozas E, Stahl Olafsson A, Gulsrud NM, Raymond CM, et al. Biocultural approaches to sustainability: A systematic review of the scientific literature. People Nat. 2020;2:643–59.
424. Lyver POB, Timoti P, Davis T, Tylianakis JM. Biocultural Hysteresis Inhibits Adaptation to Environmental Change. Trends Ecol Evol. The Authors; 2019;34:771–80.
425. Fernández-Llamazares Á, Terraube J, Gavin MC, Pyhälä A, Siani SMO, Cabeza M, et al. Reframing the Wilderness Concept can Bolster Collaborative Conservation. Trends Ecol Evol. The Authors; 2020;35:750–3.
426. Garnett ST, Burgess ND, Fa JE, Fernández-Llamazares Á, Molnár Z, Robinson CJ, et al. A spatial overview of the global importance of Indigenous lands for conservation. Nat Sustain. Springer US; 2018;1:369–74.

# CAPÍTULO 2

---

## *Do you say cookie or biscuit? Using citizen science to research language evolution*

Christielly Borges[1*], Zander Vilaça[1], Guilherme Batista Ferreira[2], Thiago Fernando Rangel[1]

[1] Departamento de Ecologia, Universidade Federal de Goiás, Goiânia, GO, Brasil

[2] Independent Researcher, Brasil

* Corresponding author: christielly@gmail.com

*Status:* Em preparação para submissão como um artigo de *Methodology* para a **Journal of Language Evolution**

**ABSTRACT**

Do you call two wafers with a crème filling a “cookie” or a “biscuit”? Your answer will surely depend on where you grew up. Because language is used to expresses culture and describe the world, it varies according to the groups that use it. Therefore, to understand language evolution it is necessary to portray the social life of the community where it occurs. We exhaustively collected relevant linguistic variables for Brazilian Portuguese from the literature, finding information for only 780 out of 5570 municipalities. Continental-sized countries, such as Brazil, tend to experience linguistic data shortfall, since nationwide surveys are difficult to produce due to the lack of financial resources and mobility issues. On a spatial macroscale, citizen science can help surpass this data shortfall, engaging the public to share the linguistic variants they use and where they are from. We developed a web-based dialect quiz for Brazilian Portuguese, called *Fuxiquera*, that aims at collecting previously chosen linguistic variants from the public for all regions of Brazil. We used simple distance- and dissimilarity-based methods to predict where a person is from based on their answers to our questions. We thoroughly describe the creation of the website and the quiz’s algorithms, which can be easily applied to any language. We further performed cluster analyses with the linguistic data available, which showed a clear north-south superdialect divide and five main regional superdialects in Brazil. Our multidisciplinary work combines linguistic, information technology, ecological and geographical knowledge and methods to further the understanding of how languages change, evolve and are structured in space.

*Keywords:* crowdsourcing, cultural-historical geography, dialects, dialect quiz, K-Means Cluster, linguistic variation, Sorensen distance

## 1. INTRODUCTION

Citizen science is scientific research done with public participation, where the public can contribute with data collection, data visualization, analyses, results interpretation, and even research questions, despite not being trained experts in the specific topic of study (Newman et al., 2012). Citizen science can be a resourceful way for scientists to address data limitations, especially large-scale data, that would otherwise be very challenging to collect due to limited time and financial resources (Theobald et al., 2015).

Citizen science projects have helped scientists discover over 30 new species (Freitag, Pangantihon, & Njunjić, 2018; Hartop, Brown, & Disney, 2015), a new brown dwarf (Kuchner et al., 2017) and inform United Nations Sustainable Development Goals (Fritz et al., 2019). Emerging technologies and the widespread use of internet-based devices, such as smartphones and laptops, have been an important ally in citizen science's success (Newman et al., 2012). These devices enable citizen science projects to reach a broad audience and remote corners of the world, allowing people to actively contribute with worldwide geographically explicit, large spatial and temporal scales data for bird sightings (Sullivan et al., 2014), species' geographical ranges (Soroye, Ahmed, & Kerr, 2018; Tiago, Pereira, & Capinha, 2017) and conservation biogeography (Devictor, Whittaker, & Beltrame, 2010). Nonetheless, beyond providing data about nature and the world around them, the public can also help scientists with information that is intrinsic and ubiquitous to each individual – the way they speak.

Smartphone applications and online surveys have indeed been used as a medium to crowdsource language data, specifically for dialect differences within a country. A dialect is a geographical or social variety of a language. This variation is not random, but structured and conditioned by different extralinguistic factors such as geographical origin, socioeconomic status, education level, age group, occupation, gender and other social networks (Labov, 1973). Each linguistic variant used by a group has regularities, or common features used by individuals

in the group. Although an individual may use his own variants, it is in contact with other speakers of his community that he will find the limits for his individual variation (Beline, 2005). Thus, a dialect community is made up of speakers who share linguistic traits that distinguish their group from others, communicate more within the group and share norms and attitudes towards language use (Beline, 2005).

A dialect group can be identified essentially through the common use of phonetic, morphological and lexical linguistic variables (Feagin, 2002). This explains why we can usually guess where a person is from based on their accent: the combined usage of these linguistic variables, each with multiple variants, tend to be geographically bounded. It is precisely this idea that underlies all available dialect apps and surveys: users answer a questionnaire with distinct linguistic variables and variants that predicts where they are from based on their answer set.

Dialect quizzes have the potential to complement existing data, fill in gaps and provide information traditional methods cannot gather in a timely manner. They can also be an amusing way to crowdsource data and engage the public. Dialect apps and surveys have been applied to speakers of English in the United Kingdom (Leemann, Kolly, & Britain, 2018) and the United States (Katz & Andrews, 2013), to speakers of Spanish from various countries (Bouzouita, Castillo, & Pato, 2018) and to speakers of German in Switzerland (Leemann, Kolly, Purves, Britain, & Glaser, 2016) and in Europe (Leemann, Derungs, & Elspaß, 2019), all with great public reception.

Here we describe the creation of "Fuxiquera", a web-based dialect quiz for the Brazilian Portuguese (BP) language. Although mutually intelligible, BP is very distinct from European Portuguese, especially in its vocabulary, syntactic constructions and pronunciation (Bagno, 2015), and is considered a dialect of the latter language. Spoken by almost all 210 million Brazilians in a territory of 8.5 million $km^2$ (IBGE, 2019), BP in itself has currently

approximately 19 dialects formally described by sociolinguistic fieldwork (Aguilera, 2018). Nonetheless, BP is known to have a high degree of diversity and variability, due to both its territorial extension, which generates regional differences, and social divides, which generates socioeconomic differences (Bagno, 2015). Thus, we theorize there are at least twice that amount of BP dialects, and that they have not yet been formally described due primarily to the monetary difficulties, mobility strains and herculean academic effort such a countrywide survey requires.

There are currently 12 well-known and clearly described dialects in Brazil, which are: Amazofonia, Serra Amazônica, Nordestino, Sertanejo, Baiano, Mineiro, Fluminente, Carioca, Paulistano, Capiria, Sulista and Gaúcho (SM 2; Fig. S3). These dialects are well recognized by Brazilians; thus, they are part of the common linguistic knowledge in the country, and are at the center of people's affections, humor, conflicts and disputes (Ferreira & Faria, 2016). For instance, most Brazilians will instantly recognize a person from the Nordeste region, and consequently the Nordeste dialect, simply by how they pronounce their words, including the glottalization of /r/, the palatalization of /d/ and /t/, and the omission of the definite article before the proper noun (de Aragão, 1999; Mota, 2008).

With this citizen science effort, we aim to crowdsource language variation data for BP. The creation of this database, in a nationwide scale, will later allow us to study historic migration events that shaped dialects across continental-sized Brazil and many aspects of language change, such as BP evolution through time and space, explain spatial patterns that emerge using macroecological tools, and uncover biogeographic effects, such as beta and gamma diversity, of linguistic variants.

## 2. DATA

### *2.1 Linguistic data selection and collection*

We relied on different literature sources to select and exhaustively collect relevant linguistic variables. We searched the ISI Web of Knowledge, Portal Periódicos CAPES and Google Scholar databases by using a combination of the terms “dialeto*” OR “sotaque*” AND “Brasil*”, in Portuguese, and “dialect*” OR “accent*” AND “Brazil*”, in English. We also screened bibliographies of studies within our scope to find more dialectal studies. This search led to us using 81 published and unpublished studies (i.e., dissertations and thesis) to build our questionnaire and database. We drew specially from the work of ALiB (Atlas Linguístico do Brasil; https://alib.ufba.br), a geo-linguistic group that aims at creating a linguistic atlas for Brazil by first creating state and regional atlases (Cardoso & Mota, 2017). Once our questionnaire was built, we also used Twitter to crowdsource chosen lexical variables from geolocated tweets (Gonçalves & Sánchez, 2014). A list of all literature used in our study, our Data Source, is provided in the Supplementary Materials.

We chose linguistic variables known to be geographically distinct (Feagin, 2002) and that were broadly described in the literature. From all the BP linguistic variables available, we narrowed our quiz to 25 variables that represent wide-ranging differentiation and geographic boundaries within the country. Overall, our quiz had 9 lexical, 5 morphosyntactic, 9 phonetic and 2 morphosyntactic-lexical specific questions (Table S1).

Lexical questions focus mainly on what a person calls something, and we structured most of these around a picture of the chosen variable and asking users to choose a name for it from one of the multiple answers. For all lexical questions, users also have the option to write in a noun that is not among the given answers. Morphosyntactic questions focus on the use of standard grammar rules, such as pronoun preference, gender-specified words and verb tense. And finally, phonetic questions focus on the pronunciations of consonants and vowels. These were structured around users having to listen to audio recordings of different possible pronunciations and choosing the one that best resembles how they speak. We also included

questions that asked if a word rhymed with another, to diversify our asking methods. We aimed to keep our questionnaire short and enjoyable, as to not discourage informants to lose their spontaneity and interest through the quis (Leemann et al., 2018).

Brazil has 5570 municipalities distributed in 26 states and a Federal District. A municipality is an administrative division of a state and the smallest autonomous units of the country (IBGE, 2019). Municipalities are unevenly distributed in the country, with states in the North having a smaller population size and fewer and bigger municipalities compared to the rest of the country. For instance, the state of Minas Gerais (Southeastern region) has 853 municipalities, while the state of Roraima (Northern region) has only 15.

We found the previously chosen linguistic information for 780 municipalities, and even for these we were unable to find all the necessary variables listed in our quiz. We removed 24 municipalities with information for less than three variables, remaining 756 municipalities with language data in our initial database (Fig. 1). We believe this scenario further corroborates the language data gap within the country and validates our citizen science effort to crowdsource linguistic data. To work around this gap, we used Euclidean distance to fill municipalities without information with the data from the closest municipality with linguistic information. This created “dummy” municipalities, allowing us to work with a complete nationwide database. It is our expectation that as people take our quiz, municipalities with dummy information will be updated to real citizen provided information.

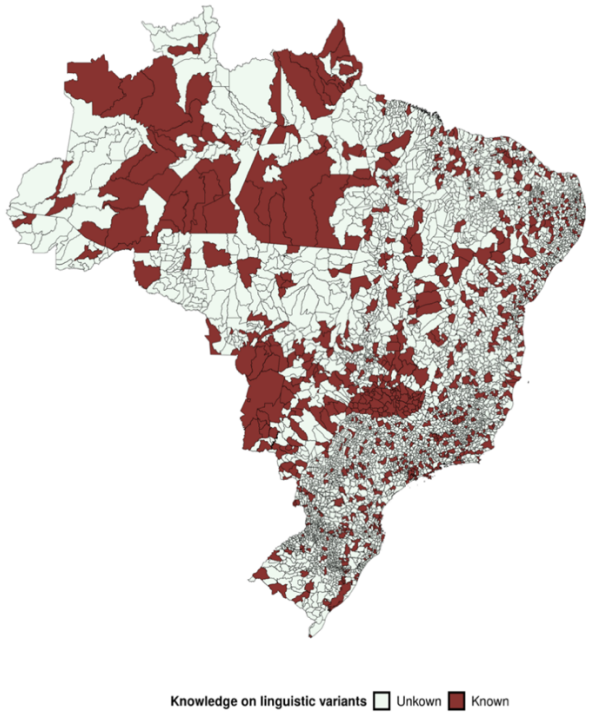


**Figure 1.** Map representing the data shortfall of our database. Municipalities in red have linguistic data available in the literature. Municipalities in white have no linguistic data available in the literature. Geographical boundaries of the Brazilian municipalities are according to the political-administrative division structure in force on the 1st of July, 2015 (IBGE, 2019).

*2.2. K-means cluster analysis*

We further investigated the spatial structure of the dummy dataset and applied a K-means clustering algorithm to identify clusters in the data. K-means is a centroid-based, unsupervised machine learning algorithm that partitions observations into k clusters with the nearest mean (Legendre & Legendre, 2012). The number of clusters are pre-defined subjectively, which is not a problem if there is domain knowledge available guiding the number of clusters to assume. Thus, we built both a two-cluster solution, to detect if the data had an expected north-south division, and a five-cluster solution, to check if the data captured the main known dialectal regions in the country.

All analyses in this section were performed using the R software (R Core Team 2017) and is publicly available on GitHub

(https://github.com/chrisborges/Clustering_dialects). The map for municipalities limits was downloaded in a 1:250.000 scale and represent the municipal political-administrative division of Brazil according to the structure in force on the 1st of July, 2015, the reference date of the 2015 Population Estimates (IBGE 2019).

The two-cluster solution captured the spatial structure of a northern superdialect (n = 2621 municipalities) and a southern superdialect (n=2951 municipalities; Fig. 2a). The five-cluster solution also captured the main regional superdialects (Fig. 2b), which we identified as the Northern (n = 1686), the Northeastern (n = 834), the Country (n = 1133), the Caipira (n = 530) and the Southern (n = 1389) superdialects. These dialects are commonly recognized in Brazil and widely described in the literature (Aguilera, 2018; Bagno, 2015).

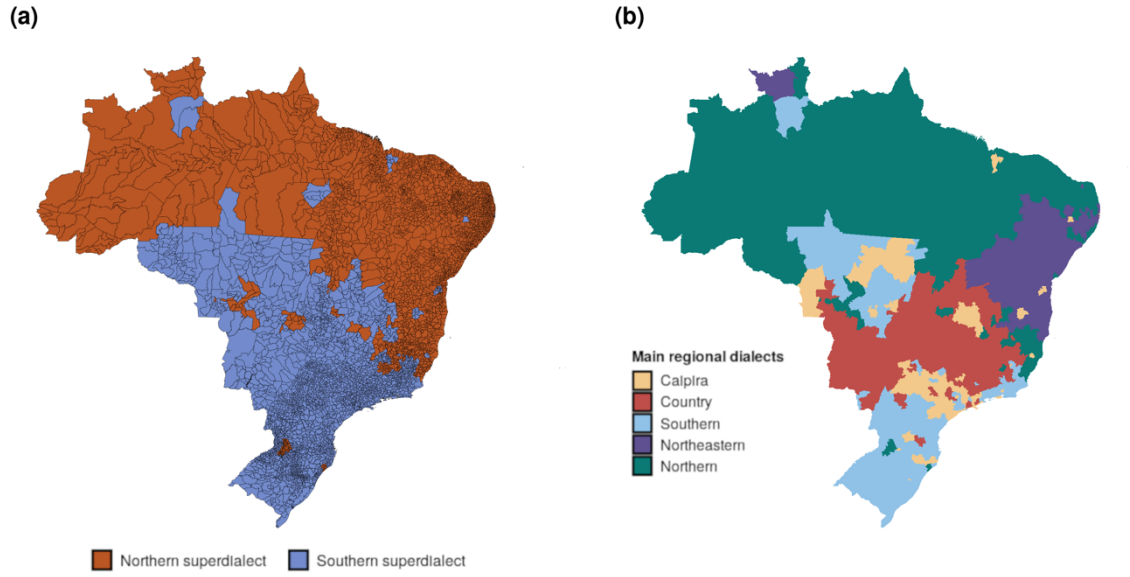


**Figure 2.** Geographical characterization of the clusters, where each colored group represents a superdialect. a) Two-cluster solution representing a Northern superdialect (green) and a Southern superdialect (blue). b) Fivecluster solution, representing the main regional superdialects: Northern (dark blue), Northeastern (purple), Country (orange), Caipira (yellow) and Southern (blue). Points are the center of municipalities.

The Northern dialect, spoken in the northern states, inherited its way of speaking mainly from the Portuguese colonizers. Because of this, a common trait of the dialect is the

raising of the middle vowel to a high vowel and a wheezing pronunciation of the /r/ and /s/ (Cruz-Cardoso, 2018). The Northeastern dialect, spoken in the coastal states in the northeast of Brazil, is quickly identified by the palatalization of the /d/ and /t/ consonants and the omission of a definite article before a first name (Mota, 2008). The Caipira dialect, is the inland São Paulo dialect, it emerged from the contact between Portuguese colonizers and native indigenous peoples (Pereira, 2014). This dialect was first described in 1884, is associated with a rural Brazil, and its main characteristic is the retroflex /R/. The Country dialect emerged from the contact between the Caipira speakers and migrants from other regions as they colonized the Midwest region of Brazil (Bueno, 2011). It shares many lexical and phonetic variables with the Caipira dialect, and is also associated with the rural and agricultural states. Finally, the Southern dialect is spoken in the three states that make up the Southern region of Brazil. This region was occupied mainly by Portuguese from the Azores island and many Italian migrants, each speaking their own Italian dialects (Görski, 2012). This mixture is reflected in the Southern dialect, and it stands out from other dialects specially because of the /e/ pronunciation at the end of the words.

Furthermore, the state of Mato Grosso (Midwestern region) displays a mixture of the Southern, Caipira, Northern and Country dialects (Fig. 2b). This is historically accurate since Mato Grosso is an agricultural frontier state and consequently experienced federal colonization programs from the 1930s until the mid 1980s (Arvor *et al.* 2018). This is a rather recent occupation and the different dialects are a reflection of this recent migration event, as the regional population has not had the time to fully form its own distinct dialect.

The Caipira dialect, specifically, shows up in many municipalities across the five regions in Brazil. This cluster structure is due to the retroflex /R/ present in those regions, though these misplaced cluster agglomerates in the Northern and Northeastern regions are probably due to the database's incompleteness.

## 3. A WAY FORWARD: THE WEB-BASED DIALECT QUIZ

### *3.1. Fuxiquera: a web-based dialect quiz*

We launched a web-based dialect quiz, called "Fuxiquera", to crowdsource data about the linguistic idiosyncrasies from as many Brazilians as possible. Our quiz predicts the municipality a person is from in Brazil based on answers to 25 questions regarding known lexical, morphosyntactic and phonetic variances present in specific regions. Previous dialect quizzes relied on historical linguistic data to underlie their questions (Katz & Andrews, 2013; Leemann et al., 2018). However, no such data exists for Brazil and available linguistic Atlases are geographically biased (Aguilera, 2018), with more data for the south and southeast regions in Brazil (where there are more universities and greater financial resources), greater sampling effort in the state's capitals, and few individuals interviewed per survey.

We created a web-based dialect quiz contrary to a smartphone application to guarantee accessibility and user-friendliness for all users within Brazil. Brazil is a developing country and we believe most users would not download an app to simply take a quiz. However, if that quiz came through a link in WhatsApp, a messaging application owned by Facebook and widely used in Brazil (Resende et al., 2019), the probability of users being interested could be much higher. Further, only 74.9% of Brazilian households had internet access in 2017, however, 93.2% of households had a mobile phone or smartphone (IBGE, 2019). Thus, our website is mobile friendly and we expect social media and WhatsApp to be its main diffusers.

We called our website "Fuxiquera", which is a noun that can be used for different objects or situations in BP and is a reoccurring option in our quiz. "Fuxiqueira" can be a name for a mandarin orange (*Citrus reticulata*), it can be a popular name for a parrot, and it can also be a way to call a loquacious or gossipy person. It is quite common in spoken

BP for monophthongization to occur, a process in which a diphthong is reduced to a single vowel (/ey/ > /e/) (Mota, 2008). This is why our name is spelled “Fuxiqu**e**ra” contrary to the grammar standard “Fuxiqu**ei**ra”, dropping the /i/ represents the verbal monophthongization of /ey/ to /e/. This was intentionally done to give users the idea that we are not interested in the grammar rules of standard Portuguese, but in how they speak when talking in a familiar and relaxed environment.

Further, our logo (Fig. S1) is a green parrot (*Amazona aestiva*) eating a mandarin orange (*Citrus reticulata*). To call someone a “parrot” in BP can also mean that someone is talkative, and Brazilians are aware of the lexical differences within the country for “mandarin orange” (Aguilera, 2009). Thus, our logo is supposed to tell users we want them to talk a lot about speech differences they may or may not be aware of.

The website lists our aims, who we are and how to contact us. It also has texts explaining what is a dialect, how it comes to be, and geographic speech differences within Brazil. Users can also navigate through a macro-scale dialect map and read about the main dialects found in Brazil (Fig. S3).

### *3.2. Algorithm behind the quiz*

Our goal is to match each answer set to the community (municipality) it has the most linguistic variants in common with. We created an algorithm that calculates a double-zero asymmetrical Sørensen Dissimilarity ($D_{sor}$) between any given set of answers and all municipalities. We chose this dissimilarity index because it is ideal for presence and absence data, and it gives a higher weight for double-presences and ignores double-absences (Legendre & Legendre 2012). The index scales values between 0 and 1, where values of 0 mean the answer set and the municipality are identical (they share all the same variants) and 1 means they are completely different. To summarize our information, we

transformed our data into $N \times K$ presence and absence matrices (PAMs), where $N$ are the linguistic variants and $K$ are the municipalities. Presence of a variant in a municipality were marked as one (1) and absences as zero (0).

In our quiz, users are required to choose only one answer that applies to how they speak. This information is saved in a vector of length 25. A $D_{sor}$ is calculated between each position of the answer vector and the corresponding PAM. Afterwards, we calculated an average between all 25 $D_{sor}$, resulting in an ordered distance matrix where the closer a dissimilarity is to zero (0) the higher the probability of the user being from that municipality. We used a threshold of 0.10 to determine the test's accuracy, where any distance smaller than 0.10 is displayed as the municipality the user is from. Results with dissimilarities above 0.10 will display the first three municipalities as possibilities. In both cases, respondents are enquired if the quiz guessed correctly and if not, for them to inform us of their correct birth location.

After the user takes the quiz and is given an answer, we ask permission to use their information for academic purposes, all the while keeping their anonymity. Users can agree or disagree by checking a box. If they agree, we then ask for their social information: gender, education level, age group, and the birthplace of their mother, father and spouse. We store each answer set and use it to continually update our PAMs. Because users can play around with the test, we wait for a municipality to have at least 5 answer sets before updating the PAMs.

All algorithms, the quiz and the website were written in the Python programming language (Python Software Foundation 2008), codes are publicly available on GitHub (https://github.com/guilhermebferreira/fuxiqueira). Final quiz's results are visually displayed as a map with the municipalities targeted with a circle (Fig. S2), using Leaflet, an open-source JavaScript library (Agafonkin 2019).

*3.3. Pilot study*

We performed a pilot study to test the efficacy of our quiz and help narrow the focus of the questions (Feagin, 2002). Overall we quizzed 45 people from all five official regional divisions in Brazil: the Midwest (n = 12), the Northeast (n = 13), the North (n = 3), the Southeast (n = 10) and the South (n = 7).

In our pilot, using only the data collected from the literature, our algorithm predicted only 6 municipalities correctly and 20 predictions were made for municipalities in the correct state. We chose 15 answer sets to update the database. Answer sets were chosen according to the user's birth municipalities, we gave preference to capitals and cities in known poorly surveyed states. After this update, our algorithm predicted 23 municipalities correctly and 9 municipalities in the correct state. When the database was updated with all user's answers it predicted 45 out of 45 precisely, as expected.

We also improved our quiz, by replacing three questions that were not adding meaningful information, since all participants answered the same option in the pilot. The final version of our quiz can be found in the Supplementary Material and (link).

## DISCUSSION

In this paper we described the use of simple distance- and dissimilarity-based methods to create an enjoyable dialectal quiz as a citizen science effort to crowdsource linguistic data throughout Brazil. To our knowledge, this is the first dialectal quiz created for Brazilian Portuguese and aimed specifically at a South American country.

As we collected data for our quiz, we noticed the current linguistic data shortfall, which further validated our efforts in creating this methodology as a way of linguistic data collection. Despite the data shortfall, our cluster solutions captured rather well the superdialects expected

from the literature. In the future we intend to re-run the cluster analyses with the completer database to see if the clustering changes. Our pilot study showed that improving the data does indeed make a difference in the algorithm's prediction outcome. Thus, we presume our quiz will continue to improve as people take it and share their answers with us.

## CONCLUSION

Language evolution occurs through the propagation of linguistic variants in a population (Sneller & Roberts, 2018), and dialect divergence is considered an initial stage of linguistic divergence (Honkola et al., 2018). Because language expresses culture, its evolution is not uniform and varies according to the groups that use it. Thus, to understand language change it is necessary to understand the social life of the community where it occurs, especially since social pressures are continuously operating upon language (Labov, 1973). On a spatial macroscale, this can be done through a citizen science project, engaging the public to share the linguistic variants they use and where they are from.

Our multidisciplinary work combines linguistic, information technology, ecological and geographical knowledge and methods to further the understanding of how languages change, evolve and are structured in space. By studying dialects as a measure of human diversity we are bridging the gap between natural and social sciences, and mainly bridging the gap between science and the public by bringing data science straight to their homes and contributing with their citizen empowerment through language identity.

## ACKNOWLEDGEMENTS

We thank Dr. Elisa Barreto for the witty and insightful suggestions to an earlier version of this manuscript.

## AUTHOR' CONTRIBUTIONS

C.B. and T.R. conceived the study. C.B. curated the data, wrote the quiz, analyzed and interpreted the data. Z.V. developed the website's frontend and its illustrations. G.B.V. developed the website's front and backend. C.B. wrote the first draft and all authors contributed to the final version.

## DATA AVAILABILITY

Data Source, a list of all the literature we used to build our quiz, is available in the Supplementary Material. Codes for data analysis can be found in GitHub (https://github.com/chrisborges/Clustering_dialects). The web-site's code, development and algorithm can be found in GitHub (https://github.com/guilhermebferreira/fuxiqueira).

## REFERENCES


Aguilera, V. de A. (2009). Léxico e áreas dialetais: o que podem demonstrar os dados do ALiB. In D. da Hora (Ed.), *VI Congresso Internacional da ABRALIN Anais* (pp. 4219–4233). ABRALIN 40 anos.

Aguilera, V. de A. (2018). A Geolingüística no Brasil: Estágio Atual. *Revista Da ABRALIN*, 5(1/2), 215–238. Retrieved from https://doi.org/10.5380/rabl.v5i1/2.52646

Bagno, M. (2015). *Preconceito Linguístico* (56th ed.). São Paulo: Parábola Editorial.

Beline, R. (2005). A variação linguística. In J. L. Fiorin (Ed.), *Introdução à Linguística: I. Objetos teóricos*. São Paulo: Contexto.

Bouzouita, M., Castillo, M., & Pato, E. (2018). Dialectos del Español. Una nueva aplicación para conocer la variación actual y el cambio en las variedades del español. *Dialectologia*, 20(April 2017), 61–83.

Bueno, E. S. da S. (2011). O falar do homem pantaneiro: um olhar sociolinguístico. *Ave Palavra*, 12(2), 1–47.

Cardoso, S. A. M., & Mota, J. A. (2017). Estudos geolinguísticos: caminhos seguidos no território brasileiro. *Lingüística*, 33(1), 89–105. Retrieved from https://doi.org/10.5935/2079-312X.20170006

Cruz-Cardoso, M. L. de C. (2018). O Atlas Linguístico do Amazonas – ALAM. In A. de Paula, D. K. Gomes, & E. F. B. da Silveira (Eds.), *Uma História de Investigações sobre a Língua Portuguesa: Homenagem a Silvia Brandão* (pp. 141–150). São Paulo: Blucher.

De Aragão, M. do S. S. (1999). A variação fonético-lexical em Atlas Lingüístico do Nordeste. *Revista do GELNE*, 1(2), 14-20.

Devictor, V., Whittaker, R. J., & Beltrame, C. (2010). Beyond scarcity: Citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions*, 16(3), 354–362. Retrieved from https://doi.org/10.1111/j.1472-4642.2009.00615.x

Feagin, C. (2002). Entering the community: Fieldwork. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The Handbook of Language Variation and Change* (pp. 20–39). Blackwell Publishing Ltd.

Ferreira, A. C. F., & Faria, J. P. de. (2016). Dialetos/Línguas do Brasil na desciclopédia. *RUA*, 22(2), 593. Retrieved from https://doi.org/10.20396/rua.v22i2.8647951.

Freitag, H., Pangantihon, C. V., & Njunjić, I. (2018). Three new species of Grouvellinus Champion, 1923 from Maliau Basin, Sabah, Borneo, discovered by citizen scientists during the first Taxon Expedition (Insecta, Coleoptera, Elmidae). *ZooKeys*, 754(754), 1–21. Retrieved from https://doi.org/10.3897/zookeys.754.24276

Fritz, S., See, L., Carlson, T., Haklay, M., Oliver, J. L., Fraisl, D., … West, S. (2019). Citizen science and the United Nations Sustainable Development Goals. *Nature Sustainability*, 2(10), 922–930. Retrieved from https://doi.org/10.1038/s41893-019-0390-3

Gonçalves, B., & Sánchez, D. (2014). Crowdsourcing Dialect Characterization through Twitter. *PLoS ONE*, 9(11), e112074. Retrieved from https://doi.org/10.1371/journal.pone.0112074

Görski, E. M. (2012). Fenômenos variáveis na Região Sul do Brasil: aspectos de comportamento sociolinguístico diferenciado entre as três capitais. *Estudos Linguísticos, São Paulo*, 41(2), 806–817.

Hartop, E. A., Brown, B. V., & Disney, R. H. L. (2015). Opportunity in our Ignorance: Urban Biodiversity Study Reveals 30 New Species and One New Nearctic Record for Megaselia (Diptera: Phoridae) in Los Angeles (California, USA). *Zootaxa*, 3941(4), 451–484. Retrieved from https://doi.org/10.11646/zootaxa.3941.4.1

Honkola, T., Ruokolainen, K., Syrjänen, K. J. J., Leino, U. P., Tammi, I., Wahlberg, N., & Vesakoski, O. (2018). Evolution within a language: Environmental differences contribute to divergence of dialect groups. *BMC Evolutionary Biology*, 18(1), 1–15. Retrieved from https://doi.org/10.1186/s12862-018-1238-6

IBGE, I. B. de G. e E. (2019). Cidades e Estados. Retrieved 4 November 2019, from https://cidades.ibge.gov.br/brasil/panorama

Katz, J., & Andrews, W. (2013). How y'all, youse and you guys talk. Retrieved 20 October 2019, from https://www.nytimes.com/interactive/2014/upshot/dialect-quiz-map.html?r=0

Kuchner, M. J., Faherty, J. K., Schneider, A. C., Meisner, A. M., Filippazzo, J. C., Gagné, J., … Stajic, T. (2017). The First Brown Dwarf Discovered by the Backyard Worlds: Planet 9 Citizen Science Project. *The Astrophysical Journal*, 841(2), L19. Retrieved from https://doi.org/10.3847/2041-8213/aa7200

Labov, W. (1973). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Leemann, A., Derungs, C., & Elspaß, S. (2019). Analyzing linguistic variation and change using gamification web apps: The case of German-speaking Europe. *PLoS ONE*, 14(12), 1–29. Retrieved from https://doi.org/10.1371/journal.pone.0225399

Leemann, A., Kolly, M.-J., & Britain, D. (2018). The English Dialects App: The creation of a crowdsourced dialect corpus. *Ampersand*, 5(August 2017), 1–17. Retrieved from https://doi.org/10.1016/j.amper.2017.11.001

Leemann, A., Kolly, M.-J., Purves, R., Britain, D., & Glaser, E. (2016). Crowdsourcing Language Change with Smartphone Applications. *PLOS ONE*, 11(1), e0143060. Retrieved from https://doi.org/10.1371/journal.pone.0143060

Legendre, P., & Legendre, L. (2012). *Numerical Ecology* (3rd ed.). Elsevier.

Mota, J. A. (2008). Como fala o nordestino: a variação fônica nos dados do Projeto Atlas

Lingüístico do Brasil. In M. C. Lima-Hernandes, M. J. Marçalo, G. Micheletti, & V. L. de R. Martin (Eds.), *A língua portuguesa no mundo* (v. 1). São Paulo: FFLCH-USP.

Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., & Crowston, K. (2012). The future of citizen science: emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, 10(6), 298–304. Retrieved from https://doi.org/10.1890/110294

Pereira, D. T. (2014). O uso do termo e do dialeto caipira nos jornais do século XIX (1838-1884). *Revista Ars Historica*, (7), 169–179.

Resende, G., Melo, P., Sousa, H., Messias, J., Vasconcelos, M., Almeida, J., & Benevenuto, F. (2019). (Mis)Information Dissemination in WhatsApp: Gathering, Analyzing and Countermeasures. In *The World Wide Web Conference on - WWW '19* (pp. 818–828). New York, New York, USA: ACM Press. Retrieved from https://doi.org/10.1145/3308558.3313688

Sneller, B., & Roberts, G. (2018). Why some behaviors spread while others don't: A laboratory simulation of dialect contact. *Cognition*, 170(October 2017), 298–311. Retrieved from https://doi.org/10.1016/j.cognition.2017.10.014

Soroye, P., Ahmed, N., & Kerr, J. T. (2018). Opportunistic citizen science data transform understanding of species distributions, phenology, and diversity gradients for global change research. *Global Change Biology*, 24(11), 5281–5291. Retrieved from https://doi.org/10.1111/gcb.14358

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., … Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, 169, 31–40. Retrieved from https://doi.org/10.1016/j.biocon.2013.11.003

Theobald, E. J., Ettinger, A. K., Burgess, H. K., DeBey, L. B., Schmidt, N. R., Froehlich, H. E., … Parrish, J. K. (2015). Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation*, 181, 236–244. Retrieved from https://doi.org/10.1016/j.biocon.2014.10.021

Tiago, P., Pereira, H. M., & Capinha, C. (2017). Using citizen science data to estimate climatic niches and species distributions. *Basic and Applied Ecology*, 20, 75–85. Retrieved from https://doi.org/10.1016/j.baae.2017.04.001

## DATA SOURCES


Aguilera, V. de A. (2009). Léxico e áreas dialetais: o que podem demonstrar os dados do ALiB. In D. da Hora (Ed.), *VI Congresso Internacional da ABRALIN Anais* (pp. 4219–4233). ABRALIN 40 anos.

Aguilera, V. de A., & Silva, H. C. da. (2011). Dois momentos do /r/ retroflexo em Lavras - MG: no Atlas Linguístico de Minas Gerais e nos dados do projeto do Atlas Linguístico do Brasil. *Revista Diadorim*, *8*(1), 125–142. doi:10.35520/diadorim.2011.v8n1a7962

Alvim, A. F. D., Rodrigues, B., Oliveira, C. R. de, Dias, C. de O., Lourenzoni, M. Q., & Pinto, V. M. R. (2016). O falar carioca, paulista e caipira: Análise fonética e fonológica. *SóLetras*, 234–240.

Andrade, C. Q. (2015). *A fala brasiliense: origem e expansão do uso do pronome 'tu'*. Universidade de Brasília.

Augusto, V. L. D. dos S. (2012). *Atlas Semântico-Lexical do estado de Goiás Tomo II*. Universidade de São Paulo.

Aurélio, R. P. (2012). *Os falares da Bahia e do Espírito Santo: Implicações sob os aspectos dialetológicos*. Universidade Federal do Espírito Santo.

Barboza, C. L. F. (2016). A Difusão Das Africadas Pós-Alveolares Em Falares Do Português Brasileiro. *ReVEL*, *14*(27), 115–150.

Battisti, E., & Dornelles Filho, A. A. (2010). A palatalização variável das oclusivas alveolares num falar de português brasileiro e sua análise pela Teoria da Otimidade. *Letras de Hoje*, *45*(1, jan./mar.), 80–86.

Benke, V. C. M. (2012). *Tabus Linguísticos Nas Capitais Do Brasil: Um Estudo Baseado Em Dados Geossociolinguísticos*. Universidade Federal de Mato Grosso do Sul.

Bertoldo, S. R. F. (2007). *Estudo semântico-lexical no Distrito Nossa Senhora da Guia*. Universidade de São Paulo.

Bianchin, V. (2015, July). O certo é "biscoito" ou "bolacha"? *Superinteressante*. Retrieved from https://super.abril.com.br/mundo-estranho/o-certo-e-biscoito-ou-bolacha/

Bisol, L., & Collischonn, G. (Eds.). (2009). *Português do sul do Brasil: variação fonológica*. Porto Alegre: EDIPUCRS.

Braga, L. M. (2012). *Ausência/presença de artigo definido diante de antropônimos na fala dos moradores de Mariana e Uberaba – MG*. Universidade Federal de Uberlândia.

Brandão, S. F. (2007). Nas trilhas do -R retroflexo. *Signum: Estudos Da Linguagem*, *10*(2), 265. doi:10.5433/2237-4876.2007v10n2p265

Bueno, E. S. da S. (2011). O falar do homem pantaneiro: um olhar sociolinguístico. *Ave Palavra*, *12*(2), 1–47.

Campos-Júnior, H. da S. (2011). *A Variação Morfossintática Do Artigo Definido Na Capital Capixaba*. Universidade Federal do Espírito Santo.

Cardoso, S. A. M., & Mota, J. A. (2013). Percursos da Geolinguística no Brasil. *Lingüística*, *29*(1), 115–142.

Cardoso, S. A. M., & Mota, J. A. (2017). Estudos geolinguísticos: caminhos seguidos no território brasileiro. *Lingüística*, *33*(1), 89–105. doi:10.5935/2079-312X.20170006

Cardoso, S. A., & Mota, J. A. (2012). Projeto Atlas Linguístico do Brasil: Antecedentes e estágio atual. *Alfa, São Paulo*, *56*(3), 855–870.

Castro, V. S. (2013). O " r caipira " em Mato Grosso do Sul – estudo baseado em dados do ALMS, Atlas linguístico do Mato Grosso do Sul. *Estudos Linguísticos, São Paulo*, *42*(1), 566–575.

Cavalcante, R. (2006). *CearensÊs: a cultura do povo cearense*. Clube de Autores.

Coelho, M. do S. V. (2012). Aspectos da linguagem falada pelos Gurutubanos. *Revista (CON)TEXTOS Linguísticos*, *6*(6), 95–114.

Costa, D. D. S. S. (2018). *Vocabulário Dialetal do Centro-Oeste: Interfaces entre a Lexicografia e a Dialetologia*. Universidade Estadual de Londrina.

Costa, L. B. da. (2013). *Variação nos pronomes 'Tu'/"Você" nas capitais do norte*. Universidade Federal do Pará. Retrieved from http://ir.obihiro.ac.jp/dspace/handle/10322/3933

Cristianini, A. C. (2007). *Atlas Semântico-Lexical da Região do Grande ABC*. Universidade de São Paulo.

Cruz-Cardoso, M. L. de C. (2018). O Atlas Linguístico do Amazonas – ALAM. In A. de Paula, D. K. Gomes, & E. F. B. da Silveira (Eds.), *Uma História de Investigações sobre a Língua Portuguesa: Homenagem a Silvia Brandão* (pp. 141–150). São Paulo: Blucher.

Cruz, R. (2012). Vogais na Amazônia Paraense. *Alfa, São Paulo*, *56*(3), 945–972.

Cuba, M. A. (2015). *Atlas linguístico topodinâmico do território incaracterístico*. Universidade Estadual de Londrina.

de Aragão, M. do S. S. (1984). *Atlas Linguístico da Paraíba: cartas léxicas e fonéticas*. Brasília: UFPB/CNPq, Coordenação Editorial.

de Aragão, M. do S. S. (2011). Atlas Linguísticos Regionais Brasileiros: Itens lexicais sinônimos e parassinônimoa. *Acta Semiótica et Lingvistica*, *16*(35), 27–48.

de Aragão, M. do S. S. (2014a). Ditongação e monotongação nas capitais brasileiras. In *XVII Congresso Internacional Asociación de Lingüistica y Filología de América Latina* (pp. 2089–2101). João Pessoa.

de Aragão, M. do S. S. (2014b). Sinónimos e parassinônimos em capitais do nordeste brasileiro: dados do ALiB. *Acta Semiótica et Linguística*, *19*, 7–20.

Dias, E. C. O. (2010). Uso variável das oclusivas alveolares /t, d/ em Florianópolis. *Work. Pap. Linguíst.*, (especial), 1–19.

Dias, M. R., & Viegas, M. do C. (2017). Vogais médias pretônicas: falares mineiros. *Caligrama, Belo Horizonte*, *22*(2), 5–31. doi:10.17851/2358-9787.22.2.5-31

Erig, G. A., Mesquita, W. R. De, Mendes, D. T., & Morais, F. A. da S. (2015). *Variações Linguísticas e a sua relação com o turismo: o caso de Palmas, Tocantins*. Instituto Federal do Tocantins.

Ferreira, A. B. H. (1999). *Novo Aurélio Século XXI: o dicionário da língua portuguesa* (2nd ed.). Rio de Janeiro: Nova Fronteira.

Ferreira, J. A. (2015). *Jogos e diversões infantis: um estudo geossociolinguístico na região norte do Brasil*. Universidade Federal do Pará.

Galvão Maia, E. (2018). Enfraquecimento do /S/ em coda silábica em dados do sul do Amazonas (Brasil). *Estudos de Lingüística Galega*, *I*, 219–236. doi:10.15304/elg.ve1.3593

Görski, E. M. (2012). Fenômenos variáveis na Região Sul do Brasil: aspectos de comportamento sociolinguístico diferenciado entre as três capitais. *Estudos Linguísticos, São Paulo*, *41*(2), 806–817.

Graebin, G. de S. (2008). *A Fala De Formosa/GO: A Pronúncia Das Vogais Médias Pretônicas*. Universidade de Brasília.

Justiniano, J. dos S. (2012). *Atlas linguístico dos falares do alto rio negro - ALFARiN*. Universidade Federal do Amazonas.

Karim, J. M. (2004). *A variação na concordância de gênero no falar da comunidade de CáceresMT*. Universidade Estadual Paulista.

Lee, S. H. (2006). Sobre as vogais pré-tônicas no Português Brasileiro. *Estudos Lingüísticos*, *35*(1), 166–175.

Medeiros, J. C. (2018). *Atlas morfossintático de parte da microrregião do rio Negro-Solimões - AMPRINES - Vol. I*. Universidade Federal do Amazonas.

Mendes, S. T. do P. (2002). Qualificativo 'Dona' e nomes próprios: análise diacrônica de dados do português mineiro de Barra Longa-MG. In M. A. A. M. Cohen & J. M. Ramos (Eds.), *Dialeto mineiro e outras falas: estudos de variação e mudança linguística* (p. 197). Belo Horizonte: Faculdade de Letras, UFMG.

Mendonça, T. A. (2017). *Atlas Linguístico de Icatu (ALinI)*. Universidade Federal do Maranhão.

Milani, S. E., & Silva, D. M. da. (2017). Fatos fonéticos e fonológicos constatados na pesquisa do atlas linguístico de Goiás - ALINGO. In *Simpósio 25 - Demonstração dos usos, normas e identidades linguísticas locais* (pp. 589–606). Atas do V SIMELP - Simpósio Mundial de Estudos de Língua Portuguesa. doi:10.1285/i9788883051272p606

Mota, J. A., & Cardoso, S. A. M. (Eds.). (2006). *Documentos 2: projeto atlas linguístico do Brasil*. Salvador: Quarteto.

Mota, J. A., & Lopes, P. H. de S. (2018). Os subfalares do Norte e o traçado das vogais médias pretônicas. *Estudos de Lingüística Galega*, *I*, 209–218. doi:10.15304/elg.ve1.3480 Neiva, I. (2017a). *Vocabulário Dialetal Baiano V.1*. Universidade Federal da Bahia.

Neiva, I. (2017b). *Vocabulário Dialetal Baiano V.2*. Universidade Federal da Bahia.

Nogueira, F. M. da S. B. (2013). *Como os falantes de Feira de Santana e Salvador tratam o seu interlocutor?* Universidade Federal da Bahia.

Oliveira, D. P. de. (2006). O atlas lingüístico de Mato Grosso do Sul. *Signum: Estudos Da Linguagem*, *9*(2), 169. doi:10.5433/2237-4876.2006v9n2p169

Oliveira, D. P. de (Ed.). (2007). *Atlas Linguístico de Mato Grosso do Sul (ALMS)*. Campo Grande, MS: Ed. UFMS.

Oliveira, J. M. De, Paim, M. M. T., & Ribeiro, S. S. C. (2018). A importância do Atlas linguístico do Brasil para o ensino de português. *Revista Tabuleiro de Letras*, *12*(3), 2176– 5782.

Pacheco, C. da S. (2010). *Padrões sociolinguísticos da concordância de gênero na baixada cuiabana*. Universidade de Brasília.

Pereira, M. das N. (2007). *Atlas geolingüistico do litoral potiguar AliPTG*. Universidade Federal do Rio de Janeiro.

Philippsen, N. I. (2013). *A Constituição do Léxico Norte Mato-Grossense na Perspectiva Geolinguística: Abordagens Sócio-Semântico-Lexicais*. Universidade de São Paulo.

Pinheiro, I. M. G. (2016). *Aspectos fonológicos do português do sul de goiás*. Universidade Federal de Goiás.

Razky, A., Ribeiro, C., & Sanches, R. (2017). O projeto atlas linguístico do Amapá (ALAP): Caminhos percorridos e estágio atual. *Alfa, São Paulo*, *61*(2), 303–317. doi:10.1590/19815794-1709-3

Razky, A., & Sanches, R. D. (2016). Variação geossocial do item lexical riacho/córrego nas capitais brasileiras. *Gragoatá, Niterói*, *40*, 70–89.

Reis, M. G. S. dos. (2010). Fenômenos linguísticos característicos do português arcaico na fala do Alto Pantanal. *Cadernos Do CNLF*, *XIV*(4), 2705–2714.

Ribeiro, S. S. C. (2012). *Brinquedos e brincadeiras infantis na área do falar baiano v. 1*. Universidade Federal da Bahia.

Ribeiro, T. L. (2016). Pesquisa Geossociolinguística no Norte do Estado do Paraná: A variação lexical na rota do café. *Cadernos Do CNLF*, *XX*(12), 191–211.

Rodrigues, A. N. (1974). *O dialeto caipira na região de piracicaba*. São Paulo: Ática.

Romano, V., & Aguilera, V. (2014). Padrões de variação lexical na região Sul a partir dos dados do Projeto Atlas Linguístico do Brasil. *Estudos Linguísticos, São Paulo*, *43*(1), 575–587.

Romano, V. P. (2012). *Atlas Geossociolinguístico de Londrina: um estudo em tempo real e tempo aparente*. Universidade Estadual de Londrina.

Romano, V. P. (2018a). Áreas lexicais brasileiras: um novo olhar sobre a proposta de Antenor Nascentes nos dados do projeto atlas linguístico do Brasil. *Lingüística*, *34*(1), 117–145. doi:10.5935/2079-312X.20180007

Romano, V. P. (2018b). Áreas lexicais no Centro-Sul do Brasil sob uma perspectiva geolinguística. *Revista De Estudos Da Linguagem*, *26*(1), 103–145. doi:10.17851/22372083.26.1.103-145

Romano, V. P., & Seabra, R. D. (2014). Menino, Guri Ou Piá? Um Estudo Diatópico Nas Regiões Centro-Oeste, Sudeste E Sul a Partir Dos Dados Do Projeto Atlas Linguístico Do Brasil. *Alfa, São Paulo*, *58*(2), 463–497. doi:10.1590/1981-5794-1405-9

Romano, V. P., & Seabra, R. D. (2015). Dados geolinguísticos sob uma perspectiva estatística: a variação lexical no Centro-Oeste, Sudeste e Sul do Brasil. *Revista de Estudos Da Linguagem*, *22*(2), 59–92. doi:10.17851/2237-2083.22.2.59-92

Romano, V. P., & Seabra, R. D. (2017). Do presente para o passado: a variação lexical em Minas Gerais a partir de corpora geolinguísticos sobre brinquedos infantis. *Revista De Estudos Da Linguagem*, *25*(1), 111. doi:10.17851/2237-2083.25.1.111-150

Sá, E. J. de. (2013). *Atlas Linguístico de Pernambuco (ALiPE)*. Universidade Federal da Paraíba.

Sanches, R. D. (2015). *Variação Lexical Nos Dados Do Projeto Atlas Geossociolinguístico do Amapá*. Universidade Federal do Pará.

Sanches, R. D., Moreira, S. T., & Razky, A. (2018). Designações para Riacho/Córrego na região norte do Brasil. *Entreletras, Araguaína*, *9*(2), 466–479.

Santos-Ikeuchi, A. C. dos. (2014). *Atlas linguístico topodinâmico do oeste de São Paulo*. Universidade Estadual de Londrina.

Silva, G. A. da. (2018). *Atlas linguístico topodinâmico e topoestático do estado do Tocantins (ALITTETO)*. Universidade Estadual de Londrina.

Soares, R. de C. da S. (2012). *Atlas semântico-lexical da região norte do alto Tietê (ReNAT) - São Paulo*. Universidade de São Paulo.

Souza, G. G. A. (2016). *Palatalização de oclusivas alveolares em Sergipe*. Universidade Federal de Sergipe.

Tavares, L. S. (2013). *Atlas Morfossintático da Microrregião do Madeira - AMSIMA*. Universidade Federal do Amazonas.

Vilefort, M. T. C. (1985). *Aspectos Sintáticos do Dialeto Caipira na Região de Morrinhos*. Goiânia, GO: Kelps.

SUPPLEMENTARY MATERIAL 1

## Supplementary figures

Figure S1. Logo for Fuxiquera, our web-based citizen science project.

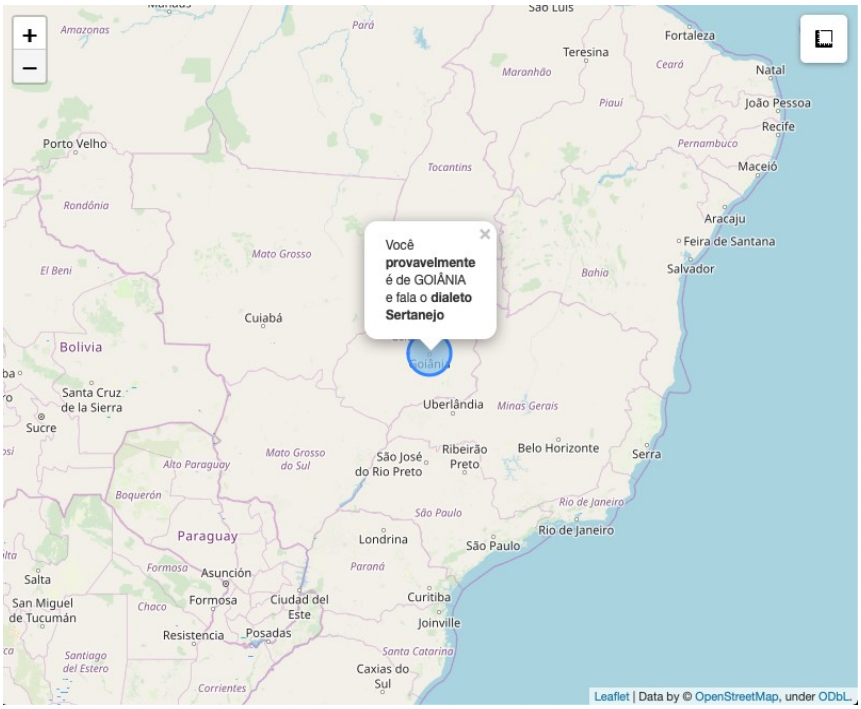


Figure S3. Visual result of the Fuxiquera dialect quiz. The Portuguese text reads "You are probably from Goiânia and speaks the Country dialect", users will then be asked to confirm or refute this statement.

**Supplementary tables**

Table S1. Lexical, morphosyntactic and phonetic linguistic variables chosen for the quiz.

| Linguistic variable | Variable (in English) | Example variants (in Portuguese) | N | Type |
|---|---|---|---|---|
| 1. Lexical variation | *boy* | Garoto; guri; piá | 17 | Lexical |
| 2. Lexical variation | *small river* | Corgo; riacho; | 33 | Lexical |
| 3. Lexical variation | *mandarin orange* | Mexerica; bergamota | 23 | Lexical |
| 4. Lexical variation | *frozen juice in a plastic bag* | Sacolé; geladinho | 22 | Lexical |
| 5. Lexical variation | *marbles* | China; biloca | 36 | Lexical |
| 6. Interjection preference | *yikes* | Ôxe; uai; bah | 23 | Morphosyntactic/ Lexical |
| 7. Lexical variation | *bread* | Pão francês; pão jacó | 16 | Lexical |
| 8. Lexical variation | *greedy* | Mão de vaca; pão duro | 43 | Lexical |
| 9. Pronoun preference. Verbal agreement with chosen pronoun in imperative mode | *You go* | Tu vai; tu vais; você vai; | 5 | Morphosyntactic |
| 10. Absence or presence of the feminine article defined before anthroponyms | *Maria's book* | De Maria; da Maria | 3 | Morphosyntactic |
| 11. Absence or presence of the masculine article defined before anthroponyms | *Carlos' book* | De Carlos; do Carlos | 3 | Morphosyntactic |
| 12. Gender of the noun. Demo article. | *The lettuce* | A alface; O alface | 2 | Morphosyntactic |
| 13. Gender morphosyntax | The female of *thief* | Ladra; ladrona | 4 | Morphosyntactic /Lexical |
| 14. Lexical variation | *Cookie* | Biscoito; bolacha | 2 | Lexical |
| 15. Verb preference for existential constructions | *There was* | Tinha; havia | 4 | Morphosyntactic |
| 16. Diphthongation before /s/ and variation of /s/ in syllable coda | Pronunciation of *three* | Diphthongation (Tre[js]); Palatal (Tre[ʃ]) | 5 | Phonetic |
| 17. Pronunciation of the pretonic mean vowel /o/ | *Tomato* | Open-mid ([tõ]mate); Closemid ([tʊ]mate) | 2 | Phonetic |
| 18. Monotongation of /ow/. Glottal /r/ | *Beetle* | bes[ow]ro; bes[oh]o | 2 | Phonetic |

| | | | | |
|---|---|---|---|---|
| 19. Diminutive use, vocabulary reduction, sound of /t/ and pronunciation of /r/ in syllable coda | Pronunciation of *little closer* | Palatized consonant and retroflex /r/ (Pe[ɽ][tʃi]inho); Alveolar occlusive and glottal /r/ (Pe[ɦ][Ti]nho) | 10 | Phonetic |
| 20. Pronunciation of /t/ | Pronunciation of *milk* | Palatized (Lei[tʃi]); Alveolar (Lei[Ti]) | 4 | Phonetic |
| 21. Pronunciation of /r/ in a syllable coda | Pronunciation of *green* | Velar (ve[x]de); Erasure (Ve[ø]de) | 7 | Phonetic |
| 22. Pronunciation of /r/ in the final position of an infinitive verb | Pronunciation of *to travel* | Alveolar tap (viaja[ɾ]); Erasure (viajá[ø]); | 5 | Phonetic |
| 23. Pronunciation of /s/ in coda syllable | Pronunciation of *indeed* | Palatal (me[ʃ]mo); Glotal fricative (me[ɦ]mo) | 6 | Phonetic |
| 24. Pronunciation of the pretonic mean vowel /e/ | Pronunciation of *television* | Open-mid (t[ɛ]l[ɛ]vizãw); Close-mid (t[e]l[e]vizãw); | 2 | Phonetic |
| 25. Lexical variation | *manihot* | Mandioca; aipim | 14 | Lexical |

N = number of known variants.

SUPPLEMENTARY MATERIAL 2

**Supplementary figure and text**

There are currently 12 well-known and clearly described dialects in Brazil (Fig. S3). Below we briefly describe each dialect (in Portuguese), a description that is also the accompanying text to the interactive figure below.

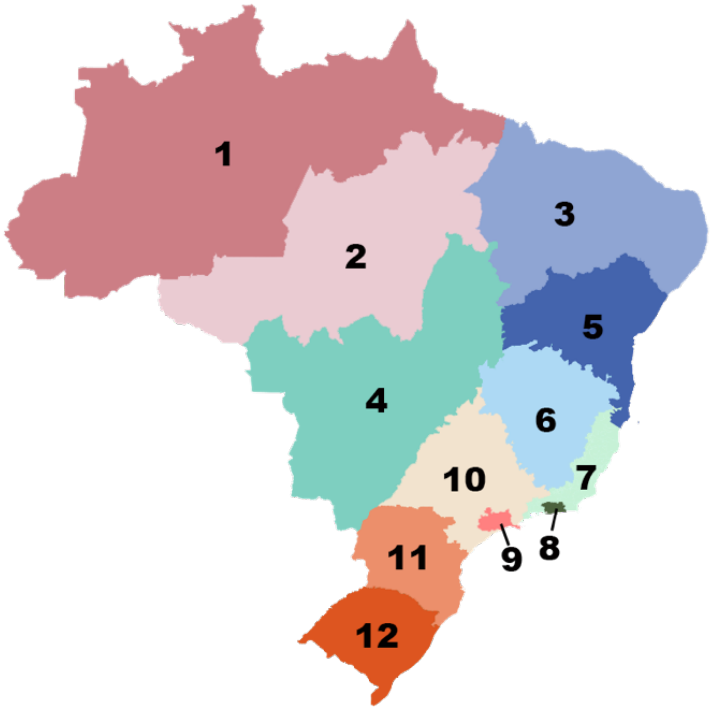


Figure S3. The main Brazilian Portuguese dialects found in Brazil. 1. Amazofonia 2. Serra Amazônia 3. Nordestino 4. Sertanejo 5. Baiano 6. Mineiro 7. Fluminense 8. Carioca 9. Paulistano 10. Caipira 11. Sulista 12. Gaúcho. On the website, upon a click, a description (in Portuguese) will be shown for each dialect.

**1. AMAZOFONIA**

*Esse dialeto é uma chibata no balde!*

Carapanã, égua, pai d'égua, até o tucupi, de bubuia, pitiú, baiacu, banzeiro, são apenas alguns exemplos de palavras e expressões empregadas no dialeto Amazofonia.

O dialeto Amazofonia ou Nortista, é um dialeto com aproximadamente 12 milhões de falantes. Concentrado na região norte do Brasil, ele ocorre em 5 estados: Acre, Amazonas, Roraima, Amapá e parcialmente no Pará.

Devido as diferentes colonizações portuguesas na região, esse dialeto herdou bastante do falar dos portugueses e um pouco do falar indígena e nordestino. Por causa da herança portuguesa, há muito em comum entre esse dialeto e o dialeto carioca, como a pronuncia do /r/ e /s/ chiados e o alteamento da vogal média /o/ em uma vogal alta /u/ (tomate = tumate). No entanto, há diferenças, como o uso adequado da norma culta no dialeto nortista, por exemplo, conjugando o tu como “tu fizeste” e “tu és”.

## 2. SERRA AMAZÔNICA

*Esse dialeto é só o milho da pipoca!*

Fim da rosca, pisero, pocar, torar a jaca, rampubanhu, apombalhado, calciná as ideias, são apenas alguns exemplos de palavras e expressões empregadas no dialeto serra amazônica.

O dialeto Serra Amazônica, também chamado de dialeto do arco do desflorestamento, é falado em Rondônia, sudeste do Pará, sudoeste do Maranhão, norte do Mato Grosso e em Tocantins.

O dialeto é justamente formado pela migração desordenada, na década de 1970, de nordestinos, goianos, sudestinos e sulistas que ocuparam a região em busca de terras baratas para o agronegócio. Assim, ele se diferencia do dialeto Amazofonia e Nordestino pela sua pronuncia mais próxima dos dialetos Caipira e Sertanejo. Uma característica desse dialeto é a forte pronuncia do /s/, similar ao dialeto Paulista.

## 3. NORDESTINO

*Êta dialeto arretado!*

Oxente, de hoje a oito, fulero, galego, pirangueiro, zoada, são apenas alguns exemplos de palavras e expressões empregadas no dialeto Nordestino.

O dialeto Nordestino possui aproximadamente 50 milhões de falantes. Sua origem se dá em Recife, capital de Pernambuco e hoje ocupa praticamente todos os estados do Nordeste: Paraíba, Alagoas, Rio Grande do Norte, Ceará, Piauí, Maranhão e Sergipe.

Características marcantes desse dialeto incluem a palatalização de /d/ e /t/ (por exemplo a pronuncia de noite como “noiti” e não “noitchi”), a glotalização do /r/, que é pronunciado com som de /h/ e a omissão de artigo definido antes de nome próprio (não se diz “A Maria me deu” e sim “Maria me deu”). O dialeto Nordestino é bastante rico em léxicos, com o Dicionário do Nordeste contendo cerca de 5 mil palavras e expressões próprias da região.

**4. SERTANEJO**

*Rensga, esse dialeto é só o ouro!*

Corguim, lereia, boi de piranha, tem base um trem desse?, anéin, paia, quando é fé, dar rata, são apenas alguns exemplos de palavras e expressões empregadas no dialeto sertanejo.

Muito além de batizar um ritmo musical para quem está na sofrência amorosa, o dialeto sertanejo é falado por aproximadamente 33 milhões de pessoas. Esse falar é encontrado principalmente na região Centro-Oeste do Brasil, nos estados do Mato Grosso, Mato Grosso do Sul, Goiás e em partes de Minas Gerais.

O dialeto se originou a partir do dialeto Caipira, que se modificou sob influência de imigrantes de outras regiões (dialetos Paulista, Mineiro, Sulista e Nordestino), principalmente após a construção de Brasília e das rodovias Belém-Brasília e BR-364. É por isso que existe

tanta similaridade entre os sotaques Sertanejos e Caipira, como o uso do “uai”, “trem”, “mio”, e a possibilidade do /r/ puxado.

No dialeto Sertanejo há sempre a palatalização de /di/ e /ti/ (exemplo: dente = dentche), as fricativas /s/ e /z/ nunca são palatalizadas e há alternância na pronuncia do /r/ podendo ser puxado ou surdo (ve**R**de ou verde).

**5. BAIANO**

*Man, esse dialeto broca!*

Barril dobrado, miserê, aooonde, lá ele, num tô comendo reggae de ninguém, batê o baba, são apenas alguns exemplos de palavras e expressões empregadas no dialeto Baiano.

O dialeto Baiano é considerado o primeiro dialeto brasileiro, uma vez que Salvador foi a primeira sede do Brasil Colônia e consequentemente sofreu influências de ondas migratórias europeias, de povos indígenas e africanos. Hoje ele é falado por aproximadamente 12 milhões de pessoas e engloba além do estado da Bahia, parte de Sergipe, norte de Minas Gerais, leste de Goiás e Tocantins.

Têm como características a inversão da colocação negativa (“Não sei” vira “Sei não”), a ausência do artigo definido antes de nome próprio e preposições (ex. “O João me ligou” = “João me ligou”, e “O carro do Mario” = “O carro de Mario”), traço que influenciou posteriormente o dialeto Nordestino. No entanto, se diferencia do dialeto Nordestino principalmente pela pronuncia de /d/ e /t/ palatalizados (pronuncia de noite como “noitchi”).

**6. MINEIRO**

*Ô trem bão dimái da conta esse dialeto sô!*

Inté, nossinhora, logali, trem, muié, bão tamém, são apenas alguns exemplos de palavras e expressões empregadas no dialeto Mineiro.

Para um bom entendedor, basta meia palavra. Aqui impera a lei do menor esforço, por isso o dialeto Mineiro é reconhecível pelo encurtamento das palavras, onde “massa de tomate” é pronunciado “mass tumátch” e os clássicos “onde que eu estou, para onde eu vou” são “oncotô, oncovô”.

O dialeto Mineiro, também chamado de Montanhês, é falado na região central de Minas Gerais, com aproximadamente 18 milhões de falantes. Ele surge em 1800, após a decadência da mineração na região. Pode ter alguma sobreposição com o dialeto caipira, mas são maneiras de falar distintas.

Suas características mais reconhecíveis são o uso amplo do diminutivo seguido de transformação do /inho/ em /im/ no final nas palavras (tadinho = tadim, pertinho = pertim), alguns hiatos passam a ser vogais longas (fio = fíi) e permutação de /e/ em /i/ e /o/ em /u/ (domingo = dumingu).

**7. FLUMINENSE**

*Esse dialeto é beleza pura!*

Gastura, qualé, komboza, pagar vecha, pocar, eitá preula, taruíra, são apenas alguns exemplos de palavras e expressões empregadas no dialeto Fluminense.

O dialeto Fluminense é falado nos estados do Rio de Janeiro e Espírito Santo, com aproximadamente 17 milhões de falantes. Já chegou a ser declarado como a pronúncia-padrão do Português e é o dialeto usado em telejornais nacionais.

Assim como o dialeto carioca, o dialeto fluminense se origina a partir da chegada da corte portuguesa ao Brasil, em 1808, e com a mudança da sede oficial de Salvador para o Rio

de Janeiro. Por isso, apresenta traços em comum com o português europeu, como reduzir as vogais /e/ para /i/ e /o/ para /u/. Outros traços marcantes são o /r/ aspirado e a ditongação das vogais (três = treis, mas = mais). Se difere do dialeto Carioca na pronuncia do /s/, que não é chiado e no uso do pronome "você".

## 8. CARIOCA

*Dá uma moral nesse dialeto sinistro, aíh!*

Asfalto, caô, irado, mermão, vacilão, caraca, já é, lek, queimar a largada, são apenas alguns exemplos de palavras e expressões empregadas no dialeto Carioca.

O dialeto Carioca é falado na cidade do Rio de Janeiro e região metropolitana, contando com 12,5 milhões de falantes. É provavelmente o jeito de falar mais reconhecível entre os brasileiros, por causa da sua grande projeção midiática no país. Está Presente na maioria das nossas novelas e filmes, é também o dialeto de português brasileiro mais reconhecível entre estrangeiros.

Esse dialeto se origina com a vinda da Família Real Portuguesa para o Brasil e a mudança da corte de Salvador para o Rio de Janeiro. Essa mudança trouxe 15 mil portugueses para o Rio de Janeiro, que outrora era habitada principalmente por escravos africanos (23 mil). Do encontro do falar europeu e do falar africano nasce o dialeto carioca.

Suas principais características são a pronuncia de /s/ chiado, uso das vogais abertas (mesmo quando há favorecimento do uso fechado) e a ditongação das vogais, exceto /a/. Assim, um carioca pronuncia a palavra "festa" como "féishta", "mesmo" como "méishmo" e "picolé" como "picoliééaa". Ainda há o uso do pronome tu conjugado em segunda pessoa "tu viu" e uso amplo do aumentativo (malzão, boladão, felizona).

## 9. PAULISTANO

*Meu, esse dialeto é daora!*

Balela, tá me tirando?, jhow, coxinha, zica, firmeza?, são apenas alguns exemplos de palavras e expressões empregadas no dialeto Paulistano.

Esse dialeto é falado na cidade de São Paulo e região metropolitana, com 25 milhões de falantes. Ele divide espaço hegemônico com o dialeto Carioca por causa dos meios de comunicação concentrados nessas duas capitais.

O dialeto Paulistano sofre forte influência das línguas de imigrantes, mas principalmente de italianos. 70% dos italianos que migraram para o Brasil se fixaram em São Paulo, e essa fala se misturou com a fala local. Outros falares de imigrantes, como sírio, libanês, japonês, e espanhol também influenciaram o falar paulistano, mas principalmente com termos léxicos e com pouco impacto nas pronuncias.

Diferente de outros dialetos, o /o/ tônico não vira nasal, sendo pronunciado como vogal aberta (tomate = tómate). A pronuncia do /r/ é acentuada, assim como nos dialetos sertanejo e sulista. Há palatalização de /d/ e /t/, com “tio” sendo pronunciado como “tchio” e “djia”.

## 10. CAIPIRA

*Este tar di dialeto é bão tamem!*

Matutano, dedin de prosa, causo, eita sô, biboca, salgar o galo, são apenas alguns exemplos de palavras e expressões empregadas no dialeto Caipira.

O dialeto Caipira é falado no interior de São Paulo, leste/sul do Mato Grosso do Sul, no sul de Minas Gerais, sul de Goiás e norte do Paraná. Ele foi descrito pela primeira vez em 1920, por Amadeu Amaral, em seu livro *O dialeto caipira*.

Reconhecido popularmente como o dialeto de pessoas “do interior”, o dialeto Caipira surge a partir do contado dos bandeirantes com os indígenas dessas regiões. Os bandeirantes eram descendentes de primeira e segunda geração de portugueses em São Paulo e a partir do meio do século 1500 penetram o interior do Brasil em busca de riquezas, como ouro e prata.

Era comum que alguns bandeirantes falassem Tupi, e usavam a língua para nomear lugares por onde passavam como Piracicaba, Taubaté, Sorocaba. Em contrapartida, os falantes nativos de Tupi não conseguiam pronunciar alguns sons da língua portuguesa, como os sons de /f/, /l/ e /r/. Assim, pronunciavam “sal” como “sar”, por exemplo.

Justamente, os traços mais marcantes do dialeto caipira são o r retroflexo (ou r puxado), a iotização do /lh/ (palha = paia, milho = mio), a redução do pronome “você” para “ocê”, a ausência de diferenciação entre singular e plural (os velhos = os vei, as mulheres = as muié) e a nasalização do /d/ em gerúndio (andando = andanu). Todas adaptações da fonética portuguesa ao Tupi.

O erre retroflexo é uma inovação desse dialeto. As pronúncias (co)R(da), (go)R(dura), (po)R(ta) são encontradas apenas nesse dialeto e em ampla distribuição na região geográfica ocupada. No entanto, existe muito preconceito contra os falantes do dialeto caipira, altamente estereotipado como “ignorante”. Lembremos que a comunidade fala assim há 400 anos e não há nada de errado com esse falar.

## 11. SULISTA

*A piazada ama esse dialeto, daí!*

Bigato, piá, capaz!, bobiça, tanso, não tem?, mandrião, disaoje, são apenas alguns exemplos de palavras e expressões empregadas no dialeto Sulista.

O dialeto Sulista é um dialeto próprio da região sul do Brasil, sendo mais falado no Paraná, Santa Catarina e norte do Rio Grande do Sul. O resto do estado de RS fala o dialeto Gaúcho, próprio da região.

O sulista “fala cantado” por causa da forte influência de portugueses da ilha de Açores na região. Essa região também recebeu grande número de migrantes italianos, cada um falando um dialeto Italiano diferente. Toda essa mistura se reflete no falar Sulista, onde o /r/ pode ser puxado ou aspirado. Podem usar o pronome “você” ou “tu” (ou ambos).

Você já deve ter pedido para um Sulista falar “LeitE quentE dá dôr nos dentEs”. Justamente, uma marca forte do dialeto é a pronuncia do /e/ ao final de palavras, que é sempre pronunciado como /e/ e nunca como /i/, como ocorre na maior parte dos outros dialetos brasileiros.

## 12. GAÚCHO

*Bah, esse sotaque é tri legal!*

Barbaridade, cacetinho, chimia, arroio, sanga, chavear a porta, são apenas alguns exemplos de palavras e expressões empregadas no dialeto Gaúcho.

O dialeto gaúcho é falado principalmente no estado do Rio Grande do Sul, com aproximadamente 12 milhões de falantes. Há diferenças entre o dialeto dentro do próprio estado por influência de diferentes colonizações, assim variações no falar entre Porto Alegre e Pelotas, por exemplo, são comuns.

Esse dialeto sofreu forte influência do Espanhol, por causa da colonização espanhola nas regiões vizinhas do Uruguai e Argentina. Por isso, a fonologia do dialeto Gaúcho é bastante próxima do Espanhol rio-platense, como o ritmo silábico de fala. Também é influenciado pela língua Guarani, que enriquece principalmente o léxico.

Características marcantes desse dialeto são o uso do pronome “tu” com o verbo em terceira pessoa, como em “Tu vai”, embora não seja raro ouvir a conjugação “Tu vais” e a extensão de vogais no meio das palavras. Assim como no dialeto sulista, as palavras terminadas em /e/ são pronunciadas como /e/.

**References**


Brandão SF. (2014) Nas trilhas do -R retroflexo. *Signum Estud da Ling*. 10:265. doi:10.5433/2237-4876.2007v10n2p265.

Bueno ES da S. (2011). O falar do homem pantaneiro: um olhar sociolinguístico. *Ave Palavra*. 12:1–47.

Collischonn G, Monaretto V de O. (2012). Banco de dados VARSUL: A relevância de suas características e a abrangência de seus resultados. *ALFA Rev Linguística*. 56:835–853. doi:10.1017/CBO9781107415324.004.

de Aragão M do SS. (1999). A variação fonético-lexical em Atlas Lingüístico do Nordeste. Rev do GELNE. 1:14–20.

Görski EM. (2012). Fenômenos variáveis na Região Sul do Brasil: aspectos de comportamento sociolinguístico diferenciado entre as três capitais. Estud Linguísticos, São Paulo. 41:806–817.

Martins EF. (2006). Atlas lingüístico do Estado de Minas Gerais: o princípio da uniformidade da mudança lingüística nas características fonéticas do português mineiro. *Rev Virtual Estud da Ling*. 4:1–13.

Mota JA. (2008). Como fala o nordestino: a variação fônica nos dados do Projeto Atlas Lingüístico do Brasil. In: Lima-Hernandes MC, Marçalo MJ, Micheletti G, Martin VL de R, editors. *A língua portuguesa no mundo*. v. 1. São Paulo: FFLCH-USP.

Nassif, Luiz. (2012). A padronização do sotaque no telejornalismo. *Jornal GGN.* Disponível em <https://jornalggn.com.br/cultura/televisao-cultura/a-padronizacao-do-sotaque-no-telejornalismo/>.

Pereira DT. (2014) O uso do termo e do dialeto caipira nos jornais do século XIX (1838-1884). *Rev Ars Hist*. 169–179.

Romano VP. (2018) Áreas lexicais no Centro-Sul do Brasil sob uma perspectiva geolinguística. *Rev Estud Da Ling*. 26:103–145. doi:10.17851/2237-2083.26.1.103-145

SUPPLEMENTARY MATERIAL 3

**Fuxiquera Quiz**

Existem vários portugueses dentro do português brasileiro.
Qual deles você fala?

**ATENÇÃO!** Antes de começar o teste saiba que não existe resposta certa ou errada. Esqueça as regras da norma culta do português. Queremos saber como você **FALA**, como se comunica verbalmente com as outras pessoas.

Não existe "falar errado" pois cada região do Brasil possui características próprias, com gírias e pronuncias regionais. A língua falada é viva e dinâmica, por isso está em constante evolução. O certo é biscoito ou bolacha? Tangerina ou fuxiquera? Depende de onde você nasceu.

[FAÇA O TESTE]

1. **Além de menino, qual outra forma de chamar uma criança entre 5 e 10 anos, do sexo masculino?**
    a) Garoto
    b) Moleque
    c) Guri
    d) Piá
    e) Rapazinho
    f) Homenzinho
    g) Pivete
    h) Pirralho
    i) Bambino
    j) Bacuri
    k) Curumim
    l) Badeco
    m) Pixote
    n) Criança
    o) Pimpolho
    p) Pequeno
    q) Nenhuma
    r) [outra]

2. **Como se chama um rio pequeno, de uns dois metros de largura?**
    a. Igarapé
    b. Garapé
    c. Corixo
    d. Corgo / Corguinho
    e. Córrego
    f. Riacho / Riachinho / Riachozinho
    g. Riozinho
    h. Ribeirão
    i. Sanga / Sanguinha
    j. Rego/Reguinho
    k. Arroio
    l. Açude
    m. Barreiro
    n. Brejo
    o. Canal
    p. Bisca
    q. Cachoeira / Cachoeirinha
    r. Lagoa
    s. Lago
    t. Vala/Valo/Valão
    u. Valeta/Valetinha
    v. Grota / Grotão
    w. Olheirinho
    x. Correnteza / Corrente
    y. Fonte / Fontezinha
    z. Baixa
    aa. Baixio
    bb. Estreito
    cc. Enseada
    dd. Cóligo
    ee. Vazante
    ff. Lajeado
    gg. Vertente
    hh. [outra]

3. **Independente da variedade, qual o nome das frutas das fotos?**

*Imagens licenciadas por Creative Commons CC0*

    a. Mimosa

b. Tanja
c. Mexerica / Mixirica
d. Poncã / Ponkã
e. Pocã / Pokan
f. Tangerina
g. Bergamota / Vergamota
h. Laranja-cravo
i. Laranja-crava
j. Mangota (Mangote, Mongote, Margota)
k. Morgota
l. Fuxiqueira
m. Mandarina
n. Carioquinha
o. Maricota/Maricote
p. Muricota/Moricota (Muricote, Morocota, Morocote)
q. Murcote/Morcote (Marcota, Morcota, Mucote)
r. Mimo-do-céu
s. Mixiriquêra
t. Azedinha
u. Mormota
v. Irredera /Enredera
w. Clementina
x. [outra]

4. **Qual o nome desse suco de frutas congelado em saquinhos?**

*Foto por Alexandra de Abreu em Flickr*

a. Brasinha
b. Tubiba
c. Dida
d. Vip
e. Sacolé
f. Laranjinha
g. Dindin
h. Apolo
i. Dudu
j. Tabu
k. Juju
l. Refresco
m. Flau / Frau
n. Chupe-chupe / Chup-chup
o. Gelinho
p. Geladinho

q. Geladinha
r. Picolé
s. Chopp
t. Chope
u. Suquinho
v. Bolão
w. [outra]

5. **Qual é outro nome para "Bolinha de gude"?**

*Imagem licenciada por Creative Commons CC0*

a. Peteca
b. Fona / Bola de fona
c. Butila
d. Tila-tila
e. Gurca / Gurquinha
f. China
g. Bolita
h. Bulita
i. Bolica
j. Bulica / Búlica
k. Burica / Búrica
l. Burca / Burquinha
m. Quilica
n. Clica
o. Peca
p. Biloca
q. Bilisco
r. Biroca
s. Birosca
t. Bilosca
u. Birola
v. Biroquê
w. Gude
x. Bila
y. Ximbra / Chimbra
z. Fubeca
aa. Boleba/ Baleba
bb. Bolinha de crica / Bolinha de crique
cc. Pinica
dd. Pilica/Tilica
ee. Aço
ff. Marraio

gg. Bolinha de vidro
hh. Bolinha
ii. Olho de gato
jj. Peloco
kk. Nenhuma
ll. [outra]

6. **Qual interjeição você mais usa?**
   a. Ôxe
   b. Oxente
   c. Uai
   d. Báh
   e. Tchê
   f. Égua
   g. Arre égua
   h. Hehein
   i. Afe / Aff
   j. Vei / Véi / Veio
   k. Vichi / Vixe
   l. Ixe
   m. Eita / Eta
   n. Meu
   o. Caraca
   p. Putz
   q. Ué
   r. Daí
   s. Mano
   t. Viu
   u. Valha
   v. Beí / Bei
   w. Nenhuma
   x. [outra]

7. **Como se pede esse pão na padaria?**

*Imagem produzida pela Agência Brasil, a agência pública de notícias do Brasil.*

   a. Carioquinha
   b. Média
   c. Pão francês
   d. Pão de sal
   e. Pão de trigo

f. Pão aguado
g. Pão d'água
h. Pão jacó
i. Pão massa grossa
j. Pão careca
k. Filão / Filãozinho
l. Cacetinho
m. Pãozinho
n. Pão de padeiro
o. Pão brotinho
p. [outra]

8. **A pessoa que não gosta de gastar seu dinheiro e, às vezes, até passa dificuldades para não gastar?**
   a. Roda presa
   b. Pirangueiro
   c. Canhenga
   d. Canguinha
   e. Amarrado
   f. Amarradinho
   g. Unha de fome
   h. Somítico
   i. Usurário
   j. Morta-fome
   k. Morto-de-fome
   l. Morto-a-fome
   m. Pechincheiro
   n. Agarrado
   o. Pica-fumo
   p. Arrochado
   q. Chula
   r. Fona
   s. Fominha
   t. Casquinha
   u. Munheca
   v. Muxiba
   w. Murrinha
   x. Miserável
   y. Tacanha
   z. Papagaio no arame
   aa. aa. Enforcado
   bb. bb. Muquirana
   cc. cc. Miserento dd.
   dd. Mão-dura ee.
   ee. Mão-apertada ff.
   ff. Mão-fechada
   gg. Mão de vaca
   hh. Mão de munheca
   ii. Mão de égua
   jj. Mão mirrada
   kk. Mão de figa

ll. Pão dura
mm. Mesquinha
nn. Avarenta
oo. Sovina
pp. Cainha
qq. Ridica
rr. [outra]

9. **Quando se quer saber se uma amiga está indo embora, como é que se pergunta?**
   a. Amiga, tu vais embora?
   b. Amiga, tu vai embora?
   c. Amiga, você vai embora?
   d. Amiga, ocê vai embora?
   e. Amiga, cê vai embora?

10. **Quando Maria tem um livro:**
   a. O livro de Maria
   b. O livro da Maria
   c. Ambos

11. **Quando Carlos tem um livro:**
   a. O livro do Carlos
   b. O livro de Carlos
   c. Ambos

12. **Qual o gênero de alface?**
   a. O alface
   b. A alface

13. **Um homem que rouba, você diz que é ladrão. E quando é uma mulher?**

   a. Ladrona
   b. Ladra
   c. Ladroa
   d. Gatuna
   e. [outra]

14. **Como você chama esse alimento?**

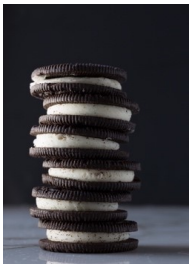

*Imagem por Denise Johnson disponível em Unsplash*

   a. Biscoito
   b. Bolacha

15. **Como você conta para alguém sobre uma praça que existia no passado?**

a. Tinha uma praça
b. Havia uma praça
c. Era uma praça
d. Via uma praça

16. **Como você fala o número 3? [OUVIR ÁUDIO]**
    a. tre[js]
    b. tre[S]
    c. tre[ʃ]
    d. tre[jʃ]
    e. trei

17. **Fale em voz alta a palavra "tomate". A primeira silaba "to" se parece mais com quando você fala a primeira sílaba de:**

a. 
*Imagem por Carmelita Rodrigues disponível em Unsplash*

**ou**

b. 

18. **Fale em voz alta o nome do bicho da foto. Ele rima com qual opção abaixo?**

*Imagem disponível em CSIRO.*

a. Touro
b. Zorro

19. **Qual áudio representa melhor a sua pronuncia de "Pertinho"? [OUVIR ÁUDIO]**
    a. Pe(h)(tch)inho
    b. Pe(x)(tch)inho
    c. Pe(ɾ)(tch)inho
    d. Pe(ɾ)(tch)inho
    e. Pe(ɾ)(tch)im
    f. Pe(h)(tch)im
    g. Pe(ɾ)(tch)im

h. Pe(h)(Ti)nho
i. Pe(r)(Ti)nho
j. Pe(x)(TI)inho

20. **Como você fala a bebida da imagem? [OUVIR ÁUDIO]**

*Imagem disponível em Peakpx.*

a. Leiti
b. LeitE
c. Leitchi
d. Leitu

21. **Como você fala a cor da imagem? Escolha o áudio que melhor se aproxima. [OUVIR ÁUDIO]**

a. Ve(ɽ)de
b. Ve(h)de
c. Ve(x)de
d. Ve(ɾ)de
e. Vede
f. Ve(j)de
g. Ve(w)de

22. **Como você fala "viajar"? [OUVIR ÁUDIO]**
a. Viaja(x)
b. Viaja(h)
c. Viaja(ɽ)
d. Viaja(ɾ)
e. Viajá

23. **Como você escreveria a forma como você fala "mesmo"?**
a. mehmo
b. mermo
c. memo
d. mes
e. mezmo
f. mexmo

24. **Quando você fala "televisão", o "te" ou "le" rima com pé?**

a. Sim
b. Não

25. **Como você chama essa raiz?**

*Foto por Codevasf em Flickr*

a. Pão de pobre
b. Castelinha
c. Mandioca
d. Aipim
e. Uaipi
f. Macaxeira
g. Maniva
h. Maniveira
i. Pau de farinha
j. Macamba
k. Pão da América
l. Xagala
m. Candinga
n. [outra]

**Perguntas extralinguísticas (ao final do teste)**

1. Gênero
2. Idade
3. Grau de escolaridade
4. Onde nasceu
5. Onde já morou
6. **6.** De onde é sua mãe?
7. De onde é seu pai?
8. De onde é seu cônjuge?

# CAPÍTULO 3

## *High mountains, wide rivers: mechanisms that shaped gradients of language diversity in the Neotropics*

Christielly Borges[1*], Marco Tulio Pacheco Coelho[1,2], Thiago Costa Chacon[3], Michael Gavin[4], Thiago Fernando Rangel[1]

[1] Departamento de Ecologia, Universidade Federal de Goiás, Goiânia, Goiás, Brasil

[2] Swiss Federal Institute for Forest, Snow and Landscape Research, WSL, Switzerland

[3] Departamento de Linguística, Universidade de Brasília, Brasília, DF, Brasil

[4] Department of Human Dimensions of Natural Resources, Colorado State University, Fort Collins, CO, USA

***** Corresponding author: christielly@gmail.com

*Status*: Em preparação para submissão à revista **Global Ecology and Biogeography**

ABSTRACT

**Aim:** Global patterns of human diversity are spatially and demographically complex and their underlying causes are still discussed. Here, though a spatially explicit mechanistic model we tested how environmental, ecological and demographic mechanisms imposed by altitude, water availability, rivers and group size shape pre-Columbian language and phylogenetic diversity.

**Location:** Neotropical realm

**Time period:** Pre-Columbian era

**Major taxa studied:** Linguistic diversity of *Homo sapiens*

**Methods:** We designed a spatially explicit mechanistic model to simulate the spread and evolution of languages under the effect of environment, demographic and ecological factors. Our model's premises are that human groups spread to occupy empty spaces, environmental carrying capacity determines group's population density, and groups have a maximum population size. The geographical domain of our model is represented by a gridded map in which hexagon cell sizes are affected by topographic and hydrographic complexity have varying carrying capacity according to the environment.

**Results:** On the one hand, our model is very accurate on the total number of languages produced in the continent and generates an average of 1016 simulated languages vs. 986observed pre-colonial languages. On the other hand, our model has an intermediate performance for spatial gradients of language diversity and explains 8% of the variation in the observed spatial patterns of language richness and 22% of the observed spatial pattern of phylogenetic language diversity.

**Main conclusions:** Our work contributes to the growing literature on processes driving language diversity patterns by offering an 'in-silico' experiment on how rivers and mountains shape different dimensions of language diversity.

**Keywords**: language diversity, macroecology, mechanistic model, simulation modelling

## INTRODUCTION

Different aspects of human culture change through descent and are rich in diversity, thus can be considered from an evolutionary perspective and as a measure of variation in human social behavior (Brewer et al., 2017). Global patterns of human diversity are also spatially and demographically complex, which allows for environmental, ecological and demographic factors to be explored as possible drivers (Mirazón Lahr, 2016). Human diversity has indeed gained attention in transdisciplinary research in recent decades (Maffi, 2005), mainly in macroecological studies focusing on human evolution through variables such as religion (Watts, Sheehan, Atkinson, Bulbulia, & Gray, 2016), subsistence strategies (Gavin et al., 2018; Vilela et al., 2020), residential mobility, resource ownership (Kavanagh et al., 2018) and most notably languages (Gavin & Sibanda, 2012; Gavin & Stepp, 2014; Hua, Greenhill, Cardillo, Schneemann, & Bromham, 2019; Levinson & Gray, 2012; Matthews, Passmore, Richard, Gray, & Atkinson, 2016; Michalopoulos, 2008; Pacheco Coelho et al., 2019).

There are approximately 7,000 extant languages spoken in the world today (Simons & Fennig, 2018), from a total of 30,000, a cautious estimate, to 500.000 languages, a bolder estimate, that may have existed since the Cognitive Revolution in *Homo sapiens* (Crystal, 2010). However, languages are unevenly distributed worldwide, with a higher density of linguistically distinct groups occurring towards the equator. Even amongst tropical areas we encounter unevenness, with the continent of Africa holding five times more languages than South America (Axelsen & Manrubia, 2014).

Many studies have tackled the question of what influences the geographical patterns of linguistic diversity (Currie & Mace, 2009; Fincher & Thornhill, 2008; Gavin et al., 2017; Gavin & Sibanda, 2012; Levinson & Gray, 2012; Michalopoulos, 2008; Sutherland, 2003), enabling the main factors to be grouped into three categories: environmental, topographic and sociocultural (Gavin et al., 2013). Nonetheless, there are still challenges in studies of cultural

evolution (Brewer et al., 2017) one in special being the mechanism underlying patterns of linguistic diversity and cultural transmission models (Gavin et al., 2013).

The first study to apply a mechanistic model to uncover the mechanisms behind language diversity was able to explain 56% of the spatial variation in language diversity in Australia and concluded that the main processes explaining the pattern are environmental carrying capacity and group size (Gavin et al., 2017). Another study found that a simple mechanistic model to simulate the effects of different carrying capacity and group size limits explained 11% of the language diversity in North America and was also one of the strongest predictors assessed in a path analysis (Pacheco Coelho et al., 2019). Nonetheless, due to the environmental, social and historical differences, mechanisms can differ amongst regions.

The Neotropics is an ideal region to apply mechanistic models to language diversity questions. The continent has a higher diversity in Central America, and not in the Amazon region as was expected according to the cooccurrence with biodiversity hypothesis (Manne, 2003). It is the region with the highest phylogenetic language diversity (Nettle, 1998), has the highest number of isolates (Blench, 2008; Crystal, 2010) and was the last continent to be populated by *Homo sapiens* (Davidson, 2013).

Higher language diversity in the Americas occurs on the west coast, following mountain chains from Sierra Nevada, in North America, to Sierra Madre, in Central America, to the Andes, in South America. Given this uneven distribution, we speculate topography to be an important mechanism in the linguistic pattern observed in the region. Mountainous areas are fragmented and topographically diverse, promoting environments with high biological diversity (resources) (Spehn & Körner, 2005) and have already been suggested as mechanisms responsible for generating and supporting language diversity (Michalopoulos, 2008). Furthermore, as an environmental barrier with limited topographic area (mountain tops and valleys), mountains contributed to an initial rapid spread and subsequent mobility restrainer

which formed and maintained small demographic ethnic and linguistically distinct groups (Michalopoulos, 2008).

Regions with high river and lake densities have also been proposed as cradles of language diversity (Axelsen & Manrubia, 2014), specially for serving as social centers, allowing contact and stablishing trade systems amongst populations (Hornborg, 2005). This systems allows the emergence of new languages within few generations, contributing to language diversity (Axelsen & Manrubia, 2014). Thus, rivers and lakes do not act as geographic barriers isolating populations, but as heterogenous environments that provide a higher number of resources and raises the carrying capacity in the region. This so happens because beyond the available aquatic resources, floodplain areas have fertile soil, which also permits high productivity and subsistence farming practices (Hornborg, 2005). Finally, the presence of rivers and lakes can be an important mechanism to explain the high language diversity given specially the occupation and immigration of different language families in different watersheds (Cabral, Martins, Corrêa da Silva, & Oliveira, 2014).

Climate has been central in many studies assessing drivers of language of diversity, where it is usually used as a productivity proxy specially because stable climatic conditions are more suitable for human settlement, provision of resources, subsistence farming and consequently, higher population growth and density (Derungs, Köhl, Weibel, & Bickel, 2018). Precipitation, specifically, has been found as a strong variable in many language diversity models, mechanistic and correlational, in different continents (Axelsen & Manrubia, 2014; Gavin et al., 2017; Moore et al., 2002; Nettle, 1996; Pacheco Coelho et al., 2019). Precipitation supplies water necessary for terrestrial organisms, thus, much like rivers and mountains, we expect precipitation also increases the carrying capacity in a given region.

Finally, group size is another important mechanism in language diversity patterns (Gavin et al., 2017). If population size can actually influence the rates of language change has

been a controversial topic, with evidence against and for it (Bromham, Hua, Fitzpatrick, & Greenhill, 2015; Greenhill, Hua, Welsh, Schneemann, & Bromham, 2018). Nonetheless, it has been argued that limits on group size may facilitate the division of social groups, which compensates the cost of maintaining social ties, such as recognizing relatives or unreliable individuals (Dunbar, 2008). Further, new languages emerging from small groups can suffer rapid bursts of change, diverging from the ancestral language (Atkinson, Meade, Venditti, Greenhill, & Pagel, 2008).

This paper aims to explain the geographic language diversity patterns observed in the Neotropics. By Neotropics we specifically refer to all the countries south of the United States, which is also the geopolitical region of Latin America. We expect altitude, rainfall, river width and group size to be important mechanisms in generating the observed pre-Columbian language diversity pattern found in the region. We built a simple mechanistic simulation model, incorporating all four mechanisms mentioned above, to recreate this observed language pattern.

## METHODS

### General Model

We created a spatially explicit mechanistic model to reconstruct the quantity and spatial pattern of the observed pre-Columbian language diversity in the Neotropical realm. This model follows the approach stablished by Gavin *et al.* (2017) and abides by the same premises: 1) human groups occupy empty spaces, 2) environmental carrying capacity determines group's population density, and 3) groups have a maximum population size. Our model is stochastic, based on hypothetical languages and the occupation of cells has no intended relationship to historic time.

We used a hexagonal hybrid grid where cell sizes are inverse to topographic and hydrographic complexity (Rangel et al., 2018), meaning smaller cells in areas of high elevation

and river width, and larger cells in lowlands (Fig. S1). We calculated for each cell a carrying capacity, meaning how many individuals can occupy it. Simulation begins with 10 individuals (representing a linguistic group) occupying a random cell in the map. As population grows and reaches the cell's maximum carrying capacity, the group can colonize neighboring cells until it reaches its maximum population size (sampled from the empirical distribution of hunter-gatherer group sizes). When the maximum population size is reached, another cell is colonized by 10 individuals (a new language group) and the cycle is repeated (Gavin et al., 2017). The simulation ends when all cells are occupied (see Movie S1). Our mechanistic model was written in the Python programing language (Van Rossum & Drake, 2009), and is publicly available on GitHub (https://github.com/chrisborges/Neotropical_Biocultural_Diversity).

**Hexagonal grid and environmental data**

We created a geodesic dome with hexagonal cells, where each cell is neighbored by six cells. Cells were classified within one of 50 linear categories created by a sum of the hydrographic (river width) and topographic (altitude's standard deviation) complexity found in each cell. Cells with low summed-rank were joined to its neighbors. Thus, cells with high hydrographic and topographic complexity were rescaled as small and cells with low complexity (no rivers or in lowlands) were joined together and rescaled as big. Homogenous cells were joined in clusters of seven to three neighboring hexic cells. Complex small cells that were isolated (not adjacent to any other complex cell) were also joined to its homogeneous neighbors. Our final American hybrid grid resulted in 19,244 different sized hexic cells, where smaller cells represent river and altitude complexity and bigger cells are clusters of non-complex cells.

We used altitude's standard deviation as a proxy for topographic complexity. The altitude data was obtained from the WorldClim database (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005), which offers an altitude grid layer (elevation above sea level in meters) compiled

from the Shuttle Radar Topography Mission (SRTM) database. We used river width as a proxy for hydrographic complexity. River data was obtained from the global river network HydroSHEDS, which is based on Shuttle Elevation Derivatives at multiple Scales (Lehner, Verdin, & Jarvis, 2008).

We obtained annual precipitation data for the pre-industrial period (1760) from the ecoClimate database at a resolution of 0.5˚ latitude/longitude (Lima-Ribeiro et al., 2015). As there are no appropriate method for an Ocean Atmospheric General Circulation Model (OAGCM) selection, we averaged the precipitation layers from five OAGCMs: CCSM4, CNRM, FGOALS, MIROC-ESM and MRI, to better represent the past climate (Yoo & Cho, 2018). Rainfall data was later extracted to the hexic grid (Fig. S2).

**Group size**

We used Binford's hunter-gatherer (HG) data to determine randomly the maximum population size of each new group (language) in the model (Binford, 2001; Marwick, 2017). Following (Gavin et al., 2017), we excluded HG data from the artic and subarctic, as these climates do not occur in the Neotropic, and from Mexico and South America, to avoid redundancy in our results. From a total of 339 observations, 233 observations remained after the exclusion (see Fig. S3 for the empirical distribution).

**Carrying capacity**

The study by Gavin *et al.* (2017) evaluated three functions (exponential, logistic and power) to determine the best relation between increase in carrying capacity (K) and precipitation. They concluded that the power function has better predictive power, thus we calculated K for each cell using only the power function:

$K = \alpha P^{\beta}$

where K is the number of individuals per $km^2$, P is mean annual precipitation in millimeters (mm), and α and β are unknown parameters to be estimated.

We estimated the carrying capacity's unknown parameters through a Markov Chain Monte Carlo (MCMC) Gibbs sampler (Gavin et al., 2017; Gelman et al., 2013). We ran replicates of the simulation to explore 3380 different parameter combinations for the above power function. We later compared these parameter combinations with the model's goodness of fit index, which converged in a log-linear relationship (Fig. S4).

**Language data**

We are mainly interested in linguistic patterns that arose from groups tied to a landscape living. At present, most Latin American countries consider as their official languages at least one Indo-European language such as Portuguese, Spanish, Dutch, English, and French (Simons & Fennig, 2018), reflecting a history of colonialism. Environmental factors may not be the most qualified mechanisms to explain the distribution of groups speaking colonial languages (Manne, 2003) since the process behind these patterns are due mainly to historically political and societal pressure, economic survival, prestige (Mufwene, 2002) or cultural imperialism (Silva, 2004). Therefore, we made a historical cut of languages occurring in the Neotropics approximately before the European colonization in the 1500s.

Observed language diversity data were obtained from the Glottolog 3.2 database (Hammarström, Forkel, Haspelmath, & Bank, 2020), which provides point occurrence for languages worldwide. We obtained data for North and South America and later refined it to our Neotropic region. We excluded languages from the Indo-European family and sign languages. Our final database had 986 languages (Fig. 2a) distributed in 157 language families in the Neotropics (Fig. 3), of which 92 are family isolates. Additionally, 15 languages had an

unclassified family, those were removed from the language family maps but kept in the observed language diversity.

We used Voronoi diagrams, through the Thiessen Polygon method, to create geographic language ranges from the point occurrence data. Voronoi diagrams partition the space by applying the rule of nearest neighbor, where each point is associated with the nearest region, and can make natural processes more explicit (Aurenhammer, 1991). Thiessen Polygons is a method created to apply Voronoi diagrams to geographic studies and serve as models in spatial process and point pattern analysis (Aurenhammer, 1991). This method is ideal for language data because pre-colonial language ranges do not overlap in space, thus Thiessen polygons produce accurately the language's distributions based on occurrence points.

**Model test and validation**

Simulation models were tested on their ability to replicate the observed pattern of language richness and the total number of languages in the Americas (Gavin et al., 2017). Following (Gavin et al., 2017), we created an ad hoc goodness of fit index (f) that evaluated the combined ability of the model to predict the number of languages and the observed spatial pattern for the region.

To assess the model's capacity in replicating the number of observed languages in the region, we calculated a standardized measure of similarity (s) as

$$s = 1 - \frac{|O-P|}{P},$$

where O is the observed number of languages and P is the number predicted by the model.

We created an equal-area grid of 300 x 300 km cell, totalizing 253 cells, to assess the replicated spatial pattern. We further created maps of language richness by counting the number of languages (polygons) that occurred within each cell. We used the coefficient of determination ($r^2$) of linear regressions to measure the fit between the observed (response

variable) and the simulated (explanatory variable) language richness (Gavin et al., 2017). The residuals map (observed - predicted; Fig. 2e) indicates regions where the model underestimated or overestimated the observed number of languages, given by positive and negative values respectively.

The goodness of fit index was then calculated as

$f = \frac{r^2+s}{2}$,

where the maximum value of f is 1 if the model predicts exactly 986 languages and the language richness matches the observed language richness precisely.

## RESULTS

Our model produced an average of 1016±84.6 simulated languages for the Neotropics, as replicated by the best 200 models (Fig. 1a). This is a difference of only 30 simulated languages from the observed 986 number of languages. The best 200 models further had an average coefficient of determination of 8.4%±0.02 (Fig. 1b), demonstrating a low fit between the observed and the simulated language richness.

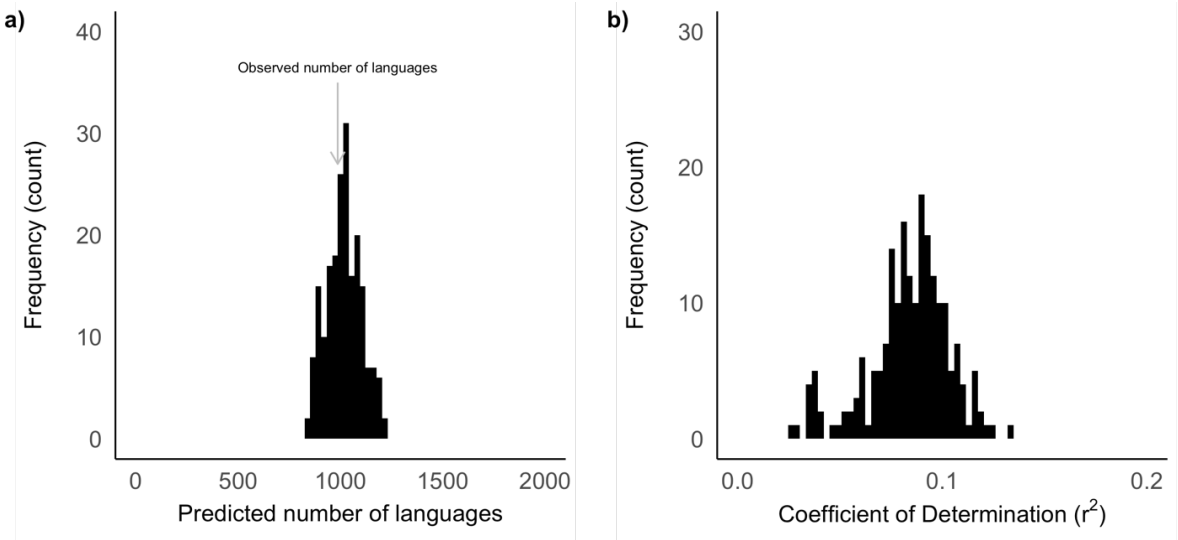


**Figure 1.** (a) Distribution of the total number of languages predicted by the best 200 models. The observed number of languages is 986 and the average number of languages predicted was

1016. (b) Distribution of the coefficient of determination ($r^2$) between the spatial pattern for the observed language richness and the predicted language richness, by the best 200 models.

Nonetheless, the simulated languages maps (Fig. 2b; representing one model replicate) are visually similar to the observed map (Fig. 2a). The language ranges are indeed smaller in the Amazon and Andean regions and larger towards the east coast and south of the continent in both maps, meaning the model was capable of replicating the observed pattern in South America. For Central America, however, the predicted language ranges are larger than expected when compared to the observed map.

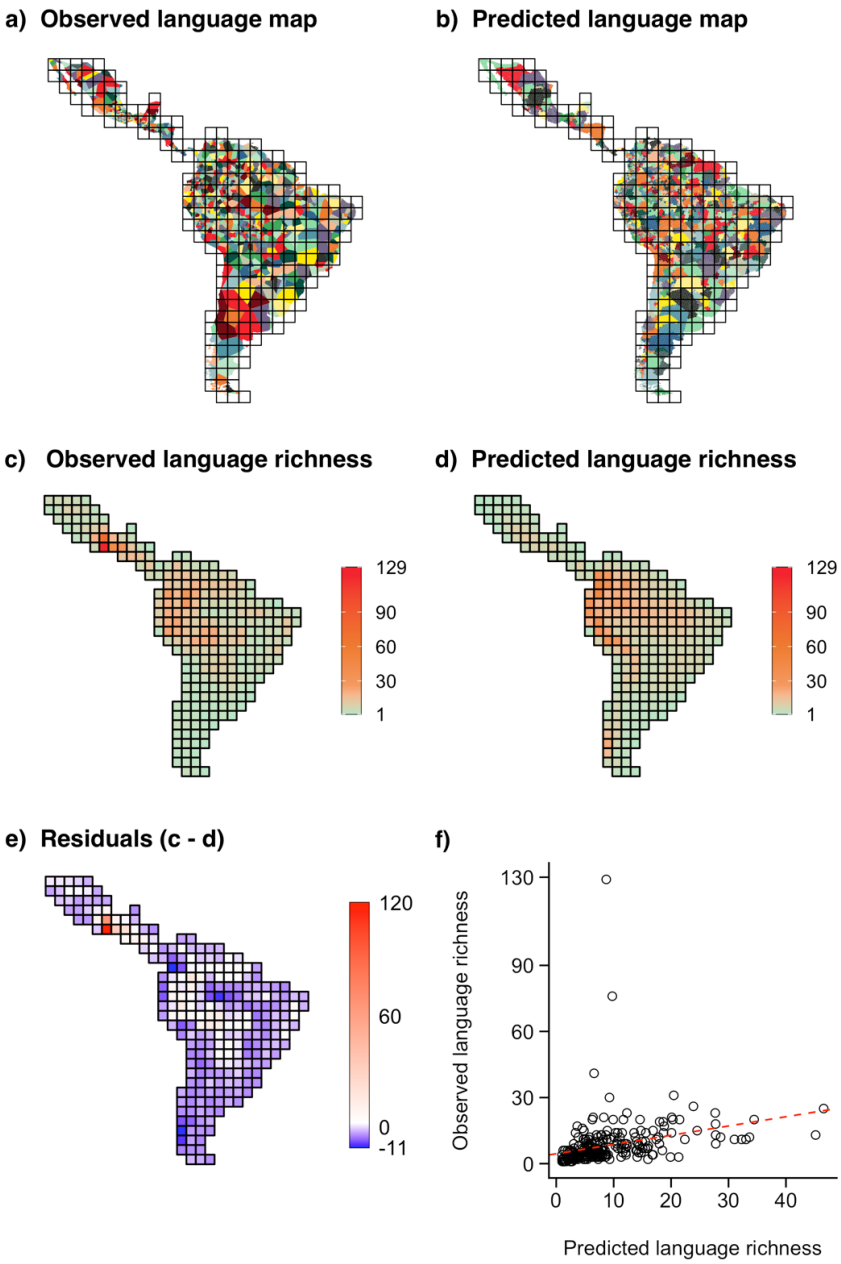


**Figure 2.** Observed and predicted language diversity patterns for the Neotropics. (a) Observed language map for a total of 985 languages. (b) Predicted language map from one model replicate. (c) The observed and (d) predicted number of languages (richness) occurring within a grid of 300 x 300 km$^2$ cells. (e) Represents fit, shows the residuals (observed – predicted) languages in each cell. (f) The model predicts 8.6% of the variation in observed language richness patterns.

The predicted and observed language richness are also very visually similar, as the model predicted correctly the regions of high and low linguistic diversity (Fig. 2cd). Nonetheless, the models predicted a higher number of languages for the Andean region of South

America (~ 50 languages), whereas the higher observed cell is located in today Guatemala (129 languages). Further, predicted richness for the southern Chilean region is higher (~30 languages) than expected by the observed richness (~1 language).

Our residuals map shows our model greatly underestimated the language richness in today Guatemala (Fig. 2e; in red) and overestimated the number of languages in most of the continent (in blue). Overall, our predicted and observed language richness maps are 30% correlated and our model's predicted language richness explains 8.7% of the variation in the observed richness (Fig. 2f). Further, we used a traditional correlation approach to assess the relationship between past mean precipitation (mm) and the observed language richness in the 300 $km^2$ grid cells, which showed a similar predictive power ($r^2 = 0.07$).

Finally, we noticed our predicted language range and language richness maps were particularly similar to the language family maps (Fig. 3), as the language family ranges are larger in Mexico and Central America and the higher diversity is in the Andes. We ran a linear regression and found the predicted language richness explains 22% of the observed language family richness.

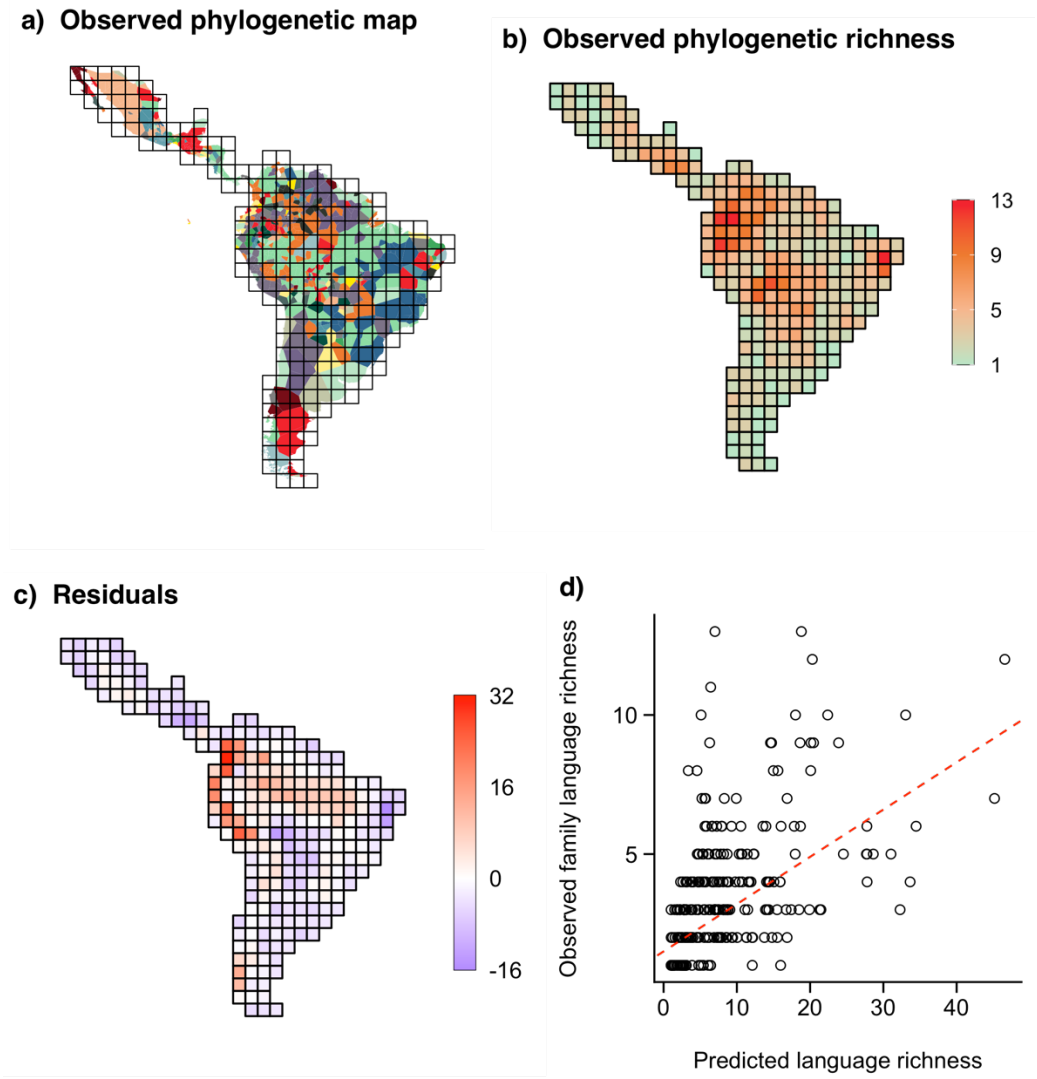


**Figure 3.** Observed phylogenetic language diversity patterns for the Neotropics. (a) Observed language family map for 157 families, of which 92 are language isolates. (b) Observed language family richness occurring within a grid of 300 x 300 km$^2$ cells, maximum is 13 family languages in the same cell. (c) Represents fit, shows the residuals (observed – predicted (Fig. 2d)) languages in each cell. (d) The model predicts 22% of the variation in observed phylogenetic richness patterns.

## DISCUSSION

Our mechanistic model incorporates environmental, topographic and sociocultural as drivers in shaping the language diversity patterns in the Americas. Our model successfully predicted the number of languages and the spatial pattern of the linguistic diversity for the Neotropics. Further, our model had a greater predictive power (8% vs 7%) than the traditional correlation

approach, upholding that underlying mechanisms can be incorporated successfully in simulation modelling for ecological and evolutionary biodiversity processes (Gavin et al., 2017; Pacheco Coelho et al., 2019; Rangel et al., 2018).

The high richness predicted in the Andean region and south of Chile can be due to these regions being compromised of mountains chains, meaning the cells were also smaller. By our models' parameters, these factors resulted in a high carrying capacity for these cells. Similarly, our model's failure in predicting an accurate number of languages for Central America is probably due to the low mean annual precipitation found for these regions (Fig. S2). This means precipitation is not an important mechanism in the linguistic pattern found for Central America. We suspect, because this continent is a strip bounded by the ocean on both sides (almost an archipelago if you consider the Panama Canal as a barrier), that oceanic resources might be the main mechanisms driving language diversity in this continent.

Nonetheless, the greater observed language richness for Central America can also be tied to the farming propensity of traditional human societies, shaped specially by richer environments that could sustain a greater diversity of domesticated plant and animal species (Vilela et al., 2020). In contrast, populations in South America were more reliant on hunting, gathering or fishing. This could explain the higher diversity of the continent, as the geographic variation in reliance on agriculture can be predicted by language family (Vilela et al., 2020).

A recent study evaluated the drivers of global language diversity and found that neither river density nor landscape roughness were significant in explaining language diversity variation (Hua et al., 2019). They argue the positive results found in other studies (Axelsen & Manrubia, 2014) are products of spatial autocorrelation and phylogenetic non-independence. This difference in results found in the literature illustrates further the advantages of using simulation models to uncover drivers of richness patterns, as our results are not the artifact of statistical misuse or spurious correlations. Thus, our work further contributes to the growing

literature on processes responsible for global language diversity since in an unprecedented way, we included rivers and mountains as underlying mechanisms directly in the simulation model.

## REFERENCES


Atkinson, Q. D., Meade, A., Venditti, C., Greenhill, S. J., & Pagel, M. (2008). Languages Evolve in Punctuational Bursts. *Science*, 319(5863), 588–588. Retrieved from https://doi.org/10.1126/science.1149683

Aurenhammer, F. (1991). Voronoi diagrams---a survey of a fundamental geometric data structure. *ACM Computing Surveys*, 23(3), 345–405. Retrieved from https://doi.org/10.1145/116873.116880

Axelsen, J. B., & Manrubia, S. (2014). River density and landscape roughness are universal determinants of linguistic diversity. *Proceedings of the Royal Society B: Biological Sciences*, 281(1784), 20133029–20133029. Retrieved from https://doi.org/10.1098/rspb.2013.3029

Binford, L. R. (2001). *Constructing Frames of Reference: An Analytical Method for Archaeological Theory Building Using Ethnographic and Environmental Data Sets*. Berkeley: University of California Press.

Blench, R. (2008). Accounting for the diversity of Amerindian languages: modelling the settlement of the New World. In *Archaeology Research Seminar* (pp. 1–17). Canberra.

Brewer, J., Gelfand, M., Jackson, J. C., MacDonald, I. F., Peregrine, P. N., Richerson, P. J., … Wilson, D. S. (2017). Grand challenges for the study of cultural evolution. *Nature Ecology & Evolution*, 1(3), 0070. Retrieved from https://doi.org/10.1038/s41559-017-0070

Bromham, L., Hua, X., Fitzpatrick, T. G., & Greenhill, S. J. (2015). Rate of language evolution is affected by population size. *Proceedings of the National Academy of Sciences*, 112(7), 2097–2102. Retrieved from https://doi.org/10.1073/pnas.1419704112

Cabral, A. S. A. C., Martins, A. M. S., Corrêa da Silva, B. C., & Oliveira, S. C. S. de. (2014). A linguística histórica das línguas indígenas do Brasil, por Aryon Dall'igna Rodrigues: perspectivas, modelos teóricos e achados. *DELTA: Documentação e Estudos Em Linguística Teórica e Aplicada*, 30, 513–542. Retrieved from https://doi.org/http://dx.doi.org/10.1590/0102-445090644999061809

Crystal, D. (2010). *The Cambridge Encyclopedia of Language* (3rd ed.). Cambridge: Cambridge University Press.

Currie, T. E., & Mace, R. (2009). Political complexity predicts the spread of ethnolinguistic groups. *Proceedings of the National Academy of Sciences*, 106(18), 7339–7344. Retrieved from https://doi.org/10.1073/pnas.0804698106

Davidson, I. (2013). Peopling the last new worlds: The first colonisation of Sahul and the Americas. *Quaternary International*, 285, 1–29. Retrieved from https://doi.org/10.1016/j.quaint.2012.09.023

Derungs, C., Köhl, M., Weibel, R., & Bickel, B. (2018). Environmental factors drive language density more in food-producing than in hunter–gatherer populations. *Proceedings of the Royal Society B: Biological Sciences*, 285(1885), 20172851. Retrieved from https://doi.org/10.1098/rspb.2017.2851

Dunbar, R. I. M. (2008). Cognitive constraints on the structure and dynamics of social networks. *Group Dynamics: Theory, Research, and Practice*, 12(1), 7–16. Retrieved from

https://doi.org/10.1037/1089-2699.12.1.7

Fincher, C. L., & Thornhill, R. (2008). A parasite-driven wedge: infectious diseases may explain language and other biodiversity. *Oikos*, 117(April), 1289–1297. Retrieved from https://doi.org/10.1111/j.2008.0030-1299.16684.x

Gavin, M. C., Botero, C. A., Bowern, C., Colwell, R. K., Dunn, M., Dunn, R. R., … Yanega, G. (2013). Toward a Mechanistic Understanding of Linguistic Diversity. *BioScience*, 63(7), 524–535. Retrieved from https://doi.org/10.1525/bio.2013.63.7.6

Gavin, M. C., Kavanagh, P., Botero, C., Bowern, C., Ember, C., Haynie, H., … Kirby, K. (2018). The global geography of human subsistence. *Royal Society Open Science*, 5, 171897. Retrieved from https://doi.org/10.1098/rsos.171897

Gavin, M. C., Rangel, T. F., Bowern, C., Colwell, R. K., Kirby, K. R., Botero, C. A., … Gray, R. D. (2017). Process-based modelling shows how climate and demography shape language diversity. *Global Ecology and Biogeography*, 26(5), 584–591. Retrieved from https://doi.org/10.1111/geb.12563

Gavin, M. C., & Sibanda, N. (2012). The island biogeography of languages. *Global Ecology and Biogeography*, 21(10), 958–967. Retrieved from https://doi.org/10.1111/j.1466-8238.2011.00744.x

Gavin, M. C., & Stepp, J. R. (2014). Rapoport's Rule Revisited: Geographical Distributions of Human Languages. *PLoS ONE*, 9(9), e107623. Retrieved from https://doi.org/10.1371/journal.pone.0107623

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Boca Raton, FL.: CRC Press.

Greenhill, S. J., Hua, X., Welsh, C. F., Schneemann, H., & Bromham, L. (2018). Population Size and the Rate of Language Evolution: A Test Across Indo-European, Austronesian, and Bantu Languages. *Frontiers in Psychology*, 9(April), 1–18. Retrieved from https://doi.org/10.3389/fpsyg.2018.00576

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2020). Glottolog 4.3. Retrieved from https://doi.org/10.5281/zenodo.4061162

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15), 1965–1978. Retrieved from https://doi.org/10.1002/joc.1276

Hornborg, A. (2005). Ethnogenesis, Regional Integration, and Ecology in Prehistoric Amazonia: Toward a System Perspective. *Current Anthropology*, 46(4), 589–620.

Hua, X., Greenhill, S. J., Cardillo, M., Schneemann, H., & Bromham, L. (2019). The ecological drivers of variation in global language diversity. *Nature Communications*, 10(1), 2047. Retrieved from https://doi.org/10.1038/s41467-019-09842-2

Kavanagh, P. H., Vilela, B., Haynie, H. J., Tuff, T., Lima-Ribeiro, M., Gray, R. D., … Gavin, M. C. (2018). Hindcasting global population densities reveals forces enabling the origin of agriculture. *Nature Human Behaviour*, 2(7), 478–484. Retrieved from https://doi.org/10.1038/s41562-018-0358-8

Lehner, B., Verdin, K., & Jarvis, A. (2008). New Global Hydrography Derived From Spaceborne Elevation Data. *Eos, Transactions American Geophysical Union*, 89(10), 93. Retrieved from https://doi.org/10.1029/2008EO100001

Levinson, S. C., & Gray, R. D. (2012). Tools from evolutionary biology shed new light on the diversification of languages. *Trends in Cognitive Sciences*, 16(3), 167–173. Retrieved from https://doi.org/10.1016/j.tics.2012.01.007

Lima-Ribeiro, M. S., Varela, S., González-Hernández, J., Oliveira, G. de, Diniz-Filho, J. A. F., & Terribile, L. C. (2015). EcoClimate: A database of climate data from multiple models

for past, present, and future for macroecologists and biogeographers. *Biodiversity Informatics*, 10, 1–21.

Maffi, L. (2005). Linguistic, Cultural, and Biological Diversity. *Annual Review of Anthropology*, 34(1), 599–617. Retrieved from https://doi.org/10.1146/annurev.anthro.34.081804.120437

Manne, L. L. (2003). Nothing has yet lasted forever: Current and threatened levels of biological and cultural diversity. *Evolutionary Ecology Research*, 5(4), 517–527.

Marwick, B. (2017). Datasets used in Binford's 2001 book 'Constructing Frames of Reference: An Analytical Method for Archaeological Theory Building Using Ethnographic and Environmental Data Sets'. Retrieved from https://github.com/benmarwick/binford

Matthews, L. J., Passmore, S., Richard, P. M., Gray, R. D., & Atkinson, Q. D. (2016). Shared cultural history as a predictor of political and economic changes among nation states. *PLoS ONE*, 11(4). Retrieved from https://doi.org/10.1371/journal.pone.0152979

Michalopoulos, S. (2008). The Origins of Ethnolinguistic Diversity: Theory and Evidence. *SSRN Electronic Journal*, 0–54. Retrieved from https://doi.org/10.2139/ssrn.1286893

Mirazón Lahr, M. (2016). The shaping of human diversity: filters, boundaries and transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1698), 20150241. Retrieved from https://doi.org/10.1098/rstb.2015.0241

Moore, J. L., Manne, L., Brooks, T., Burgess, N. D., Davies, R., Rahbek, C., … Balmford, A. (2002). The distribution of cultural and biological diversity in Africa. *Proceedings of the Royal Society B: Biological Sciences*, 269(1501), 1645–1653. Retrieved from https://doi.org/10.1098/rspb.2002.2075

Mufwene, S. (2002). Colonisation, Globalisation, and the Future of Languages in the Twenty-First Century. *International Journal of Multicultural Societies*, 4(2), 162–193.

Nettle, D. (1996). Language Diversity in West Africa: An Ecological Approach. *Journal of Anthropological Archaeology*, 15(4), 403–438. Retrieved from https://doi.org/10.1006/jaar.1996.0015

Nettle, D. (1998). Explaining Global Patterns of Language Diversity. *Journal of Anthropological Archaeology*, 17, 354–374.

Pacheco Coelho, M. T., Pereira, E. B., Haynie, H. J., Rangel, T. F., Kavanagh, P., Kirby, K. R., … Gavin, M. C. (2019). Drivers of geographical patterns of North American language diversity. *Proceedings of the Royal Society B: Biological Sciences*, 286(1899), 20190242. Retrieved from https://doi.org/10.1098/rspb.2019.0242

Rangel, T. F., Edwards, N. R., Holden, P. B., Diniz-Filho, J. A. F., Gosling, W. D., Coelho, M. T. P., … Colwell, R. K. (2018). Modeling the ecology and evolution of biodiversity: Biogeographical cradles, museums, and graves. *Science*, 361(6399), eaar5452. Retrieved from https://doi.org/10.1126/science.aar5452

Silva, N. K. (2004). *Aloha Betrayed: Native Hawaiian Resistance to American Colonialism*. Durham, NC: Duke University Press Books.

Simons, G. F., & Fennig, C. D. (Eds.). (2018). *Ethnologue: Languages of the World* (Twenty-fir). Dallas, Texas: SIL International.

Spehn, E. M., & Körner, C. (2005). A global assessment of mountain biodiversity and its function. In U. M. Huber, H. K. M. Bugmann, & M. A. Reasoner (Eds.), *Global Change and Mountain Regions* (pp. 393–400). Springer, Dordrecht.

Sutherland, W. J. (2003). Parallel extinction risk and global distribution of languages and species. *Nature*, 423(6937), 276–279. Retrieved from https://doi.org/10.1038/nature01607

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Vilela, B., Fristoe, T., Tuff, T., Kavanagh, P. H., Haynie, H. J., Gray, R. D., … Botero, C. A. (2020). Cultural transmission and ecological opportunity jointly shaped global patterns of reliance on agriculture. *Evolutionary Human Sciences*, 2, e53. Retrieved from https://doi.org/10.1017/ehs.2020.55

Watts, J., Sheehan, O., Atkinson, Q. D., Bulbulia, J., & Gray, R. D. (2016). Ritual human sacrifice promoted and sustained the evolution of stratified societies. *Nature*, 532(7598), 228–231. Retrieved from https://doi.org/10.1038/nature17159

Yoo, C., & Cho, E. (2018). Comparison of GCM precipitation predictions with their RMSEs and pattern correlation coefficients. *Water (Switzerland)*, 10(1). Retrieved from https://doi.org/10.3390/w10010028

SUPPLEMENTARY MATERIAL

**Supplementary video**

**Supplementary figures**

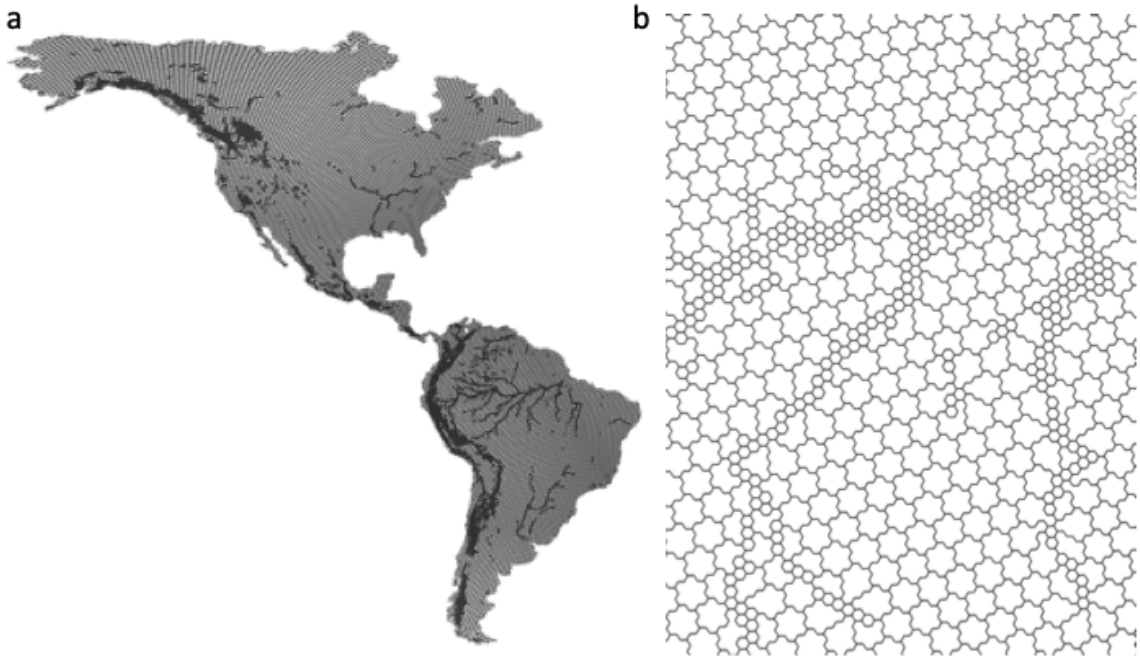


**Figure S1**. A) Hybrid hexagonal grid for the American continent. Smaller cells represent river and altitude complexity, while larger cells are clusters of non-complex (plain) cells. b) A region of the map zoomed-in to better illustrate the rescaling and cell sizes of the grid.

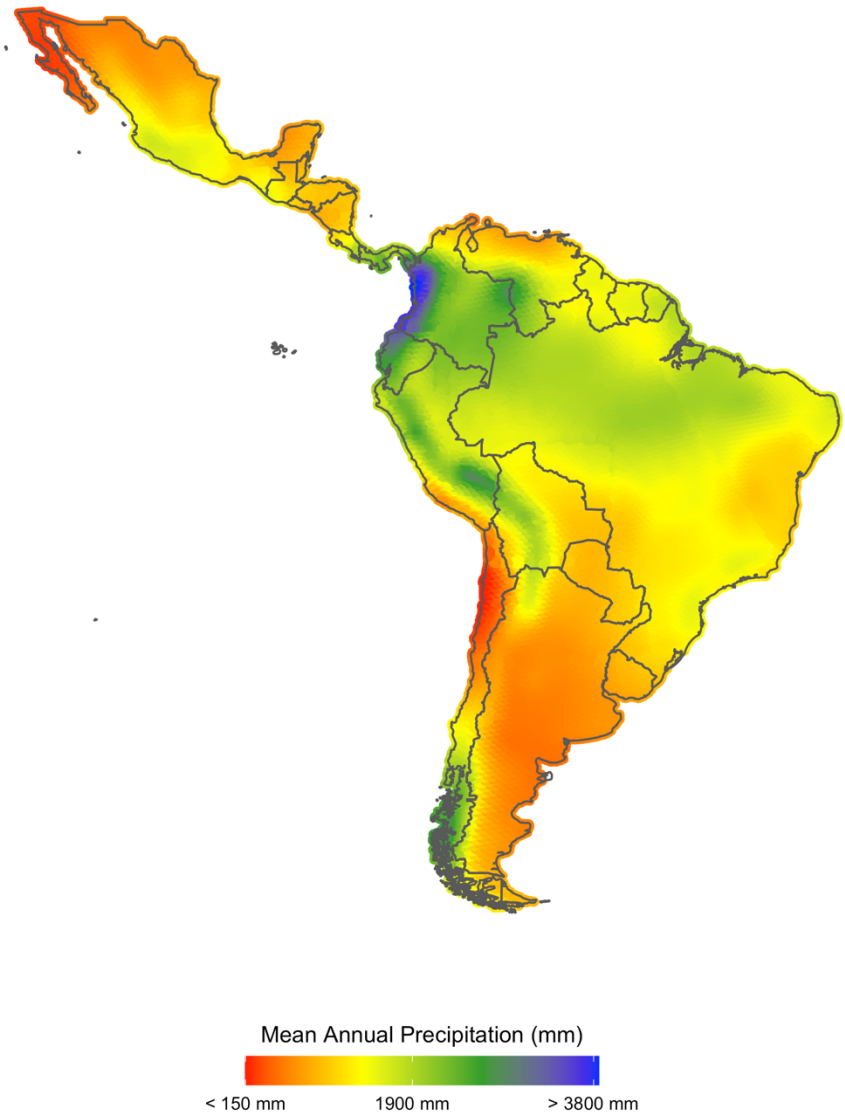


**Figure S2.** Map of mean annual precipitation in Central and South America for the pre-industrial period (1760). Data was extracted from the ecoClimate database at a resolution of 0.5˚ x 0.5˚ latitude/longitude (Lima-Ribeiro *et al.* 2015).

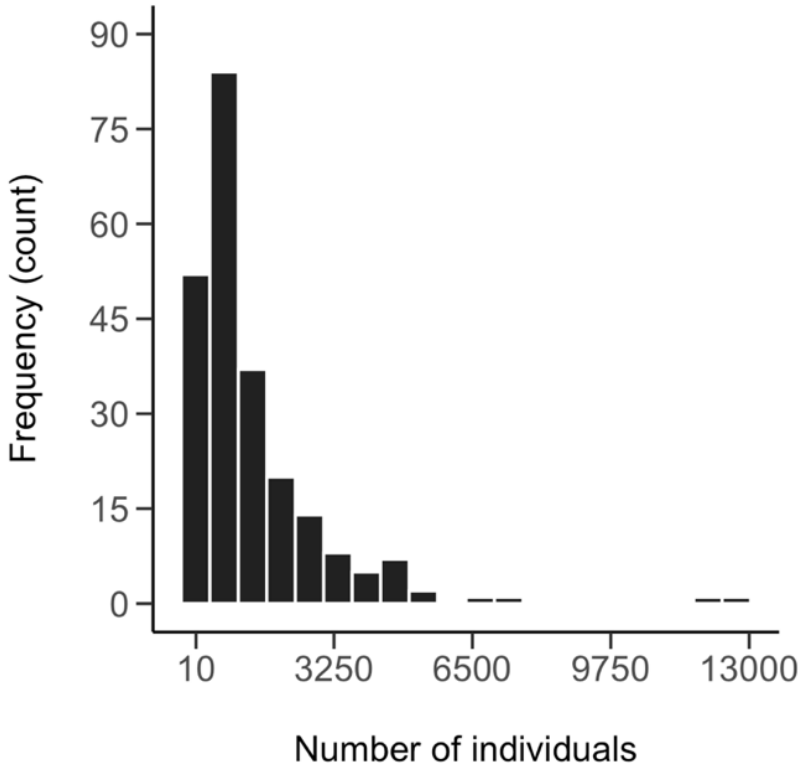


**Figure S3.** Hunter-gatherer (HG) groups sizes' empirical distribution (Binford, 2001). We excluded HG data from the artic and subarctic, as these climates do not occur in the Neotropics, and from Mexico and South America, to avoid redundancy in our results. From a total of 339 observations, we were left with 233 observations.

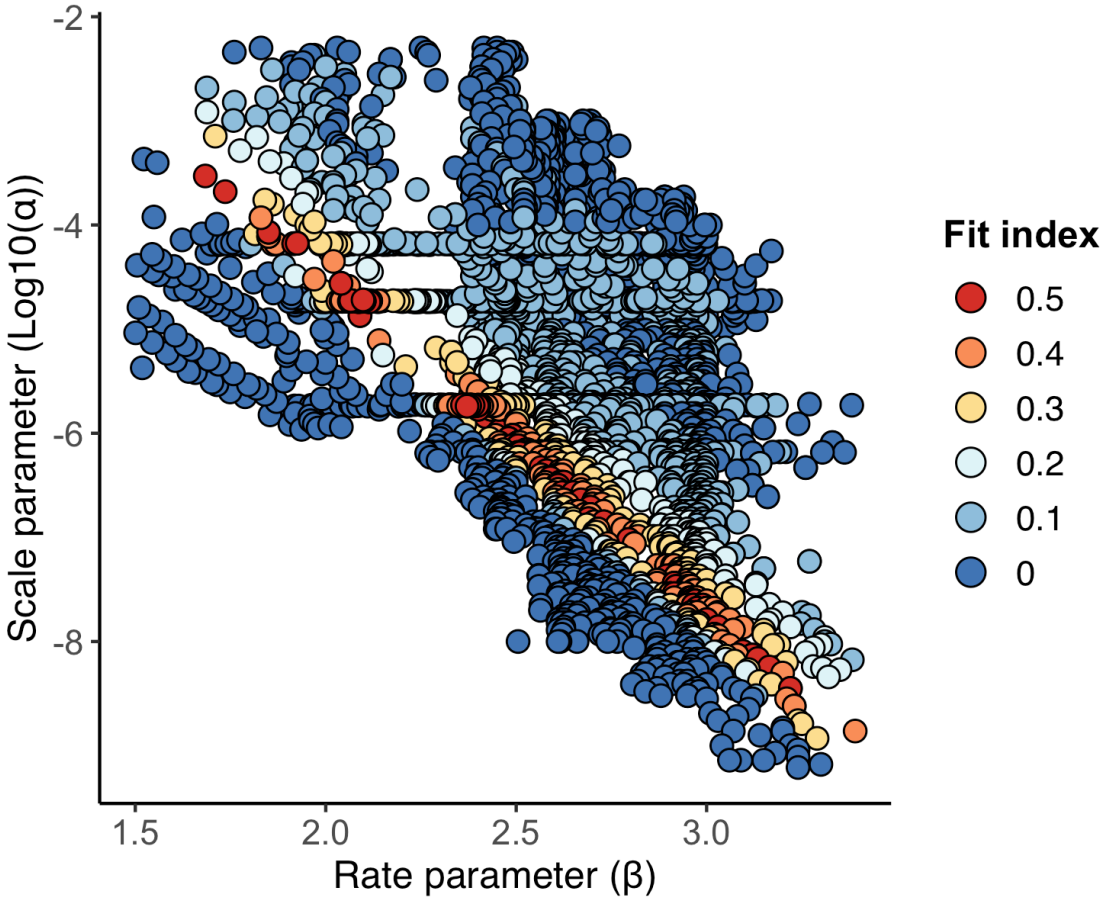


**Figure S4.** Carrying capacity parameters, scale and rate, which define the carrying capacity in each cell given the precipitation in the cell. The dots represent the goodness of fit for the parameter combination and model results, converging clearly in a log-linear relationship. Parameter combinations were evaluated by a Gibbs sampler.

**CAPÍTULO 4**

---

*The effectiveness of state-level interventions on the early spread of COVID-19*

Christielly Mendonça Borges[1*], Marco Túlio Pacheco Coelho[1], José Alexandre Felizola Diniz-Filho[1], Thiago Fernando Rangel[1]

[1]Departamento de Ecologia, Universidade Federal de Goiás, Goiânia, Goiás 74.001-970, Brasil.

[*]Corresponding author

*Status*: Em preparação para submissão

**ABSTRACT**

Brazil was deeply affected by the COVID-19 pandemic, nonetheless, the federal government never proposed a coordinated action to control it. The state of Goiás declared strict social distancing measures at an early stage, but gradually relaxed many of them due to public pressure. We seek to detect if change points in the effective growth rate of COVID-19 can be explained by the governmental interventions made in Goiás in 2020. We use a Susceptible-Infected-Recovered model combined with Bayesian inference and a time-dependent spreading rate to assess how past state-level interventions affected the spread of COVID-19 from March to May of 2020. The interventions succeeded in decreasing the transmission rate in the state, however, after the third intervention the rate remained positive and exponential. Our results reflect the efficiency of governmental interventions, but also the population's low compliance with the measures. This highlights the need for cooperation between governments and public in the fight against COVID-19. As long as there is no effective treatment or widely available vaccination, the safest way to fight this pandemic is still non-pharmaceutical interventions.

**Keywords:** Bayesian inference, public health policy**,** SARS-COV-2, time-dependent SIR model

## 1. Introduction

The coronavirus disease 2019 (COVID-19) outbreak caught the world by surprise, as in three months it went from a public health emergency of international concern to a global pandemic[1]. This is the first pandemic caused by a coronavirus, since the severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) is highly transmissible and causes a pathogenic viral infection[2]. Human to human spread occurs mainly through respiratory droplets and contact routes[2], but the virus can remain infections in aerosols and surfaces[3]. While vaccines were available by December 2020, no effective treatment exists and preventions of disease spread and mitigation of COVID-19 still relies on non-pharmaceutical interventions (NPIs) such as social distancing, isolation, face covering and quarantine measures[4].

Despite being recommended by the World Health Organization (WHO), the adoption of NPIs varied greatly between countries. The virus emerged first in China, where strict social distancing rules were enforced early and only three months later the spreading was contained[5]. Timing of intervention during the outbreak can explain discrepancies in the number of deaths in European countries, where Italy had 525 deaths per million population in contrast to Germany's 95 deaths per million population in the same month[6]. As of August 2020, the three countries with the highest number of COVID-19 related deaths were the US, Brazil and Mexico[7], respectively, all countries where heads of state and government openly spoke against NPIs[8–10].

The first confirmed COVID-19 case in Brazil was registered on February 25 of 2020, in the city of São Paulo[11]. By March, it had already reached all 26 states and the Federal District[12]. By August 17, Brazil reported over three million confirmed cases and a total of 107,852 deaths[7]. The states with the higher and lower confirmed cases were São Paulo and Mato Grosso do Sul,

with 699,493 and 36,836 cases, respectively[7]. States are autonomous under the Brazilian constitution[13], nonetheless, there was no coordinated action to control COVID-19 by the federal government.

Most state governors enforced restrictive contact measures in mid-March, when the virus began spreading. However, the president of the Federal Government has been vocally against state-level NPI policies, citing frequently his fear of an economic collapse[14]. In fact, president Bolsonaro passed a provisional measure in April which entrusted to the Union prerogatives concerning isolation, quarantine and the interdiction of locomotion, public services and essential activities during the pandemic[15]. This measure was quickly overruled by the Supreme Federal Court of Justice, instating the Union could legislate on the subject but must always safeguard the autonomy of states and municipalities[15]. This political confrontation deepened an already existing rift in the population, with Bolsonaro's supporters positioning themselves against states and municipalities' COVID-19 containment measures.

Brazil is a continental sized country with substantial regional socioeconomic inequalities, all factors that further reduce support for NPIs[12], and results in different containment measures in different states. In April 2020, while cities such as Manaus, Fortaleza, Brasília, Rio de Janeiro and São Paulo faced an exponential growth of COVID-19 cases, southern states Santa Catarina and Rio Grande do Sul started to reopen their economies without many registered cases[16,17]. The state of Goiás was amongst the leaders of social isolation, registering in March over 60% of reduced mobility monitored via geolocation in smartphones[18]. That percentage eventually dropped and in July, Goiás recorded only 37% of reduced mobility, one of the worst rates of isolation in the country[18]. As a consequence, cases and deaths started to increase fast.

The first confirmed COVID-19 case in Goiás was registered on March 12, 2020. By March 13, the state government issued a decree declaring public health emergency and instituted strict social distancing measures. However, due to public and economic pressure, the government gradually relaxed many of its first measures[19]. Thus, in this paper we assess the early transmission dynamics and evaluate the effectiveness of state-level interventions. Short-term forecasts such as this are key to estimate medical requirements and capacities, and here we use it to assess how past mitigations affected the spread of COVID-19 in the state.

## 2. Materials and methods

We reproduced the framework established by[20] and available at[21]. They combined SIR models with Bayesian parameter inference with Markov Chain Monte Carlo (MCMC) sampling and augmented the model by a time-dependent spreading rate, which is implemented via potential change points that characterize governmental interventions[20]. We adjusted the model, initially created for Germany, for the state of Goiás, by choosing the three main state-level interventions and parameters accordingly. Here we ran (1) a SIR model for the onset period with stationary spreading rate (simple SIR model) and (2) a time-dependent SIR model with weekend correction (full SIR model)[20].

SIR models have been used broadly to model epidemic spreads[22,23], and recently gained strength in efforts to model the spread of COVID-19 worldwide[20,24]. SIR models specify the rates that population recover and become infected by a disease. Bayesian inference with MCMC sampling assimilates prior knowledge available and accounts for data uncertainties into forecasts. An integration between Bayesian inference and SIR models provide a better assessment of more complex and realistic models[25].

*2.1. Goiás characterization and data*

The state of Goiás is located in the mid-west region of Brazil and has a population of approximately 7 million people[26]. The Federal District, along with the country's capital Brasília, is geographically embedded within the State of Goiás, but due to administrative differences and independence of public health policies, here we analyze only data for Goiás. Daily number of COVID-19 confirmed cases in Goiás came from the Goiás State Health Department (SES-GO; acronym in Portuguese). SES-GO systematically monitors suspected cases throughout the state and provides daily updates of confirmed cases[27]. We used data until May 22 of 2020.

*2.2. Governmental interventions*

As of May 22, Goiás had amounted 14 decrees regarding the coronavirus pandemic. Most of these decrees are relaxations of the first decree, such as reopening churches and temples. To implement and maintain the model simple, we chose three main decrees capable of influencing public behavior (Fig. 1).

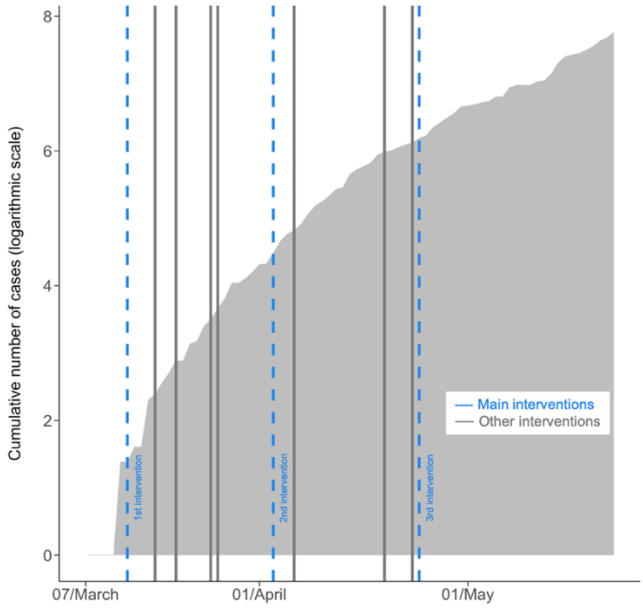


**Figure 1.** The cumulative cases of COVID-19 in the state of Goiás (logarithmic scale) and the 14 interventions made by the state's government (until May 22). We chose the three main

decrees capable of influencing public behavior as our three main interventions (dashed lines). The first intervention was on March 13; The second intervention was on April 3; and the third intervention was on April 24. Other interventions are represented by the gray solid lines. The curve represents confirmed cases.

The first intervention chosen was the first decree announced on March 13. In this decree the state declares a public health emergency and institutes strict social distancing measures, such as the shutdown of public and private events of any nature, including educational institutions at all levels, daycares, suspension of commercial activities such as malls, fairs, gyms, dental health services, religious meetings and all other non-essential services and activities[28].

The second intervention chosen was the decree from April 3rd, which was already the eighth decree announced and the fourth relaxing the measures stated in the first one. This decree accumulates all prior flexibilizations, including reopening of religious activities, beauty salons, vegetable and fruit fairs, car workshops and restaurants on highways, administrative activities in public and private educational institutions[28].

On April 19, the government launched a new decree extending the health emergency in Goiás for another 150 days. However, on April 24, they announced another decree legislating on the private sphere as well (suspending activities of common use in closed condominiums), regulating a channel for reporting disobediences to any of the decrees, and legislating specific days for religious celebrations[28]. We chose the intervention on April 24 as the third changing point, as we see it to be more rigorous than previous ones.

*2.3. Simple SIR model: stationary spreading rate*

We considered the initial onset transmission phase as being between March 6 and 20, approximately seven days before and after the first confirmed COVID-19 case in Goiás. Central

epidemiological parameters for this model are the spreading rate (λ), recovery rate (μ), reporting delay ($D$) and number of initially infected people ($I_0$)[20]. We chose informative log-normal priors of λ = 0.3 and μ = 0.11 (Table 1), as these priors cannot be estimated independently and these values maintain the effective growth rate ($\lambda^* = \lambda - \mu$) with a median of 0.19 and the basic reproduction number ($R_0 = \lambda / \mu$) with a median of 2.72, consistent with global[29] and local[30] estimates. We chose for the reporting delay a prior that incorporates the virus' incubation period between 1-14 days and the delay of infected people awaiting tests confirmation or medical appointments. Flat priors were chosen for the $I_0$ and scale factor.

**Table 1.** Priors for the simple SIR model with stationary spreading rate

| **Parameter** | **Variable** | **Prior distribution** |
|---|---|---|
| Spreading rate | λ | LogNormal(log ( 0.3 ), 0.5) |
| Recovery rate | μ | LogNormal(log ( 1/9 ), 0.2) |
| Reporting Delay | $D$ | LogNormal(log ( 8 ), 0.2) |
| Initially infected | $I_0$ | HalfCauchy( 100 ) |
| Scale factor | σ | HalfCauchy( 10 ) |

*2.4. Full SIR model: weekly reporting modulation and change points in spreading rate*

To simulate the effect of governmental interventions, we use a SIR model with incorporated change points capable of altering the transmission rate[20]. The aim of interventions is to reduce the effective growth rate, thus if the rate becomes negative, new infections will begin to decrease. The model assumes new spreading rates for each change point, inferred after supposed behavioral changes in the population.

We chose the same log-normal distributed priors for $\lambda_0$, $\mu$ and $D$ as in the simple model, with added parameters for the change points and their spreading rates (Table 2). We assumed the first government intervention reduced the spreading rate by 50% from the initial estimate $\lambda_0 = 0.3$, so the prior for the first change point is $\lambda_1 \sim$ 159ognormal (log (0.15), 0.5). Given the flexibilizations in the following decrees, we assumed the spreading rate would increase again by 15%, thus $\lambda_2 \sim$ 159ognormal (log (0.22), 0.5). For the third intervention, there was more rigidity in the social distancing measures, which we presumed was embraced by the population. Therefore, we assumed the prior decreases the spreading rate and is closer (but slightly inferior) to the rate of the first intervention $\lambda_3 \sim$ 159ognormal (log (0.11), 0.5).

**Table 2.** Priors for the full SIR model with change points and weekly reporting modulation

| **Parameter** | **Variable** | **Prior distribution** |
|---|---|---|
| Change points | $t_1$ | Normal(2020 / 03 / 06, 3) |
| | $t_2$ | Normal(2020 / 04 / 03, 1) |
| | $t_3$ | Normal(2020 / 04 / 24, 1) |
| Change duration | $\Delta t_i$ | LogNormal(log ( 3 ), 0.3) |
| Spreading rates | $\lambda_0$ | LogNormal(log ( 0.3 ), 0.5) |
| | $\lambda_1$ | LogNormal(log( 0.22 ), 0.5) |
| | $\lambda_2$ | LogNormal(log( 0.22 ), 0.5) |
| | $\lambda_3$ | LogNormal(log (0.11), 0.5) |
| Recovery rate | $\mu$ | LogNormal(log (1/9), 0.2) |
| Reporting delay | $D$ | LogNormal(log (8), 0.2) |
| Weekly modulation amplitude | $f_w$ | Beta(mean = 0.7, std = 0.17) |
| Weekly modulation phase | $\Phi_w$ | vonMises(mean = 0, $k$ = 0.01) |
| Initially infected | $I_0$ | HalfCauchy( 100 ) |

We chose normal distributed priors for the timing of change points (Table 2). Respectively $t_1$ ~ Normal (2020/03/13, 3), $t_2$ ~ Normal (2020/04/03, 1) and $t_3$ ~ Normal (2020/04/24, 1), where 3, 1 and 1 are the respective transient days. Following the logic of the aforementioned decrees, we assumed the first intervention as a strict contact ban, the second as a mild contact ban and the third as a strict contact ban. The change points take effect after a period of time ($\Delta t_i$), for which we chose a median of 3 days. During these 3 days, spreading rates are expected to change for interventions to take effect. Furthermore, time is needed to ensure a smooth transition capable of absorbing the changes in the population's behavior [20].

Priors chosen for the recovery rate, reporting delay and initial number of infected people were the same as those applied in the simple SIR model (Table 1). The number of tests and reported cases varies throughout the week, with the number of records expected to be lower on weekends[20]. To implement the weekend effect in the model, we modulate the number of cases inferred by the absolute value of a sine function with the total period of 7 days[20]. We chose flat priors for the $I_0$, scale factor and weekly modulation phase.

*2.5. Model comparison*

Following[20], we ran a model comparison using the leave-one-out (LOO) cross-validation method to avoid an over-fitting forecast. We compared four full SIR models with zero, one, two, and three change points, respectively. The full SIR model with three change points presented a better match between model and data (Table 3), as indicated by a lower LOO score. Full SIR models with zero, one and two change points performed poorly and will not be further discussed (but see Supplementary Fig. S1-S2).

**Table 3.** LOO cross-validation for SIR models (with weekend correction) with a different number of change points

| Model | LOO-score | Effective number of parameters (pLOO)[a] |
|---|---|---|
| Zero change points | 598.9 ± 13.35 | 9.32 |
| One change point | 597.18 ± 12.91 | 8.14 |
| Two change points | 595.52 ± 12.48 | 9.04 |
| Three change points | 592.26 ± 12.95 | 9.88 |

[a] Lower LOO-score indicates a better match between model and data.

We also ran a model comparison with three change points but no weekend modulation and three sensitivity analyses by choosing wider priors for different parameters[20]. The model without a weekend modulation removes the assumption that daily reporting of new cases happens mainly during weekdays, inferred parameters change only for the number of initial infections (Fig. S3). For the sensitivity analysis, all parameters and priors were maintained exactly as the full SIR model, except were indicated. We ran a model with a prior four times wider for the reporting delay, a model with a prior 14 days wide for the change times and a model with a prior four times wider for the change duration (Fig. S4-S6). The full SIR model with three change points and weekly modulation again performed better than other models given the lower LOO-score (Table S1).

## 3. Results

The daily reported cases in Goiás did not present an exponential curve in the simple SIR model with stationary spreading rate (Fig. 2A), and the total reported cases (accumulated cases) show a tendency to be exponential (Fig. 2B). The spreading rate was adjusted by the model as $\lambda = 0.16$ (95% credible interval (CI [0.07, 0.33]; Fig. 2E)) and the effective growth rate as $\lambda^* = 0.04$

(Fig. 2H), values lower than our prior. Further, μ and *D* histograms match the priors (gray line), as expected by the model[20]. The data for the initial phase is scarce and noisy, partly because the initial cases were not local infections, but of contaminated people arriving in Goiás from other states. This phenomenon is not captured by our model; thus, we will not be discussing these results further.

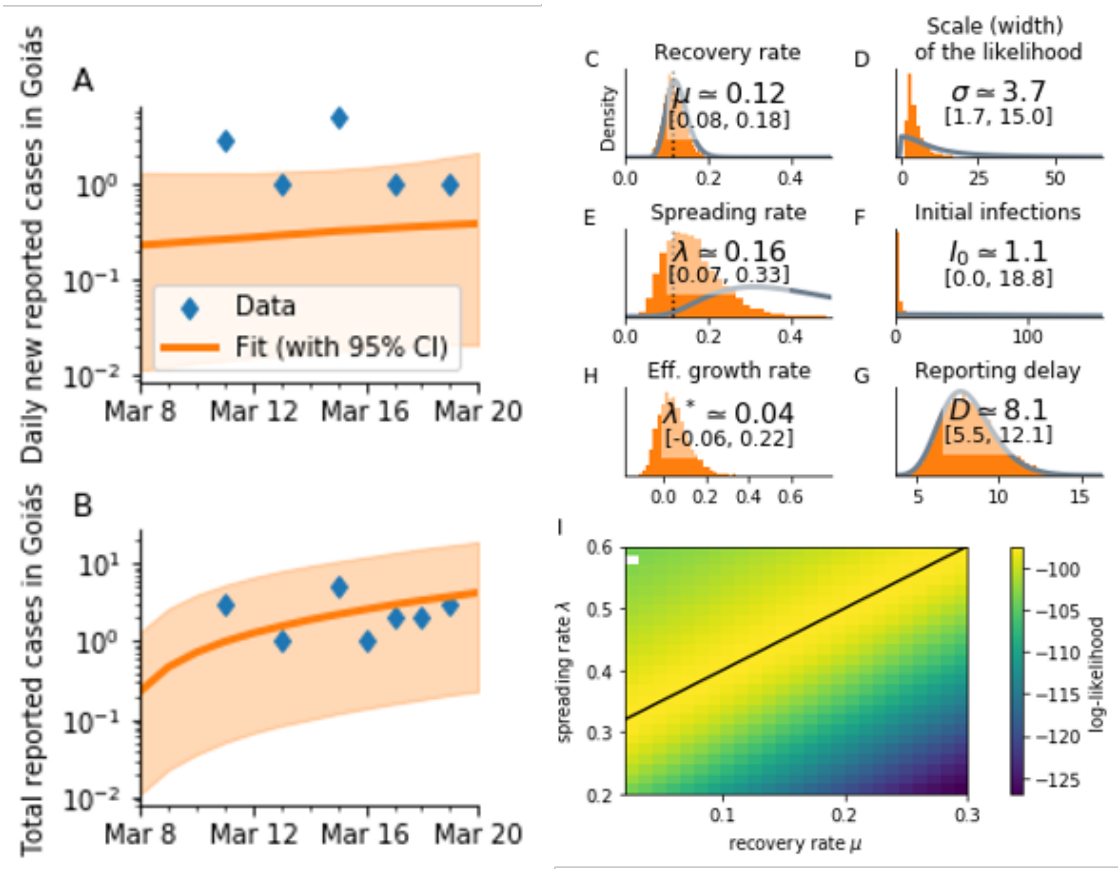


**Figure 2.** Results for the simple SIR model with stationary spreading rate during the initial onset period, March 8-20. (A) Daily new reported cases in Goiás and; (B) Total (cumulative) reported cases in Goiás. (C-G) Inference of central epidemiological parameters: prior (solid line) and posterior distributions (shaded region); (C) Recovery rate μ; (D) Scale-factor of the width of the likelihood distribution σ; © estimated spreading rate λ; (F) Initial infections $I_0$; (H) Effective growth rate $\lambda^*$; (G) reporting delay *D*. (I) Log-likelihood distribution for different combinations of λ and μ, the black line indicates a linear combination that yields the same maximal likelihood and the white dot indicates where inference did not converge.

We found clear evidence of the influence of the three interventions in the full SIR model with change points and weekly reporting modulation (Fig. 3). First, the spreading rate decreased

from $\lambda_0 = 0.28$ (CI [0.12, 0.42]) to $\lambda_1 = 0.21$ (CI [0.16, 0.29]). The date for the first change point was inferred as March 13 (CI [9, 21]), a date that marks the first state decree with strict contact ban measures. After this intervention, the effective growth rate was a median of $\lambda_0 - \mu = 0.18$ to median $\lambda_1 - \mu = 0.11$, given $\mu$ was inferred as 0.10 (CI [0.07,0.14]).

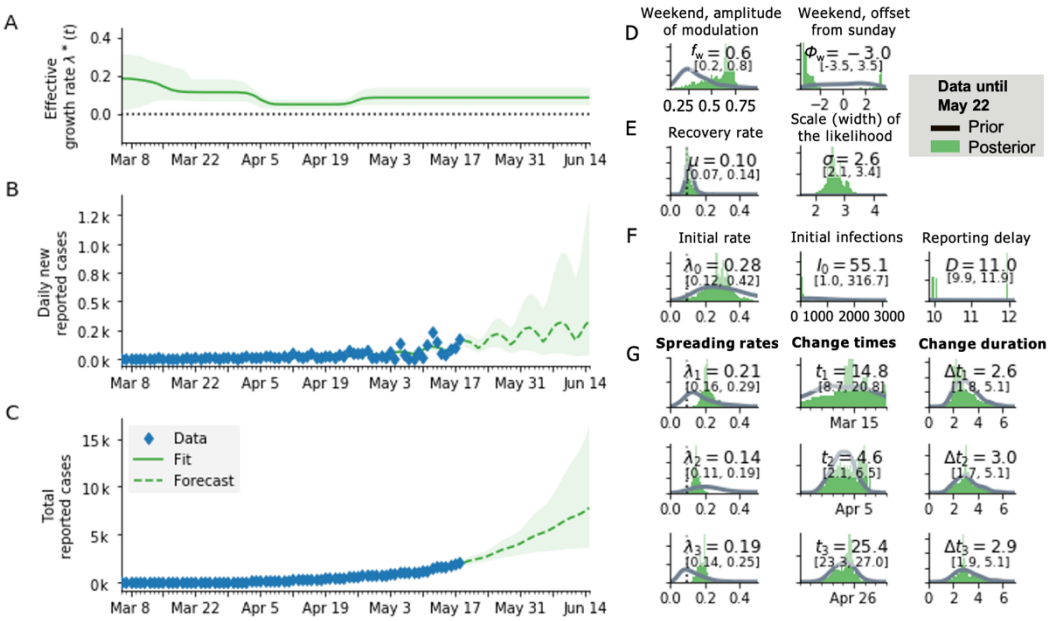


**Figure 3.** Results for the full SIR model with three change points and weekly reporting modulation. (A) Estimate of the effective spreading rate; (B) Daily new reported cases (diamonds) and the model (solid line for median fit with 95% credible intervals). Green dashed line is the median forecast with 95% CI. (C) Total reported cases and the model (representation symbols same as in B). (D-F) Inference of central epidemiological parameters: prior (solid line) and posterior distributions (shaded region), inset values indicate the median and 95% CI of posteriors. (G) Spreading rates, change times and change duration for the three change points, respectively.

At the second change point, $\lambda_t$ decreased from $\lambda_1 = 0.21$ to $\lambda_2 = 0.14$ (CI [0.11, 0.19]), lower than assumed by our prior. This date was inferred as April 3 (CI [2, 6]), which marks the accumulation of flexibilizations from four decrees to first decree of March 13, including the reopening of religious events and fruit and vegetable fairs. After the second intervention, the median growth rate was $\lambda_2 - \mu = 0.04$, in the vicinity of a critical point (close to zero), but still positive.

The third change point increased $\lambda_2 = 0.14$ to $\lambda_3 = 0.19$ (CI [0.14, 0.25]). This change point was inferred to be April 24 (CI [23, 27]), a stricter decree compared to the previous ones. After this measure, the effective growth rate was of $\lambda_3 - \mu = 0.09$, indicating an increase in the growth, remaining above zero and thus not decreasing the number of new infections.

**4. Discussion**

Given our results, the first two state-level interventions drastically reduced the COVID-19 spreading rate in Goiás. We expected the second intervention to increase the spreading rate, given prior relaxations, but the rate dropped. A plausible explanation could be that despite the relaxations of non-essential stores and services, the population complied to NPIs. Nonetheless, the third intervention, although stricter than the second, brought the transmission rate back to a rate similar to that of the first intervention. We expected the third intervention to result in a similar transmission rate to that of the first intervention, but in the model, the transmission rate increased. This result probably reflects an accumulation of all fourteen state-level decrees and the population's fatigue of being isolated, which resulted in more people eventually circulating in public spaces. Significant increase in mobility after March 2020 has been evidenced for all Brazilian states following state policies relaxations[31].

The transmission rates found in our model further match the patterns found in a study that calculated time-series of the effective reproductive number (*Rt*) in Goiás[32]. They found *Rt* to be around 2.0 in mid-March, dropping to approximately 1.2 in mid-April and increasing slowly to 1.4-1.5 in May. If we convert our transmission rates to R ($R = 1 + \lambda * 5.2$; where 5.2 is the serial interval) we get R = 2.1 for the first intervention in March 13, R = 1.2 for the second

intervention in April 3 and R = 1.5 for the third intervention in April 24. Both our models reflect the effects of the early NPIs implemented in Goiás in reducing the diseases' onset transmission.

When this model was applied to Germany it demonstrated that following a gradual linear path of interventions, first banning major public events, later announcing mild social distancing measures and finally a strict contact ban[20], aided in decreasing the transmission rate, bringing it to almost zero. Goiás followed an almost inversed path, imposing a first decree with strict social distancing measures at an early stage, but eventually reopening many specific services and activities. Nonetheless, if no NPI had been imposed in Goiás, up to 62% of the population would have been infected by June 2nd [33], representing approximately 4 million people in the state. Further, it was also estimated that the interventions in Goiás prevented between 2.834 and 3.407 COVID-19 deaths[33].

Although state-level interventions succeeded in decreasing the transmission rate, it remained high and exponential. Thus, other stricter interventions were made necessary to avoid the growth of new cases and a collapse in the health system. Nonetheless, more restrictive measures for the containment of COVID-19 were not adopted and were only discussed again in late June, when confirmed cases spiked. Our model forecasted for June 14 approximately 8,187 total reported cases (Fig. 3C), at that date, the state registered 7,944 confirmed cases[27], a difference of 243 cases that could be explained by under testing and reporting delays.

Governmental interventions need to be taken seriously by the public in order for them to have the proposed outcome. Our results reflect the efficiency of NPIs in containing the spread of COVID-19, even if short-lived, but also the population's disregard with the measures imposed. Factors previously identified with low compliance to NPIs include lack of willpower to follow

interventions, peers actively discouraging use of masks, fake news and misinformation, political polarization and social inequality[34,35]. Solutions to these problems are not straightforward, nonetheless, better communication efforts from governments can help mitigate community compliance and distrust[35]. It has further been shown that compliance to NPIs is more successful when a coherent policy set is sustained over time[31].

The COVID-19 outbreak poses itself as the biggest public health challenge in the last 100 years. Many countries around the world took drastic measures of social distancing and even complete lockdowns to contain it. The year 2021 started with several countries vaccinating their population against COVID-19. Brazilians showed themselves to be quite receptive to the COVID-19 vaccine[36], however the federal government initially acquired a low number of doses and was reluctant to buy more from specific institutions[37]. This further demonstrates the necessary cooperation between governments and public in the fight against COVID-19. As vaccines are not widely and equitably available worldwide[38], and to contain the spread of COVID-19 and the emergence of new variants, the safest way to successfully fight this pandemic is still social distancing and face covering.

**Authors' contributions**

All authors conceived and designed the study. TFR curated the data. CMB performed the analyses and wrote the first draft. All authors provided critical feedback, revised and approved the manuscript's final version.

**Conflict of Interests**

The authors declare no conflict of interests.

**Acknowledgments**

We thank Mario Joaquim dos Santos Neto for the discussions and comments on the Brazilian legal system. We thank the Goiás State and Goiânia City Health Departments for support and access to original data. CMB is supported by a CAPES scholarship. JAFDF and TFR are continuously supported by the National Institute of Science and Technology (INCT) in Ecology, Evolution and Biodiversity Conservation, supported by CNPq and FAPEG. JAFDF and TRF are also supported by CNPq's research productivity scholarships and grants.

**References**

1. WHO. Timeline of WHO's response to COVID-19. *World Health Organization*. 2020.
2. Shereen MA, Khan S, Kazmi A, Bashir N, Siddique R. COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. *J Adv Res*. 2020;24:91-98. Doi:10.1016/j.jare.2020.03.005
3. van Doremalen N, Bushmaker T, Morris DH, et al. Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1. *N Engl J Med*. 2020;382(16):1564-1567. Doi:10.1056/NEJMc2004973
4. Chu DK, Akl EA, Duda S, et al. Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis. *Lancet*. 2020;395(10242):1973-1987. Doi:10.1016/S0140-6736(20)31142-9
5. Kraemer MUG, Yang C-H, Gutierrez B, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science (80- )*. 2020;368(6490):493-497. Doi:10.1126/science.abb4218
6. Okell LC, Verity R, Watson OJ, et al. Have deaths from COVID-19 in Europe plateaued due to herd immunity? *Lancet*. 2020;395(10241):e110-e111. Doi:10.1016/s0140-

6736(20)31357-x

7. JHU. Coronavirus Resource Center. John Hopkins University. Published 2020. Accessed August 17, 2020. https://coronavirus.jhu.edu/
8. Dyer O. Covid-19: Trump stokes protests against social distancing measures. *BMJ*. 2020;369(April):m1596. Doi:10.1136/bmj.m1596
9. The Lancet. COVID-19 in Brazil: "So what?" *Lancet*. 2020;395(10235):1461. Doi:10.1016/S0140-6736(20)31095-3
10. Ibarra-Nava I, Cárdenas-de la Garza VER, Ruiz-Lozano RE, Salazar-Montalvo RG. Mexico and the COVID-19 Response. *Disaster Med Public Health Prep*. Published online July 27, 2020:1-5. Doi:10.1017/dmp.2020.260
11. Rodriguez-Morales AJ, Gallego V, Escalera-Antezana JP, et al. COVID-19 in Latin America: The implications of the first confirmed case in Brazil. *Travel Med Infect Dis*. 2020;35(January):101613. Doi:10.1016/j.tmaid.2020.101613
12. Aquino EML, Silveira IH, Pescarini JM, Aquino R, de Souza-Filho VER. Social distancing measures to control the COVID-19 pandemic: Potential impacts and challenges in Brazil. *Cienc e Saude Coletiva*. 2020;25:2423-2446. Doi:10.1590/1413-81232020256.1.10502020
13. Brasil. *Art. 18*. Constituição da República Federativa do Brasil de 1988; 1988.
14. Phillips D. Bolsonaro ignored by state governors amid anger at handling of Covid-19 crisis. The Guardian. Published April 1, 2020. Accessed July 8, 2020. https://www.theguardian.com/world/2020/apr/01/brazil-bolsonaro-ignored-by-state-governors-amid-anger-at-handling-of-covid-19-crisis
15. Brasil. *Medida Cautelar Na Ação Direta de Insconstitucionalidade 6.341 Distrito Federal. Relator: Min. Marco Aurélio.* Supremo Tribunal Federal; 2020.
16. Sobiech Pellegrini I. Untimely Reopening? Change in the Number of New COVID-19

Cases after Reopening in One Brazilian State. *SSRN Electron J*. Published online 2020:1-30. Doi:10.2139/ssrn.3623930

17. Silva LLS da, Lima AFR, Polli DA, et al. Medidas de distanciamento social para o enfrentamento da COVID-19 no Brasil: caracterização e análise epidemiológica por estado. *Cad Saude Publica*. 2020;36(9). Doi:10.1590/0102-311x00185020
18. Inloco. Mapa brasileiro da COVID-19. Inloco. Published 2020. Accessed July 5, 2020. https://www.inloco.com.br/covid-19
19. Túlio S, Martins V. Caiado recua sobre medidas mais rígidas e reclama de falta de apoio: "Não vale a pena fazer um decreto por fazer." G1 GO. Published May 14, 2020. Accessed July 2, 2020. https://g1.globo.com/go/goias/noticia/2020/05/14/governador-de-goias-recua-sobre-medidas-mais-rigidas-e-reclama-de-falta-de-apoio-nao-vale-a-pena-fazer-um-decreto-por-fazer.ghtml
20. Dehning J, Zierenberg J, Spitzner FP, et al. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science (80- )*. 2020;9789:eabb9789. Doi:10.1126/science.abb9789
21. Dehning J, Zierenberg J, Spitzner FP, et al. Bayesian inference and forecast of COVID-19. Published online 2020.
22. Hufnagel L, Brockmann D, Geisel T. Forecast and control of epidemics in a globalized world. *Proc Natl Acad Sci*. 2004;101(42):15124-15129. Doi:10.1073/pnas.0308344101
23. Bjørnstad ON, Finkenstädt BF, Grenfell BT. Dynamics of Measles Epidemics: Estimating Scaling of Transmission Rates Using a Time Series SIR Model. *Ecol Monogr*. 2002;72(2):169-184.
24. Atkeson AG. On Using SIR Models to Model Disease Scenarios for COVID-19. *Q Rev DC Nurses Assoc*. 2020;41(1). Doi:10.21034/qr.4111
25. Clancy D, O'Neill PD. Bayesian estimation of the basic reproduction number in

stochastic epidemic models. *Bayesian Anal*. 2008;3(4):737-757. Doi:10.1214/08-BA328

26. IBGE. Cidades e Estados. Goiás. Published 2019. Accessed June 20, 2020. https://www.ibge.gov.br/cidades-e-estados/go.html

27. SES-GO. Atualização dos casos de doença pelo coronavírus (Covid-19) em Goiás. Published 2020. Accessed June 20, 2020. http://covid19.saude.go.gov.br/

28. Goiás G do E de. Conheça os decretos e normas sobre o combate à pandemia do coronavírus. Published 2020. Accessed May 30, 2020. https://www.casacivil.go.gov.br/noticias/9033-legislação-sobre-o-coronavírus-covid-19.html

29. Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *J Travel Med*. 2020;27(2):1-4. Doi:10.1093/jtm/taaa021

30. Rangel TF, Diniz-Filho JAF, Toscano CM. *Nota Técnica 01. Modelagem Da Expansão Espaço-Temporal Da COVID-19 Em Goiás*.; 2020.

31. Barberia LG, Cantarelli LGR, Oliveira MLC de F, Moreira N de P, Rosa ISC. The effect of state-level social distancing policy stringency on mobility in the states of Brazil. *Ver Adm Pública*. 2021;55(1):27-49. Doi:10.1590/0034-761220200549

32. Diniz-Filho JAF, Jardim L, Toscano CM, Rangel TF. The effective reproductive number (Rt) of COVID-19 in Goias State and its relationship with social distancing. Published online 2020. Doi:10.1101/2020.07.28.20163493

33. Rangel TF, Diniz-Filho JAF, Toscano CM. *Nota Técnica 5. Avaliação Do Impacto de Medidas de Distanciamento Social Na Epidemia de COVID-19 Em Goiás Até 02/06/2020*. Universidade Federal de Goiás; 2020.

34. Seale H, Dyer CEF, Abdi I, et al. Improving the impact of non-pharmaceutical interventions during COVID-19: examining the factors that influence engagement and

the impact on individuals. *BMC Infect Dis*. 2020;20(1):607. Doi:10.1186/s12879-020-05340-9

35. Bavel JJ Van, Baicker K, Boggio PS, et al. Using social and behavioural science to support COVID-19 pandemic response. *Nat Hum Behav*. 2020;4(5):460-471. Doi:10.1038/s41562-020-0884-z

36. Lin C, Tu P, Beitsch LM. Confidence and Receptivity for COVID-19 Vaccines: A Rapid Systematic Review. *Vaccines*. 2021;9(1):16. Doi:10.3390/vaccines9010016

37. Domingues CMAS. Challenges for implementation of the COVID-19 vaccination campaign in Brazil. *Cad Saude Publica*. 2021;37(1). Doi:10.1590/0102-311x00344620

38. Mahase E. Covid-19: Where are we on vaccines and variants? *BMJ*. Published online March 2, 2021:n597. Doi:10.1136/bmj.n597

SUPPLEMENTARY MATERIAL

## Supplementary tables

**Table S1.** Model comparison using the leave-one-out (LOO) cross-validation method.

| Model | Change points | LOO-score | pLOO |
|---|---|---|---|
| SIR main | 0 | 598.9 ± 13.4 | 9.32 |
| SIR main | 1 | 597.2 ± 12.9 | 8.14 |
| SIR main | 2 | 595.5 ± 12.5 | 9.04 |
| *SIR main* | *3* | *592.3 ± 13* | *9.88* |
| SIR without weekend modulation | 3 | 603.6 ± 13.5 | 7.87 |
| SIR with wider delay prior | 3 | 601.8 ± 12.9 | 13.21 |
| SIR with wider change points prior | 3 | 698.0 ± 10.6 | 41.27 |
| SIR with wider transient length prior | 3 | 597.9 ± 13.6 | 12.09 |

We compared the full SIR model with other model variants, including SIR models with a different number of change points (Fig. S1-S2), without the weekend modulation (Fig. S3), and models from a sensitivity analysis with wider priors for the delay (Fig. 4), change points (Fig. S5) and change durations (Fig. S6). Lower LOO-score indicate better fit between model and data.

## Supplementary figures

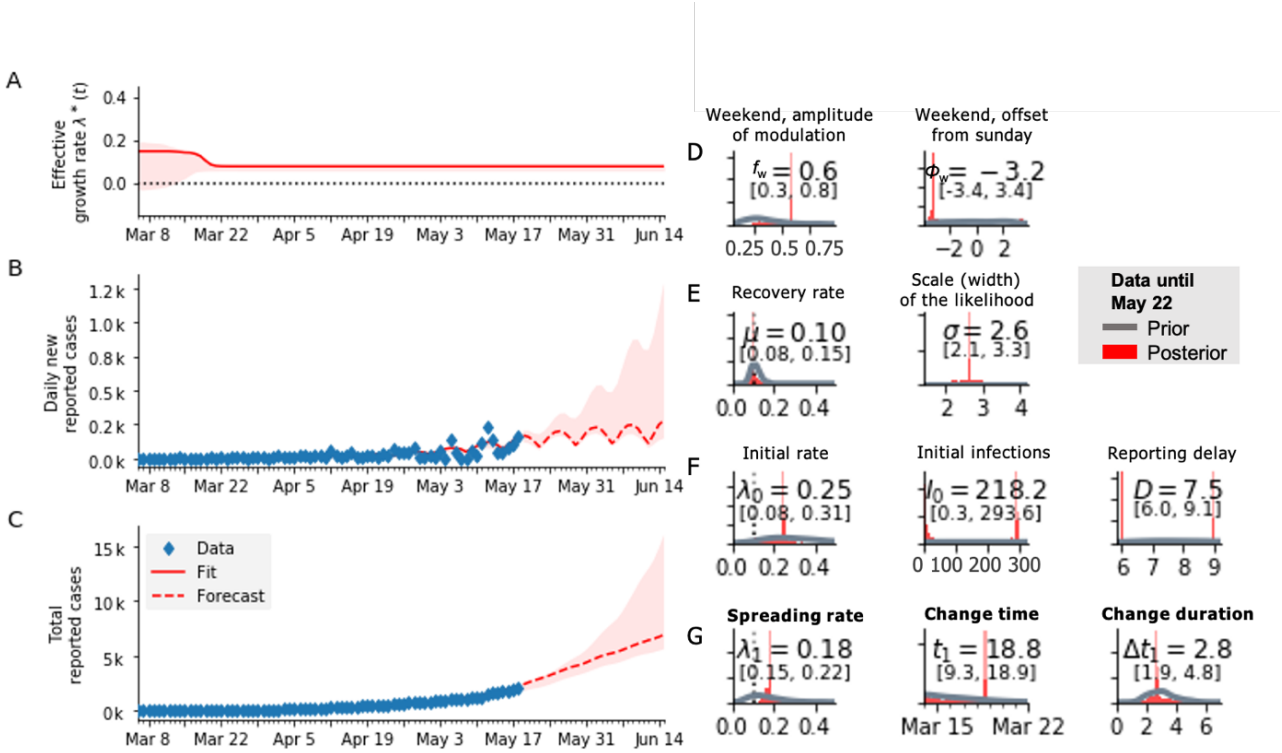


**Figure S1.** Full SIR model with **one change point** and weekly reporting modulation. All parameters and priors are the same as in the main text model (Fig.3), except number of change points. A: Estimate of the effective spreading rate; B: Daily new reported cases (blue diamonds) and the model (red solid line for median fit with 95% credible intervals). Red dashed line is the median forecast with 95% CI. C: Total reported cases and the model (color representation same as in B). D-F: Inference of central epidemiological parameters: prior (gray) and posterior distributions (red), inset values indicate the median and 95% CI of posteriors. G: Spreading rate, change time and change duration for the one change point, respectively.

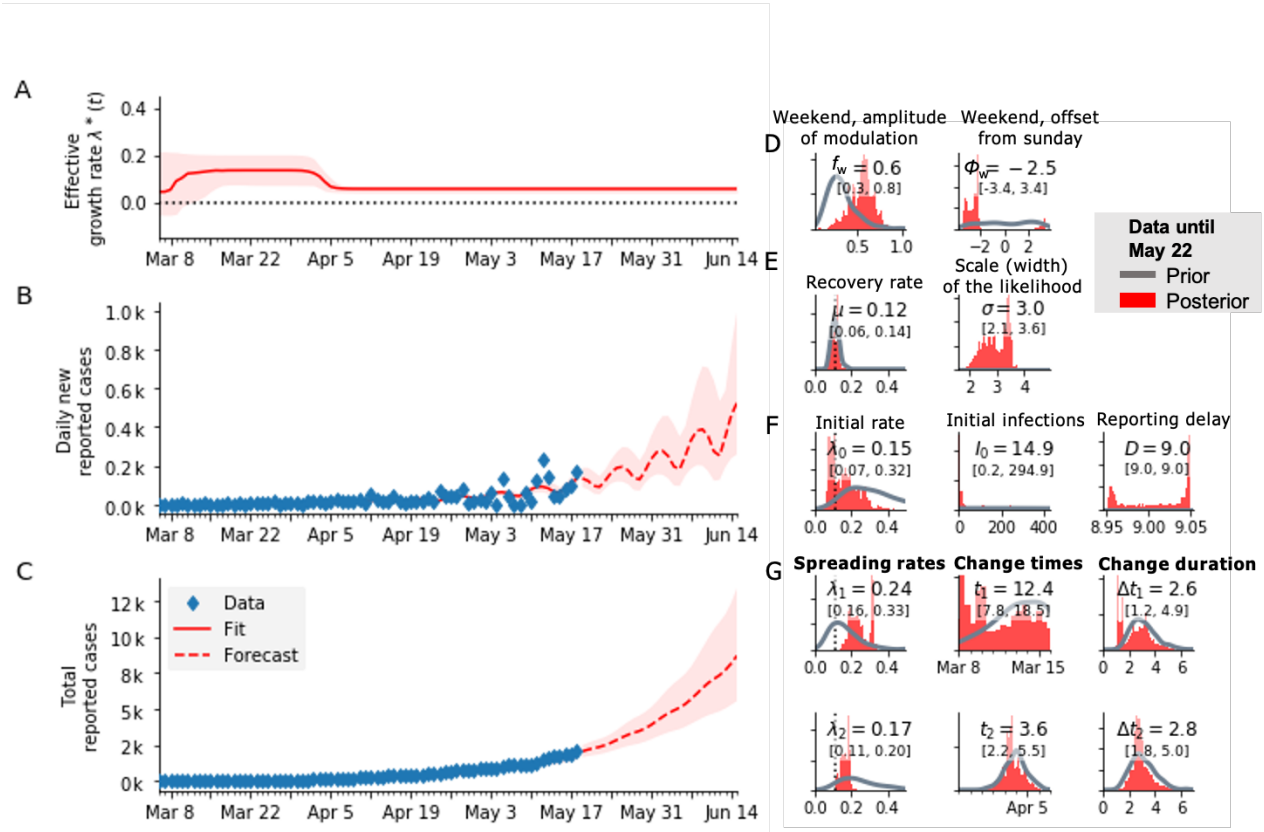


**Figure S2.** Full SIR model with **two change points** and weekly reporting modulation. All parameters and priors are the same as in the main text model (Fig.3), except number of change points. A: Estimate of the effective spreading rate; B: Daily new reported cases (blue diamonds) and the model (red solid line for median fit with 95% credible intervals). Red dashed line is the median forecast with 95% CI. C: Total reported cases and the model (color representation same as in B). D-F: Inference of central epidemiological parameters: prior (gray) and posterior distributions (red), inset values indicate the median and 95% CI of posteriors. G: Spreading rates, change times and change duration for the two change points, respectively.

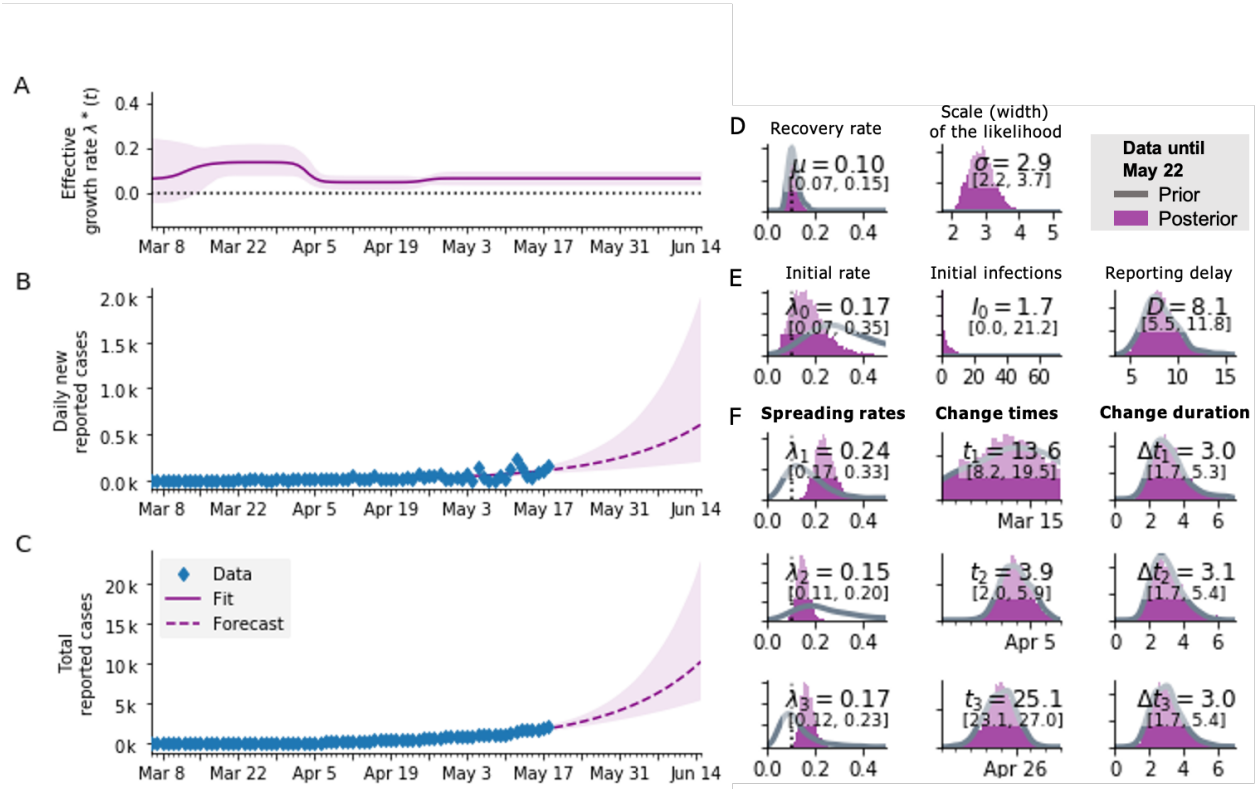


**Figure S3.** Full SIR model with three change points that **removes the weekly reporting modulation**. This model excludes the assumption that daily reported cases are lower during the weekend. All parameters and priors are the same as in the main text model (Fig.3). A: Estimate of the effective spreading rate; B: Daily new reported cases (blue diamonds) and the model (purple solid line for median fit with 95% credible intervals). Purple dashed line is the median forecast with 95% CI. C: Total reported cases and the model (color representation same as in B). D-F: Inference of central epidemiological parameters: prior (gray) and posterior distributions (purple), inset values indicate the median and 95% CI of posteriors. G: Spreading rates, change times and change duration for the three change points, respectively.

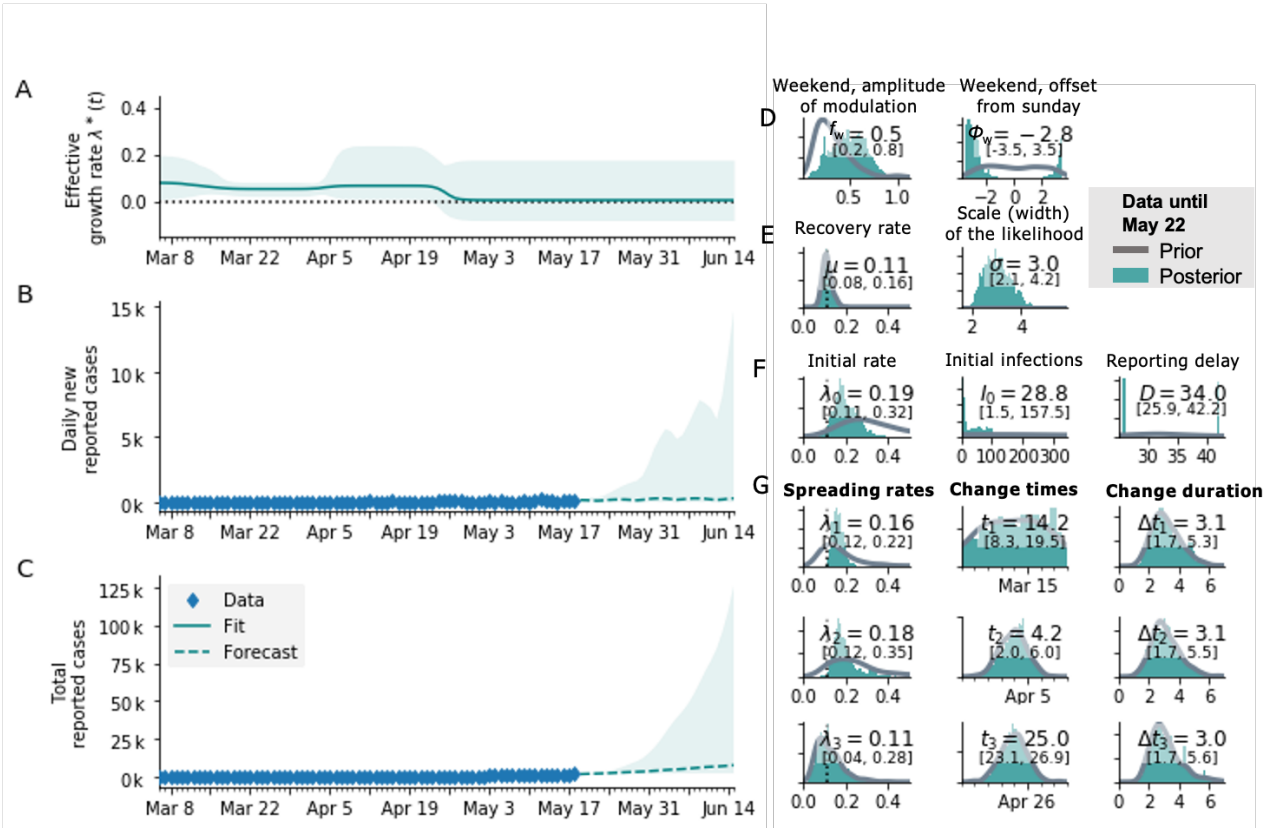


**Figure S4.** Sensitivity analysis. Full SIR model with three change points and weekly reporting modulation, but with a **wider prior for the reporting delay** (now 4 times wider). Other priors are the same as in the main text model (Fig.3). A: Estimate of the effective spreading rate; B: Daily new reported cases (blue diamonds) and the model (teal solid line for median fit with 95% credible intervals). Teal dashed line is the median forecast with 95% CI. C: Total reported cases and the model (color representation same as in B). D-F: Inference of central epidemiological parameters: prior (gray) and posterior distributions (teal), inset values indicate the median and 95% CI of posteriors. G: Spreading rates, change times and change duration for the three change points, respectively.

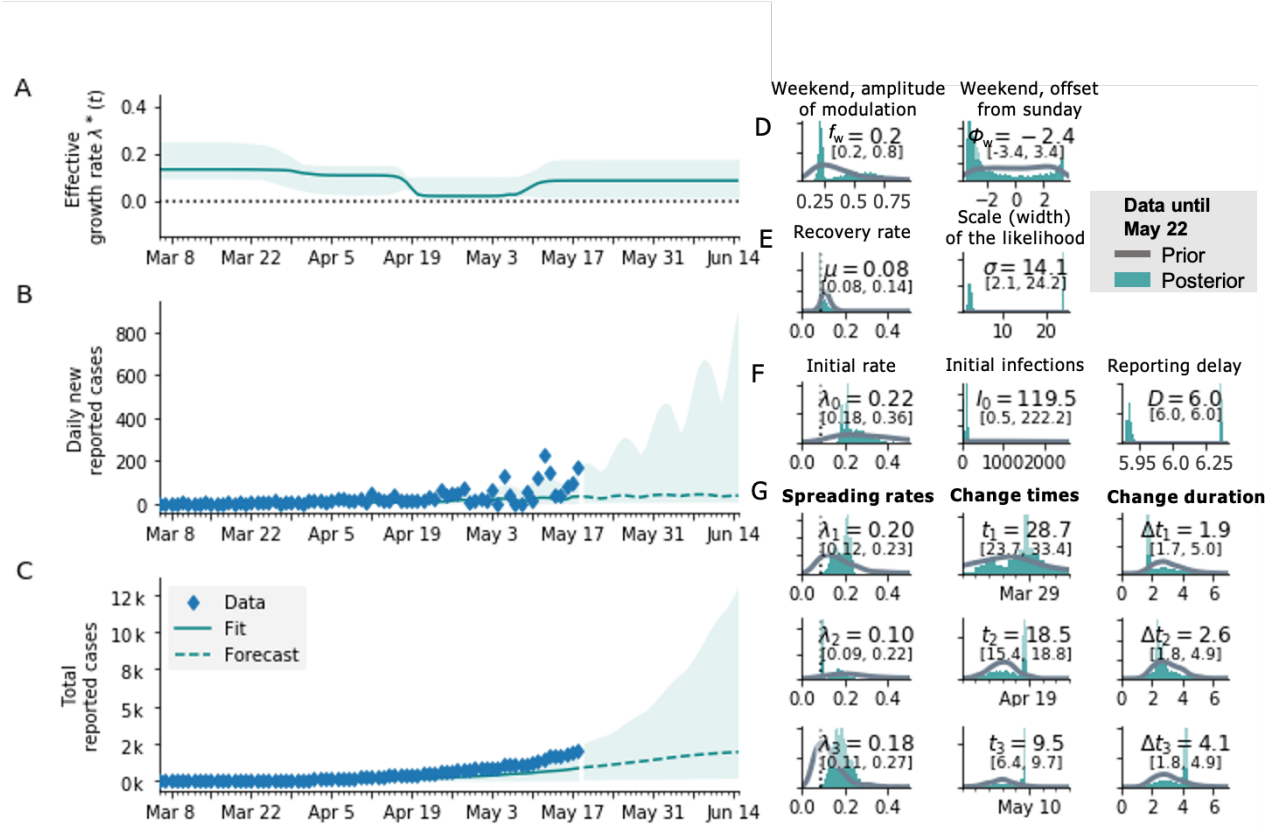


**Figure S5.** Sensitivity analysis. Full SIR model with three change points and weekly reporting modulation, but with a **wider prior for change points** (now 14 days instead of ~2 days). Other priors are the same as in the main text model (Fig.3). A: Estimate of the effective spreading rate; B: Daily new reported cases (blue diamonds) and the model (teal solid line for median fit with 95% credible intervals). Teal dashed line is the median forecast with 95% CI. C: Total reported cases and the model (color representation same as in B). D-F: Inference of central epidemiological parameters: prior (gray) and posterior distributions (teal), inset values indicate the median and 95% CI of posteriors. G: Spreading rates, change times and change duration for the three change points, respectively.

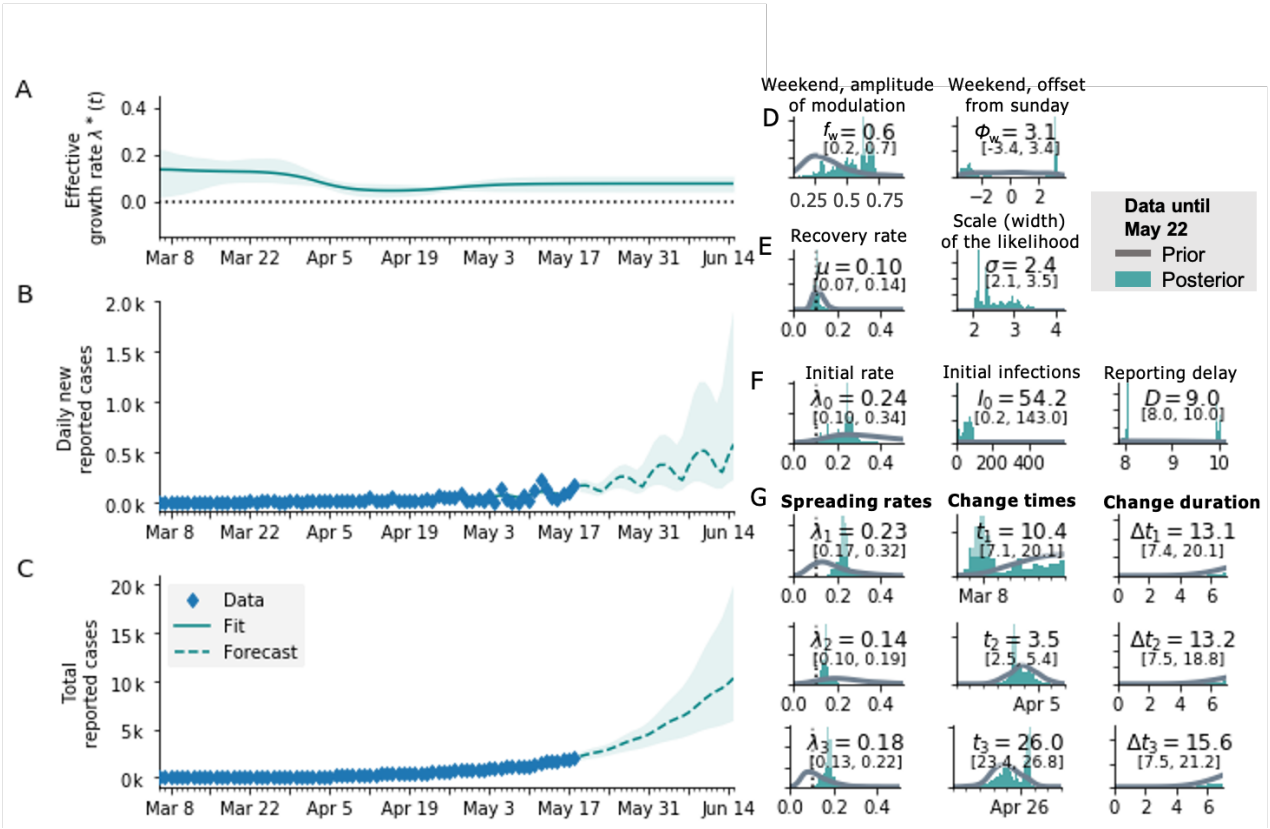


**Figure S6.** Sensitivity analysis. Full SIR model with three change points and weekly reporting modulation, but with a **wider prior for change durations** (now 4 times wider). Other priors are the same as in the main text model (Fig.3). A: Estimate of the effective spreading rate; B: Daily new reported cases (blue diamonds) and the model (teal solid line for median fit with 95% credible intervals). Teal dashed line is the median forecast with 95% CI. C: Total reported cases and the model (color representation same as in B). D-F: Inference of central epidemiological parameters: prior (gray) and posterior distributions (teal), inset values indicate the median and 95% CI of posteriors. G: Spreading rates, change times and change duration for the three change points, respectively.

# CONSIDERAÇÕES FINAIS

Nesta tese demonstramos como teorias e métodos desenvolvidos nas áreas de ecologia e evolução podem ser aplicados em áreas de humanidades, em especial na linguística e na saúde pública. Analisamos várias facetas da evolução de línguas, desde sua concepção teórica, à formação de dialetos dentro de uma "única língua", à geração dos padrões de diversidade de línguas em uma macroescala geográfica. Adicionalmente, analisamos como um vírus se propagou durante sua fase inicial de proliferação em uma população.

Além de dados oriundos de humanos, seja a língua que falam ou o vírus que os infectam, presentes em todos os quatro capítulos, outro fio condutor desta tese são principalmente as teorias de ecologia de populações. No capítulo 1 fica claro que a evolução de línguas em escala populacional é darwiniana. Em um nível individual, a língua se replica de forma memética, colonizando cérebros assim como um parasita infecta um corpo. No entanto, a língua se adapta ao ambiente físico, social e cultural, sendo considerada, portanto, um híbrido bio-cultural. Em um nível populacional, mecanismos macroevolutivos como dispersão, interação, especiação e extinção, claramente influenciam os padrões espaciais e temporais de diversidade linguística. Portanto, é necessário *scale thinking* para falar de evolução de línguas, pois os mecanismos em nível individual, populacional e comunitário sofrem pressões diferentes.

É um fato para a ciência linguística moderna que não existe uma língua no mundo que seja uniforme e homogênea (Bagno, 2015). A língua está viva, por isso ela está em constante evolução. Se olharmos para apenas um país, como o Brasil, encontraremos dentro de sua língua oficial, o Português, várias outras línguas distintas, que variam em todos níveis estruturais (i.e., fonologia, morfologia, sintaxe e léxico). Os dialetos do Brasil não foram formados de forma aleatória, eles são reflexo direto dos processos de migração

e dispersão de populações distintas pelo país. Portanto, os dialetos possuem estruturação histórica e geográfica, por isso, podem ser mapeados no tempo e espaço. Nosso teste dialetal tem justamente essa intenção: coletar a maior quantidade de *falares* possível dentro do território.

É interessante que mesmo com informação incompleta, encontramos estruturação espacial nos dados. Claro que descobrir uma divisão dialetal norte-sul para o Brasil é trivial do ponto de vista do conhecimento linguístico disponível. Sabe-se dessa divisão desde 1920, pelo menos. No entanto, a partir do momento em que se recupera um padrão espacial conhecido com métodos estáticos simples, como uma cluster espacial, se abrem precedentes para aplicação de métodos mais avançados, possibilitando que se descubram padrões desconhecidos pelos linguistas. É justamente avanços metodológicos como esses que permitem métodos mais ousados, como nosso modelo mecanístico estocástico e espacialmente explícito.

A engrenagem deste modelo, inclusive, é um cálculo de capacidade de suporte (K), um fator limitante ao crescimento populacional e conhecimento amplamente difundido na ecologia de populações. Usando métodos populacionais e computacionais, aliados à uma abordagem macroecológica, contribuímos para a discussão do porque existem tantas línguas e porque elas ocorrem com maior riqueza em algumas regiões e não em outas, demonstrando que mecanismos ambientais, topográficos e socioculturais são responsáveis por uma parte considerável dos padrões de diversidade para a região Neotropical.

Por fim, utilizamos um modelo epidemiológico, que combina métodos biológicos aos métodos sociais, para analisar a propagação do vírus SARS-CoV-2 pela população do estado de Goiás. É um capítulo que não fala de línguas, mas de vírus, que é outro organismo que também depende de humanos para se replicar e se propagar. Especificadamente, encontramos que as políticas públicas de distanciamento social e de quarentena tiveram

sucesso em conter o vírus em sua fase inicial de propagação, no entanto, conforme essas regras eram flexibilizadas e a população voltava a circular, as taxas de transmissão voltaram a aumentar. De forma muito trivial, podemos dizer que ao impedir o contato entre indivíduos, impede-se que o vírus consiga infectar novos hospedeiros, pois seu ciclo de propagação é interrompido. Infelizmente, esse conceito se provou bastante complexo para grande parte da população brasileira.

De maneira geral, demonstramos como os métodos e teorias desenvolvidos nas áreas biológicas podem ser aplicados e ajudar a avançar conhecimentos das áreas de humanidades, principalmente na linguística e na administração pública. O contrário é igualmente verdadeiro, e temos exemplos importantes de teorias linguísticas aplicada à problemas biológicos (Semple, Ferrer-i-Cancho, & Gustison, 2022).

Contribuímos também demonstrando a importância e eficácia de estudos multidisciplinares, principalmente para um objeto de estudo tão complexo quanto o *Homo sapiens*. Não se faz ciência sozinha e é desnecessário reinventar rodas, por isso, além de subir nos ombros de gigantes das nossas disciplinas, é igualmente essencial o diálogo com outras áreas do conhecimento, para subir nos ombros de gigantes deles.

## REFERÊNCIAS

Bagno, M. (2015). *Preconceito Linguístico* (56th ed.). São Paulo: Parábola Editorial.

Semple, S., Ferrer-i-Cancho, R., & Gustison, M. L. (2022). Linguistic laws in biology. *Trends in Ecology and Evolution*, 37(1), 53–66. https://doi.org/10.1016/j.tree.2021.08.012.