SHCHERBAKOVA, O.; MICHAELIS, S. M., HAYNIE, H. J.; PASSMORE, S.; GAST, V.; GRAY, R. D.; GREENHILL, S. J.; BLASI, D. E.; SKIRGÅRD, H. As sociedades de estranhos não falam línguas menos complexas. Tradução de Everton Gehlen Batista e revisão de Sara Luiza Hoff. *ReVEL*, v. 22, n. 42, 2024. [www.revel.inf.br].

As sociedades de estranhos não falam línguas menos complexas[1]

Societies of Strangers Do Not Speak Less Complex Languages

Olena Shcherbakova[2]

Susanne Maria Michaelis[3]

Hannah J. Haynie[4]

Sam Passmore[5]

Volker Gast[6]

Russell D. Gray[7]

Simon J. Greenhill[8]

Damián E. Blasi[9]

Hedvig Skirgård[10]

olena_shcherbakova@eva.mpg.de

[1] O texto original foi publicado na revista *Science Advances* (2023, v. 9, n. 33) com o nome "Societies of strangers do not speak less complex languages". O artigo recebeu uma errata e teve alguns dados corrigidos em 19 de janeiro de 2024, por isso a tradução se baseia nesta última versão (disponível em https://doi.org/10.1126/sciadv.adn6113). Para os Materiais suplementares em inglês, acesse https://www.science.org/doi/suppl/10.1126/sciadv.adf7704/suppl_file/sciadv.adf7704_sm.pdf. Os autores, gentilmente, autorizaram a publicação dessa tradução, pelo que os editores da revista agradecem.

[2] Departamento de Evolução Linguística e Cultural, Instituto Max Planck de Antropologia Evolutiva.
[3] Departamento de Evolução Linguística e Cultural, Instituto Max Planck de Antropologia Evolutiva.
[4] Departamento de Linguística, Universidade do Colorado Boulder.
[5] Iniciativa de Evolução da Diversidade Cultural, Escola de Cultura, História e Língua, Faculdade da Ásia e do Pacífico, Universidade Nacional Australiana.
[6] Departamento de Estudos Ingleses e Americanos, Universidade Friedrich-Schiller de Jena.
[7] Departamento de Evolução Linguística e Cultural, Instituto Max Planck de Antropologia Evolutiva, e Escola de Psicologia, Universidade de Auckland.
[8] Escola de Ciências Biológicas, Universidade de Auckland, e Departamento de Evolução Linguística e Cultural, Instituto Max Planck de Antropologia Evolutiva.
[9] Departamento de Biologia Evolutiva Humana, Museu Peabody, Universidade de Harvard; Departamento de Evolução Linguística e Cultural, Instituto Max Planck de Antropologia Evolutiva; e Arquivos da Área de Relações Humanas, Universidade de Yale.
[10] Departamento de Evolução Linguística e Cultural, Instituto Max Planck de Antropologia Evolutiva.

RESUMO: Diversas propostas recentes defendem que as línguas se adaptam ao ambiente. A hipótese do nicho linguístico alega que as línguas com um grande número de falantes nativos e uma proporção substancial de falantes não nativos (sociedades de estranhos) tendem a perder as distinções gramaticais. Em contrapartida, as línguas de comunidades pequenas e isoladas deveriam manter ou expandir os marcadores gramaticais. Neste artigo, nós testamos essas alegações com o uso de uma base de dados global de estruturas gramaticais, o Grambank. Modelamos o impacto do número de falantes nativos, a proporção de falantes não nativos, o número de línguas vizinhas e o status de uma língua em relação à complexidade gramatical com um controle de autocorrelação espacial e filogenética. Separamos a "complexidade gramatical" em duas dimensões: a quantidade de morfologia de uma língua ("fusão") e a quantidade de informações obrigatórias codificadas na gramática ("informatividade"). Encontramos diversos exemplos de associações positivas moderadas ou fracas, mas nenhuma correlação inversa entre complexidade gramatical e fatores sociodemográficos. Nesse sentido, os nossos resultados levantam certas dúvidas sobre as alegações muito difundidas de que a complexidade gramatical é moldada pelo ambiente sociolinguístico.

ABSTRACT: Many recent proposals claim that languages adapt to their environments. The linguistic niche hypothesis claims that languages with numerous native speakers and substantial proportions of nonnative speakers (societies of strangers) tend to lose grammatical distinctions. In contrast, languages in small, isolated communities should maintain or expand their grammatical markers. Here, we test these claims using a global dataset of grammatical structures, Grambank. We model the impact of the number of native speakers, the proportion of nonnative speakers, the number of linguistic neighbors, and the status of a language on grammatical complexity while controlling for spatial and phylogenetic autocorrelation. We deconstruct "grammatical complexity" into two separate dimensions: how much morphology a language has ("fusion") and the amount of information obligatorily encoded in the grammar ("informativity"). We find several instances of weak or moderate positive associations but no inverse correlations between grammatical complexity and sociodemographic factors. Our findings cast doubt on the widespread claim that grammatical complexity is shaped by the sociolinguistic environment.

Introdução

As sociedades variam muito em relação ao tamanho, homogeneidade e grau de contato com outras sociedades. A variação dessas propriedades é capturada no contínuo entre dois extremos: as "sociedades de íntimos" (sociedades esotéricas) e as "sociedades de estranhos" (sociedades exotéricas) (Thurston 1987; Thurston 1989; Givón; Young 2002; Givón 2005). As sociedades de íntimos são compostas por grupos pequenos, unidos e homogêneos, em que os membros compartilham grandes quantidades de conhecimento sobre a vida em comunidade e não se envolvem muito com pessoas de fora dela (Thurston 1987; Thurston 1989; Wray; Grace 2007; Lupyan; Dale 2010; Kusters 2003; Kusters 2008). No outro extremo, encontramos as sociedades de estranhos: grupos grandes e heterogêneos com proporção significativa de pessoas de fora (pessoas que usam uma língua diferente ou, pelo menos, pessoas de fora da comunidade local), redes frouxas e, em consequência, menos quantidade de

informação comunitária compartilhada. Fazemos referência a esse contínuo que se estende das sociedades de íntimos às sociedades de estranhos como exotericidade. Dentro desse contínuo, as sociedades com baixa exotericidade são o protótipo das sociedades de íntimos (sociedades esotéricas) e aquelas com alta exotericidade correspondem às características das sociedades de estranhos.

Há hipóteses sobre a influência de diferentes graus de exotericidade nas sociedades no modo de comunicação entre indivíduos — o que resultaria, por fim, em efeitos observáveis nas estruturas gramaticais. Existem duas orientações principais que ligam a exotericidade à estrutura da língua. Primeiro, os membros de sociedades que são esotéricas e homogêneas raramente se comunicam com pessoas de fora e, por isso, as línguas nessas sociedades são aprendidas e usadas quase exclusivamente pelos seus membros. Alega-se que a falta de contato com falantes não nativos molda as línguas de modo a desenvolver e reter mais marcações gramaticais obrigatórias e explícitas (Thurston 1987; Wray; Grace 2007; Thurston 1992). Por exemplo, o tariano, uma língua aruaque ameaçada de extinção falada no Amazonas, tem um sistema de marcação evidencial: os verbos carregam uma informação gramatical que distingue entre situações em que o falante viu ou ouviu a ação a que se referem ou em que essa ação é inferida ou presumida a partir de uma informação de segunda ou terceira mão (Aĭkhenvald 2013). Sugere-se que esse traço gramatical ocorre em línguas com baixa exotericidade em vez daquelas com alta exotericidade (Trudgill 2011).

Segundo, propôs-se que o ambiente social de comunidades exotéricas com alta proporção de pessoas de fora e graus de contato com falantes não nativos de L2 (segunda língua) incita simplificações morfológicas nas línguas (Wray; Grace 2007; Lupyan; Dale 2010). Os falantes de L2 acham muito difícil de processar e produzir estruturas gramaticais de fusão fonológica, como sufixos de caso e marcadores de concordância verbal (Wray; Grace 2007; Lupyan; Dale 2010; Clahsen *et al.* 2010, Trudgill 2011; Dale; Lupyan 2012). Por isso, foi sugerido que essas línguas passam por um processo de simplificação, como a perda de categorias morfológicas e concordância. Por exemplo, desde o período do inglês antigo, o inglês perdeu a concordância do adjetivo em caso, número e gênero assim como as distinções de caso nominal, que foram associadas à adoção do inglês por falantes não nativos (Trudgill 2011). De modo parecido, foi proposto que os sistemas de gênero, outra característica que é mais típica de línguas com baixa exotericidade (Trudgill 2011), tendem a

diminuir em línguas que entram em contato com outras línguas, sobretudo aquelas sem gênero, ou seja, quando as sociedades que falam essas línguas se tornam mais exotéricas (Dahl 2004), ou mesmo a desaparecer, como em osseto e grego da Capadócia, em que a perda de gênero foi relacionada ao aprendizado de L2 e contato (Igartua 2019). Além dos falantes adultos de L2 que não conseguem aprender uma língua estrangeira com precisão, a simplificação (perda ou redução de marcadores de fusão fonológica) pode ser o resultado de falantes de L1 (primeira língua) ajustarem, consciente ou inconscientemente, suas falas às necessidades de pessoas de fora por meio da redução de marcadores gramaticais que representem dificuldades de aquisição (Wray; Grace 2007).

Uma quantidade significativa de atenção foi dada às relações entre língua e estrutura social, principalmente pelas lentes de comparações específicas de pequena escala. Uma série de estudos qualitativos analisou variedades próximas do quéchua (Kusters 2008), inglês (Trudgill 2011; Szmrecsanyi; Kortmann 2009) e alemão (Maitz; Németh 2014; Baechler 2014) assim como línguas tibeto-birmanesas (DeLancey 2015) e escandinavas (Kusters 2003). Esses estudos parecem corroborar que, entre variedades próximas, as línguas que foram mais expostas ao contato com falantes de L2 tendem a ter marcadores gramaticais menos opacos e menos irregulares nos domínios estudados.

No entanto, ainda não está claro até que ponto é possível generalizar esses resultados para além de um punhado de casos. Cada uma dessas pesquisas recorre a variáveis linguísticas e sociodemográficas diferentes (e às vezes idiossincráticas), que põem em questão a homogeneidade de causas e seus mecanismos subjacentes. Essa capacidade de comparação limitada foi abordada, em parte, por estudos comparativos que objetivaram avaliar essas hipóteses em escala global. Os pesquisadores Lupyan e Dale (2010) mostraram que aspectos diferentes da complexidade morfológica têm correlação inversa com o tamanho populacional (o número de falantes de L1), a distribuição geográfica e o número de línguas vizinhas, o que se tornou conhecido como a hipótese do nicho linguístico. Um estudo de acompanhamento (Bentz; Winter 2013) não encontrou correlação entre o número de casos em substantivos e o tamanho populacional, mas mostrou que esse traço linguístico tem correlação negativa com a proporção de falantes de L2. Além disso, outros estudos (Sinnemäki; Di Garbo 2018) revelam uma correlação negativa entre síntese verbal e ambas variáveis demográficas

(o número de falantes de L1 e a proporção de falantes de L2) dentro do mesmo modelo. Por fim, alguns estudos não encontraram uma relação entre a complexidade morfológica e a presença ou ausência de uma proporção significativa de falantes de L2 (Koplenig 2019). Contudo, ainda não está claro se qualquer uma dessas medidas de exotericidade são associadas de modo significativo com a estrutura linguística. As grandes amostras linguísticas desses estudos fizeram com que colocar as comunidades estudadas em um contínuo de exotericidade de forma confiável seja um desafio. As informações confiáveis em todos os critérios (homogeneidade da população, tamanho da comunidade, densidade da rede social, isolamento relativo etc.) que delineiam as distinções entre comunidades mais ou menos exotéricas não estavam disponíveis, então, em vez disso, variáveis sociodemográficas diferentes serviram como representantes da exotericidade.

Os resultados inconsistentes de estudos anteriores podem ter ocorrido por três razões. Primeira, a cobertura interlinguística desses estudos varia consideravelmente e, assim, traz a questão de quão representativas são essas amostras de diferenças globais na complexidade gramatical. O tamanho limitado de uma amostra frequentemente é decorrente de uma cobertura irregular de traços linguísticos na base de dados do *The World Atlas of Language Structures — WALS* [Atlas Mundial de Estruturas Linguísticas] (Dryer; Haspelmath 2013). O WALS compreende 2662 línguas, mas as informações disponíveis sobre mais de 50% dos traços estão disponíveis em apenas algumas centenas delas. Assim, a pesquisa de vários traços associados com a complexidade se torna impossível sem a diminuição da amostra ou uma situação de maior incerteza nos dados. Por exemplo, o problema da escassez de dados afeta a pontuação da complexidade morfológica apresentada por Lupyan e Dale (2010), que é calculada para todas as línguas com ao menos 3 de 28 traços. Ao mesmo tempo, as amostras linguísticas usadas para a maior parte das análises de traços gramaticais individuais na mesma pesquisa também são modestas: o valor mediano é de 218 línguas por traço. Segunda, os estudos anteriores envolvem fenômenos linguísticos bastante diferentes supostamente comparáveis somente pelas lentes do termo guarda-chuva “complexidade gramatical”. A complexidade gramatical tem muitas facetas: número de marcadores, irregularidade, obrigatoriedade, composicionalidade, redundância e dependência em formas de fusão fonológica em vez de formas independentes (Wray; Grace 2007; Lupyan; Dale 2010; McWhorter 2007;

Miestamo 2008). A natureza multifacetada da complexidade significa que uma língua é vista como mais complexa à medida que aumenta o número de casos gramaticais e determinantes, formas verbais irregulares, construções não composicionais, padrões de concordância e/ou marcadores de fusão fonológica que expressam funções diferentes. No entanto, diferentes mecanismos subjacentes podem ser os responsáveis pelas mudanças nessas dimensões distintas de complexidade em sociedades exotéricas. Por exemplo, quando a complexidade é vista em termos de composicionalidade, alega-se que as estruturas linguísticas se tornam mais composicionais (elas consistem de várias partes interpretáveis em vez de uma parte interpretável independente) em sociedades exotéricas, em que altas proporções de falantes de L2 se beneficiam de uma maior transparência (Wray; Grace 2007). Todas essas diferentes dimensões podem mudar em ritmos diferentes e sob pressões diferentes. Por isso, a combinação de várias dessas dimensões em uma métrica de complexidade gramatical (por exemplo, em Koplenig (2019) e Bentz *et al.* (2016)), pode não elucidar a relação entre estruturas gramaticais e as variáveis sociodemográficas escolhidas para refletir a exotericidade.

Para testar a associação entre estruturas gramaticais e exotericidade das sociedades relevantes, nós introduzimos métricas que quantificam duas dimensões distintas de complexidade gramatical que foram consideradas reduzidas em sociedades exotéricas: (i) o grau de marcadores gramaticais de fusão fonológica ("fusão") e (ii) o número de marcação gramatical obrigatória ("informatividade"). Nós coletamos os dados de traços gramaticais incluídos nas duas métricas do Grambank (Skirgård *et al.* 2023a; Skirgård *et al.* 2023b), uma grande base de dados global (cf. Tabela S1 para a lista de traços do Grambank de ambas as métricas).

A pontuação de fusão reflete o grau que as línguas dependem de marcadores vinculados à fonologia (por ex., prefixos e sufixos) em oposição aos marcadores fonológicos independentes. Enquanto os marcadores fonológicos independentes são independentes de outros morfemas em relação à tonicidade e forma, os marcadores de fusão fonológica dependem de outros morfemas nesse aspecto, o que faz a tarefa de aquisição de marcadores de fusão fonológica por aprendizes adultos de L2 ser mais difícil. As línguas com marcadores de fusão fonológica para as categorias verbais de tempo, aspecto e modo, as categorias substantivas e pronominais de caso, e outros traços, como concordância de gênero e número, posse, negação, vão pontuar mais alto

nessa métrica. Por exemplo, a língua de pontuação mais alta, o tariano (aruaque), tem marcação morfológica de plural explícita em substantivos; casos oblíquos e nucleares em substantivos e pronomes; marcação morfológica de modo explícita, distinções de aspecto, tempos presente, passado e futuro; passiva morfológica em verbos; e concordância de número e gênero em alvos diferentes, entre outros.

A métrica da informatividade explica a quantidade de distinções gramaticais explícitas e obrigatórias feitas pelas línguas. Os traços a seguir, nesse sentido, aumentam a informatividade: pronomes de distinção de formalidade, distinção de tempo remoto no passado e futuro, artigos definidos e indefinidos, e marcação de número em substantivos (singular, dual, plural, trial, paucal; plural associativo). As línguas vão pontuar mais alto na métrica da informatividade se a gramática tiver, por exemplo, demonstrativos diferentes usados para objetos visíveis e não visíveis, como a língua nenets da tundra (urálico), em que *tay°kuy° teda*, "aquela rena", se refere a uma rena visível, e *t'exa teda*, "aquela rena", seria usado quando a rena não está visível, por exemplo, se ela estivesse atrás de algo (Nikolaeva 2014). Não incluímos os traços nessa métrica se eles são distinções geralmente consideradas como universais, como negação (Zeijlstra 2017) e posse (Nichols *et al.* 2013). Não temos o conhecimento de uma língua que não faça a distinção entre uma oração afirmativa e negativa nem mesmo uma língua que não tenha nenhum padrão produtivo para a marcação de posse. Se incluíssemos esses traços, não haveria nada de significativo para ser dito sobre a variação na expressão de significado gramatical nas línguas do mundo.

Outro traço que contribui para a informatividade é a presença de distinções entre construções inclusivas e exclusivas em sistemas pronominais e dêixis verbais. Por exemplo, o maori (austronésio) diferencia dois significados potenciais para "nós": para enunciar uma frase como "Nós vamos caminhar", o falante de maori obrigatoriamente escolhe entre *tātou* se o interlocutor vai participar e *mātou* se o interlocutor não está incluído no "nós" e o falante vai caminhar com outras pessoas (Harlow 1996). Além de excluir da métrica os traços que são marcados em todas ou quase todas as línguas, considerando que o Grambank foi planejado para apreender os traços encontrados com frequência nas línguas do mundo, nossa métrica não contém traços extremamente raros que são marcados em apenas algumas poucas línguas. A métrica da informatividade não distingue se a informação é marcada ou não por um marcador de fusão.

A terceira razão em potencial para os resultados inconsistentes nas pesquisas anteriores é que elas não controlam totalmente a não independência espacial e filogenética (mas cf. Bentz *et al.* 2015), apesar de ambos os fatores poderem influenciar a interpretação dos resultados. Por exemplo, foi demonstrado que o poder preditivo da maioria das relações negativas entre o WALS e o tamanho populacional diminuiu após um embaralhamento das línguas dentro das famílias (Lupyan; Dale 2010). Em particular, o modo como os estudos anteriores fazem o controle dessas confusões acaba por implicar o agrupamento das línguas em grandes grupos com base na ancestralidade e localização, o que simplifica demais as suas relações. Exceto por Bentz *et al.* (2015) e Verkerk e Di Garbo (2022), os trabalhos anteriores tendem a tratar a inclusão das línguas na mesma família como um efeito aleatório. Entretanto, essa perspectiva ignora as relações entre as línguas dentro das mesmas famílias. Como alternativa, as pesquisas anteriores coletam amostras de línguas de famílias e localizações diferentes, mas essa amostragem nem sempre garante a independência dos dados (Eff 2004; Bromham 2017) e leva, invariavelmente, a amostras mais restritas, com uma perda de poder estatístico em seguida (Levinson; Gray 2012). De forma parecida, os efeitos aleatórios de áreas geográficas, como as macroáreas do Glotolog ou as 24 áreas do AUTOTYP (Bickel *et al.* 2022), são usadas no controle da não independência espacial. Com macroáreas grandes, todas as línguas que são faladas em continentes diferentes são agrupadas em conjunto, e o efeito diferenciador da distância entre duas línguas vizinhas e duas línguas em lados diferentes do continente é negligenciado. Os efeitos aleatórios de áreas detalhadas têm o mesmo problema que as distâncias individuais entre línguas dentro da mesma área, elas não informam a análise. Além disso, o contato entre línguas vizinhas se elas pertencem a duas áreas diferentes não é abordado. Por exemplo, ainda que o ucraniano e o polonês sejam vizinhos geográficos muito próximos, considera-se que pertencem a áreas diferentes (Ásia Interior e Europa) quando as áreas do AUTOTYP são modeladas como efeitos aleatórios.

Testagem da hipótese

Nós testamos a hipótese que as línguas em sociedades muito exotéricas têm (i) menos marcadores gramaticais de fusão fonológica (fusão) e (ii) menos marcadores

explícitos obrigatórios em geral (informatividade) comparados às línguas em sociedades pouco exotéricas. Nosso objetivo é superar limitações anteriores com (i) o uso de uma amostra diversa e abrangente das línguas ao redor do mundo que ultrapassem as amostras de estudos anteriores, (ii) a fundamentação das variáveis envolvidas na relação entre a exotericidade e a estrutura da língua e (iii) a construção de um modelo estatístico de última geração que considere a complexa dependência histórica entre as línguas.

Amostra das línguas e sociedades

Nossa amostra compreende 1291 línguas (cf. Figuras 1 e 2). A maioria dessas línguas pertencem às seguintes famílias linguísticas: austronésio (291), sino-tibetano (144), atlântico-congo (140), afro-asiático (57), austro-asiático (53) e indo-europeu (44). A maioria das línguas na nossa amostra são faladas na Oceania (212), no Sudeste Asiático (155), na savana africana (136) e no subcontinente indiano (106). As línguas indo-europeias não estão representadas em excesso na nossa amostra de escala ampla, o que é um problema em muitos estudos interlinguísticos.

Variáveis sociodemográficas de exotericidade

Nós examinamos se a variação da pontuação de fusão e informatividade é explicada pelos fatores sociodemográficos associados com sociedades mais exotéricas. Essa dimensão é complexa e até hoje ela tem resistido a uma quantificação simples; entretanto, existe um número de variáveis sociais e demográficas que estão disponíveis globalmente e correspondem a diferentes graus de exotericidade. Em geral, uma sociedade é considerada mais exotérica se as métricas a seguir são maiores ou se as variáveis binárias estão presentes:

1. número de falantes de L1 (cf. Lupyan e Dale (2010) e Sinnemäki e Di Garbo (2018));
2. veicularidade (como um indicador da proporção de falantes de L2) (cf. Bentz e Winter (2013), Sinnemäki e Di Garbo (2018) e Koplenig (2019));
3. número de línguas vizinhas (cf. Lupyan e Dale (2010));
4. *status* oficial (cf. Chen, 2023);

5. uso na educação.

As duas últimas variáveis relacionadas ao *status* de uma língua (oficial/não oficial; uso ou não na educação) são utilizadas raras vezes para a previsão de estruturas gramaticais. Nós as incluímos porque elas ajudam no entendimento de outro lado da exotericidade: as comunidades exotéricas têm maior probabilidade de usar as línguas que são reconhecidas como oficiais e de educação. Prevemos que essas duas variáveis, *status* oficial e uso na educação, têm um papel crucial na hipótese do nicho linguístico. Elas permitem que a forma escrita se torne mais elaborada enquanto a forma falada fica mais simples (Lupyan; Dale 2016) ou elas deveriam mitigar o efeito negativo hipotético do número de falantes de L1 na complexidade gramatical. Ambos fatores deveriam agir para tornar a língua dominante mais conservadora, evitando, assim, a perda de traços complexos e aumentando a fidelidade de transmissão. Além disso, estudos recentes (Bromham *et al.* 2022) indicaram que a língua de educação, em particular, é uma das causas principais da perda das línguas minoritárias, o que reforça a pressão seletiva para aprender a língua dominante.

A veicularidade é defendida como um indicador confiável se existe a expectativa de uma língua ter falantes de L2 (Koplenig 2019). Essa variável foi construída seguindo a abordagem descrita por Koplenig (2019), que é baseada na Escala Gradual de Ruptura Intergeracional Expandida (em inglês, EGIDS — Expanded Graded Intergenerational Disruption Scale) disponível no Ethnologue (Eberhard; Simons; Fennig 2021). A escala EGIDS reflete quanto uma língua está ameaçada de extinção; o nível 0 significa “Internacional” e o nível 10 representa “Extinta”. Se uma língua tem um nível alto na EGIDS, como 0 (Internacional), 1 (Nacional), 2 (Provincial) ou 3 (Comunicação mais ampla), a língua é considerada “veicular”, ou seja, espera-se que tenha falantes de L2. Em contrapartida, não é provável que línguas de nível 4 (Educacional) e acima (Em desenvolvimento, Vigorosa, Ameaçada, Em mudança, Moribunda, Quase extinta, Dormente e Extinta) sejam usadas por falantes de L2.

Todas as variáveis são modeladas isoladamente devido à alta probabilidade de multicolinearidade: é mais provável que uma língua com muitos falantes de L1 tenha mais línguas vizinhas e aja como língua oficial e língua de educação.

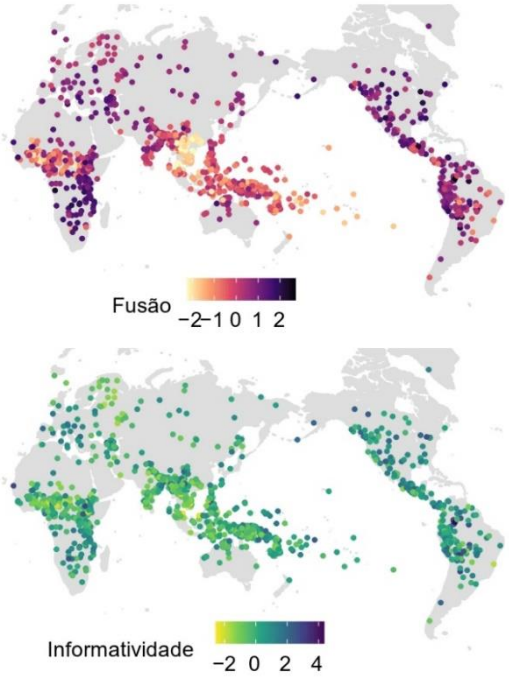

Figura 1: A distribuição global da pontuação de fusão e informatividade

Nota: as pontuações com um mínimo de 0 (ausência de todos os traços das métricas) e um máximo de 1 (presença de todos os traços das métricas) foram padronizadas para uma média de 0 e uma variância de 1. Os *hotspots* de baixa fusão estão localizados no Oeste Africano e no Sudeste Asiático. Muitas línguas austronésias também tiveram uma baixa classificação de fusão. Os padrões geográficos das pontuações da informatividade são menos claros comparados à fusão. Entre as línguas de baixa pontuação, estão aquelas faladas na África Ocidental e no Sudeste Asiático e diversas línguas urálicas e línguas faladas na Índia (indo-ariano e dravidiano).

Embora tenha sido sugerido que é mais provável que grandes populações tenham altas proporções de falantes de L2 (Lupyan; Dale 2010), é possível que as sociedades com tamanhos populacionais similares podem, ainda assim, ter diferentes proporções de falantes de L2, o que terá implicações diferentes na evolução dessas línguas se a ligação entre exotericidade e estruturas gramaticais for verdadeira. Por exemplo, uma sociedade com uma grande população de falantes de L1 e poucos falantes de L2 vai estar mais baixa na escala de exotericidade do que uma língua com tamanho populacional parecido, mas uma proporção extrema de falantes de L2. Em razão disso e da importância de levar em conta múltiplos fatores linguísticos e sociais (Sinnemäki; Di Garbo 2018; Sinnemäki 2020), ajustamos também dois modelos que usam o número de falantes de L1 e a veicularidade como (i) dois efeitos fixos separados e (ii) um termo de interação entre eles. Em uma subamostra de 120 línguas com dados

disponíveis no Ethnologue para o cálculo da proporção de falantes de L2, adequamos o mesmo grupo de modelos com a proporção de falantes de L2 em vez da veicularidade.

Modelagem filogenética espacial

Usando um enfoque bayesiano para filogenia, mapeamos as pontuações de fusão e informatividade obtidas pelo Grambank com informações disponíveis sobre os locais do Glottolog 4.5 (Hammarström *et al.* 2021) e as variáveis sociodemográficas para a árvore global (Bouckaert *et al.* 2022) das línguas distintas evolutivamente, ameaçadas de extinção globalmente, conhecidas como EDGE (do inglês *evolutionarily distinct, globally endangered*). Como resultado, temos uma amostra global de 1291 línguas disponíveis na filogenia em que ambas as pontuações das métricas foram calculadas e os dados sociodemográficos estavam presentes.

Adotamos uma modelagem filogenética espacial (Dinnage; Skeels; Cardillo 2020) que nos permite estudar a relação entre fatores sociodemográficos e linguísticos ao mesmo tempo que leva em conta as relações espaciais e genealógicas complexas entre línguas e sociedades. As duas relações, espaciais e genealógicas, estão representadas como efeitos aleatórios construídos com base nas matrizes de covariância que representam processos históricos relevantes.

Primeiro, ajustamos diferentes combinações de efeitos aleatórios para determinar se a distribuição das pontuações de fusão e informatividade depende de questões filogenéticas e geográficas das línguas. Nessa etapa, construímos sete modelos que contêm os efeitos aleatórios e de interceptação como preditores.

Efeitos filogenéticos:

1. efeitos espaciais: a difusão “local” de pontuações por muitas centenas de quilômetros é possível;
2. efeitos espaciais: a difusão “regional” de pontuações por distâncias de até 1000 km (cf. as seções Métodos e materiais e Efeitos espaciais e a Figura S1 para mais detalhes);
3. efeitos espaciais: 24 áreas linguísticas do AUTOTYP;
4. efeitos filogenéticos + efeitos espaciais: local;
5. efeitos filogenéticos + efeitos espaciais: regional;
6. efeitos espaciais: 24 áreas linguísticas do AUTOTYP.

Segundo, escolhemos o modelo mais forte para testar se a adição de variáveis sociodemográficas melhorarão o ajuste. Essa decisão implica o ajuste de sete modelos com diferentes variáveis sociodemográficas ou suas combinações tratadas como efeitos fixos:

1. número de falantes de L1;
2. veicularidade;
3. número de falantes de L1 e veicularidade (modelo combinado) (cf. Sinnemäki e Di Garbo (2018));
4. o termo de interação entre essas duas variáveis (número de falantes de L1 * veicularidade);
5. número de línguas vizinhas;
6. *status* oficial (binário);
7. língua de educação (binário).

Ajustamos esses sete modelos sem qualquer efeito aleatório para comparar em que medida o controle da não independência influencia os resultados. Então, comparamos os modelos de fusão e informatividade para determinar os preditores influentes das pontuações das métricas. Comparamos todos os modelos concorrentes em nossas análises, que são baseadas nos valores do critério de informação amplamente aplicável (WAIC, do inglês *widely applicable information criterion*) (Watanabe, 2010; Gelman; Hwang; Vehtari 2014).

Resultados

Do grupo de modelos com efeitos aleatórios, a previsão das pontuações de fusão e informatividade são melhores quando os efeitos filogenéticos e espaciais são combinados (Tabela 1 e 2). Os modelos filogenéticos espaciais que incorporam os dois efeitos aleatórios têm melhores resultados de modo significativo em relação a outros modelos, em específico, os modelos em segundo lugar, que são apenas filogenéticos. Esses resultados indicam que os dois efeitos explicam a variação nas pontuações melhor do que os efeitos filogenéticos isolados. As diferenças entre os valores do WAIC (Watanabe 2010), entre os modelos filogenéticos espaciais mais fortes e outros modelos são maiores que 45, tanto para fusão como para informatividade. A preferência pela versão local em oposição à regional de efeito espacial aleatório sugere

a provável difusão das pontuações em curtas distâncias de várias centenas de quilômetros. Nesse sentido, os efeitos aleatórios que admitem uma difusão possível muito mais ampla, ou seja, efeitos espaciais regionais (> 1000 km) ou os efeitos aleatórios de 24 áreas linguísticas, tiveram um pior desempenho.

Esse modelo mais adequado que incorpora os efeitos filogenéticos e espaciais é usado, então, para testar se a adição de qualquer um dos preditores sociodemográficos (ou combinações deles entre si) contribuíram para o entendimento da distribuição da fusão e da informatividade. Constatamos que os efeitos desses preditores variam de negligenciáveis a baixos. Os modelos mais fortes de previsão de fusão e informatividade são aqueles que incorporam esses efeitos aleatórios, o número de falantes de L1 e a veicularidade. Os modelos que incluem outras variáveis sociais são comparáveis com os modelos filogenéticos espaciais.

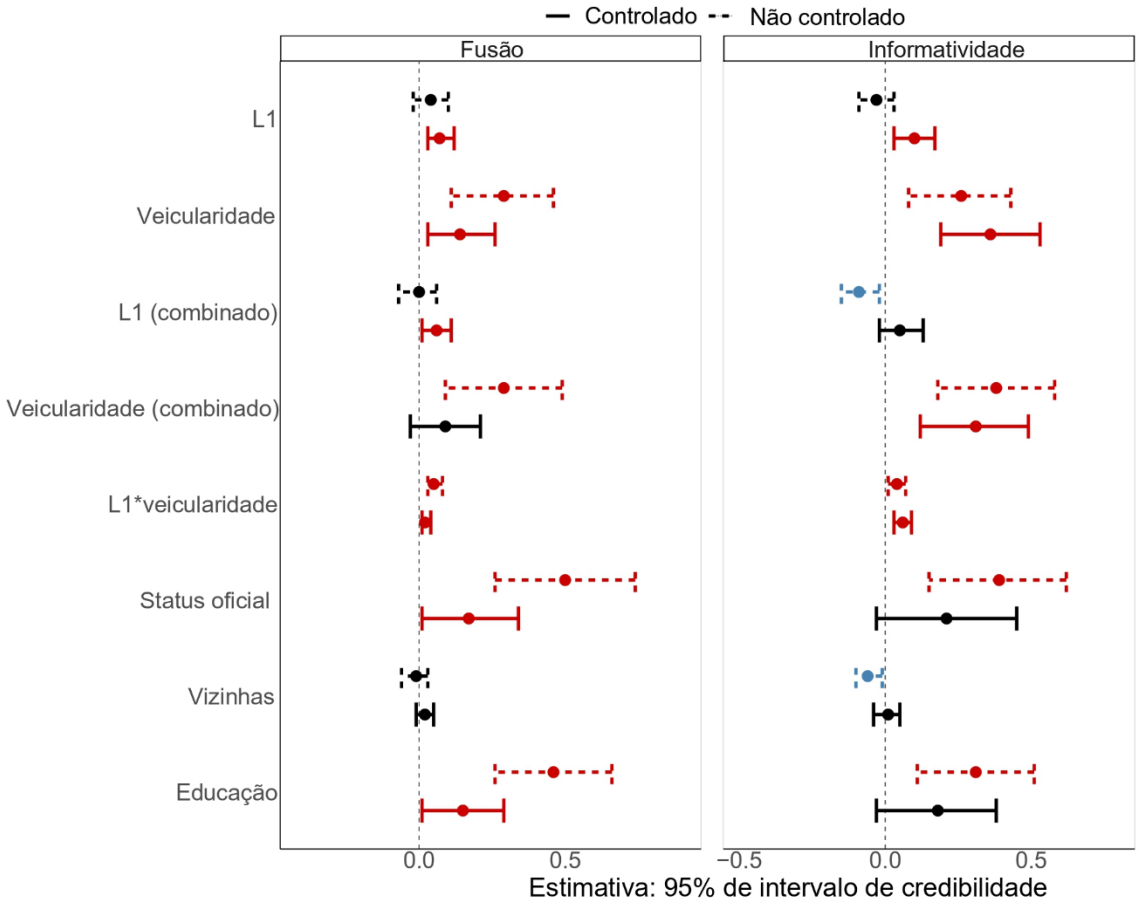

Figura 2: Os coeficientes e intervalos com 95% de credibilidade para efeitos fixos em seis modelos de regressão bivariados e um modelo multivariável (L1 combinado e veicularidade combinado) para fusão e informatividade, sem e com efeitos filogenéticos espaciais aleatórios (linhas tracejadas e contínuas, respectivamente).

Nota: os coeficientes de regressão linear de efeitos fixos que representam exotericidade nos modelos filogenéticos espaciais são representados com as barras de erro. As barras de erro em preto passam de zero, enquanto as barras vermelhas e azuis indicam relações robustas positivas e negativas, respectivamente. Muitos efeitos que parecem influentes (nas cores vermelha e azul) nos modelos que não apresentam um controle de efeitos aleatórios de genealogia e geografia (barras de erro tracejadas) desaparecem depois que fazemos o controle dessas fontes de não independência (barras de erro contínuas). As exceções são os efeitos positivos fracos dos falantes de L1 na fusão e informatividade que se revelam no modelo completo com efeitos fixos e aleatórios, mas permanecem escondidos no modelo em que o número de falantes de L1 é o único preditor das pontuações das métricas. Outras exceções são os efeitos positivos de veicularidade fracos ou moderados na fusão e informatividade, e de *status* oficial e uso na educação na fusão. Essa situação mostra que o controle da não independência das línguas é indispensável para revelar as relações de dependência entre estruturas gramaticais e fatores sociodemográficos propostas. Os nomes encurtados dos efeitos do número de falantes de L1 e da veicularidade diferem de acordo com o modelo em que estão incorporados: os modelos em que essas variáveis estão modeladas de forma isolada (efeitos de L1 (número de falantes de L1 que passaram por uma transformação logarítmica e padronização) e de veicularidade), o modelo combinado com esses dois efeitos — efeitos de L1 (combinado), (número de falantes de L1 que passaram por uma transformação logarítmica e padronização) e de veicularidade (combinado) — e o modelo com um termo de interação (L1*veicularidade), que usou o número de falantes de L1 que passou por uma transformação logarítmica.

No entanto, os coeficientes de regressão linear de muitos efeitos fixos são insignificantes: eles se sobrepõem ou não são suficientemente diferentes de zero (cf. Figura 2). Essa situação também se aplica aos resultados das subamostras de 120 línguas com dados disponíveis na proporção de falantes de L2 que não demonstram que essa variável é correlacionada com a fusão e a informatividade (cf. Tabela S2). Apenas algumas variáveis sociodemográficas testadas são preditoras fracas ou moderadas de pontuações das métricas. Para a fusão, constatamos uma correlação positiva fraca com a variável do número de falantes de L1 em dois modelos em que ela é (i) a única variável sociodemográfica e (ii) parte de uma combinação com a veicularidade (os intervalos de credibilidade de 95% para os efeitos de falantes de L1 na fusão em ambos os modelos: 0,03 a 0,12 e 0,01 a 0,11, respectivamente). De forma parecida, o *status* oficial e uso na educação demonstram uma correlação positiva fraca com a fusão quando são usados como os únicos preditores nos modelos (os intervalos de credibilidade de 95%: 0,01 a 0,34 e 0,01 a 0,29, respectivamente). Para a informatividade, encontramos um efeito positivo fraco do número de falantes de L1 (os intervalos de credibilidade de 95%: 0,03 a 0,17) apenas quando ela é a única variável sociodemográfica no modelo. Esse efeito desaparece (os intervalos de credibilidade de 95%: -0,02 a 0,13) quando tanto os números de falantes de L1 quanto a veicularidade são incluídos no mesmo modelo (os intervalos de credibilidade de 95% para os efeitos de veicularidade nesse modelo combinado para a informatividade: 0,12 a 0,49). Além disso, constatamos um efeito positivo fraco da interação entre o número de falantes de L1 e veicularidade tanto na fusão quanto na informatividade (os intervalos de credibilidade de 95%: 0,01 a 0,04 e 0,03 a 0,09, respectivamente). Nenhuma dessas relações é negativa como previsto em pesquisas anteriores. Esse resultado contradiz o argumento apresentado pela hipótese do nicho linguístico de que haveria uma relação inversa entre a complexidade gramatical e os fatores sociodemográficos associados com a exotericidade.

Modelo	Efeito	2,5%	50%	97,5%	WAIC
Filogenético + espacial: local + falantes de L1	**DP filogenético**	1,53	1,75	2,02	1796,34
	DP espacial	0,28	0,34	0,41	
	Interceptação	–0,02	0,00	0,02	
	L1	0,03	0,07	0,12	
Filogenético + espacial: local + falantes de L1 + veicularidade	**DP filogenético**	1,53	1,75	2,02	1798,71
	DP espacial	0,27	0,34	0,40	
	Interceptação	–0,04	–0,01	0,02	
	L1	0,01	0,06	0,11	
	Veicularidade	–0,03	0,09	0,21	
Filogenético + espacial: local + L1_log10:veicularidade	**DP filogenético**	1,52	1,74	2,01	1814,65
	DP espacial	0,27	0,34	0,40	
	Interceptação	–0,04	–0,01	0,01	
	L1*veicularidade	0,01	0,02	0,04	
Filogenético + espacial: local + veicularidade	**DP filogenético**	1,52	1,74	2,01	1814,82
	DP espacial	0,27	0,34	0,40	
	Interceptação	–0,04	–0,01	0,01	
	Veicularidade	0,03	0,14	0,26	
Filogenético + espacial: local + vizinhos	**DP filogenético**	1,52	1,74	2,01	1818,78
	DP espacial	0,27	0,34	0,41	
	Interceptação	–0,03	0,00	0,02	
	Vizinhos	–0,01	0,02	0,05	
Filogenético + espacial: local + oficial	**DP filogenético**	1,50	1,72	1,99	1820,45
	DP espacial	0,28	0,34	0,41	
	Interceptação	–0,03	–0,01	0,02	
	Status oficial	0,01	0,17	0,34	
Filogenético + espacial: local + educação	**DP filogenético**	1,50	1,72	1,99	1821,36
	DP espacial	0,28	0,34	0,41	
	Interceptação	–0,04	–0,01	0,01	
	Educação	–0,01	0,15	0,29	

Tabela 1: Os valores do WAIC e os quantis (0,025, 0,5 e 0,975) das estimativas nos modelos ajustados apenas com efeitos aleatórios e a interceptação nos modelos preditores de fusão

Nota: o texto em negrito indica que os efeitos são substanciais (não incluem zero).

Todos os modelos com efeitos fixos que excluem os efeitos aleatórios das relações filogenéticas e espaciais pontuam mais baixo do que os mesmos modelos que também implementam os efeitos aleatórios (cf. Tabela S2). Isso significa que o desempenho preditivo dos modelos, sem os efeitos aleatórios, é inferior em comparação com os modelos que incorporam os dois efeitos, fixos e aleatórios.

Em vez disso, a distribuição da pontuação de fusão e informatividade é explicada em grande parte por efeitos filogenéticos aleatórios — 92 e 68% da variância total de fusão e informatividade — e efeitos espaciais aleatórios que representam 4 e 9% da variância total das pontuações correspondentes (cf. Tabela S3). Medimos o sinal filogenético de ambas as dimensões de complexidade na árvore global usando a estimativa do *lambda* de Pagel (λ) (Pagel 1999). Os valores dessa métrica variam de 0, que indica nenhum sinal filogenético (distribuição aleatória de pontuações em relação à filogenia), a 1, que indica um sinal filogenético forte (línguas com relação próxima

compartilham pontuações parecidas), ou maiores que 1. Tanto fusão como informatividade apresentam sinais filogenéticos fortes: o sinal filogenético da fusão ($\lambda = 0,97, < 0,001$) é mais forte que o da informatividade ($\lambda = 0,86 < 0,001$). Em outras palavras, as pontuações de fusão e informatividade são explicadas pela herança de um ancestral comum e pela difusão espacial entre vizinhos próximos. Contudo, o papel das relações espaciais pode ter sido subestimado, considerando que a informação das localizações das línguas informou a estrutura da árvore global EDGE, em que os antecedentes geográficos fracos foram impostos nas relações linguísticas prováveis dentro do modelo filogeográfico. Nesse sentido, a contribuição relativa do preditor espacial pode ser maior.

Uma explicação alternativa para não encontramos um efeito de exotericidade considerável na fusão e informatividade é que essas relações não são lineares. Talvez seja o que ocorre, por exemplo, se os efeitos do número de falantes de L1 em estruturas gramaticais somente são significativos para comunidades extremamente grandes ou extremamente pequenas. Para abordar essa possibilidade, além de adequar regressões lineares com o número de falantes de L1, operacionalizamos essa variável na forma de efeitos não lineares dentro dos modelos de passeio aleatório de ordem 2 (RW2, do inglês *random walk models of order 2*). Descobrimos que esses modelos pontuam de modo parecido aos seus homólogos com efeitos lineares correspondentes. Uma vez que o desvio padrão (DP) desses efeitos aleatórios não lineares é pequeno (< 0,04 para fusão e informatividade), relatamos apenas os resultados dos modelos de regressão linear no texto principal e apresentamos a Tabela S2, que resume os valores do WAIC e os efeitos de todos os modelos com adequação nos Materiais suplementares.

Modelo	Efeito	2,5%	50%	97,5%	WAIC
Filogenético + espacial: local + falantes de L1 + veicularidade	**DP filogenético**	0,92	1,24	1,60	3128,91
	DP espacial	0,36	0,44	0,53	
	Interceptação	−0,04	0,00	0,04	
	L1	−0,02	0,05	0,13	
	Veicularidade	0,12	0,31	0,49	
Filogenético + espacial: local + falantes de L1	**DP filogenético**	0,97	1,28	1,66	3132,76
	DP espacial	0,35	0,43	0,53	
	Interceptação	−0,01	0,03	0,07	
	L1	0,03	0,10	0,17	
Filogenético + espacial: local + L1_log10:veicularidade	**DP filogenético**	0,89	1,20	1,57	3135,58
	DP espacial	0,36	0,44	0,54	
	Interceptação	−0,05	0,00	0,04	
	L1*veicularidade	0,03	0,06	0,09	
Filogenético + espacial: local + veicularidade	**DP filogenético**	0,87	1,18	1,55	3136,39
	DP espacial	0,36	0,44	0,54	
	Interceptação	−0,05	0,00	0,04	
	Veicularidade	−0,19	0,36	0,53	
Filogenético + espacial: local + vizinhos	**DP filogenético**	0,88	1,20	1,57	3153,06
	DP espacial	0,36	0,44	0,54	
	Interceptação	−0,01	0,03	0,07	
	Vizinhos	−0,04	0,01	0,05	
Filogenético + espacial: local + educação	**DP filogenético**	0,85	1,15	1,54	3156,96
	DP espacial	0,37	0,45	0,54	
	Interceptação	−0,02	0,02	0,06	
	Educação	−0,03	0,18	0,38	
Filogenético + espacial: local + oficial	**DP filogenético**	0,83	1,14	1,52	3157,39
	DP espacial	0,37	0,45	0,54	
	Interceptação	−0,02	0,02	0,06	
	Status oficial	−0,03	0,21	0,45	

Tabela 2: Os valores do WAIC e os quantis (0,025, 0,5 e 0,975) das estimativas nos modelos ajustados apenas com efeitos aleatórios e a interceptação nos modelos preditores de informatividade

Nota: o texto em negrito indica que os efeitos são substanciais (não incluem zero).

Discussão

A alegação específica da "hipótese do nicho linguístico" de que a complexidade gramatical se reduziria com o aumento do número de falantes não nativos não é amparada por nossos resultados. Ao contrário da correlação inversa esperada entre as pontuações de complexidade e as variáveis sociodemográficas que refletem exotericidade, os únicos efeitos que encontramos são positivos, fracos ou moderados.

Em vez disso, verificamos que a previsão das duas dimensões de complexidade que modelamos — fusão e informatividade — é melhor de acordo com a genealogia e a difusão geográfica do que com medidas de exotericidade. A medição do sinal filogenético desses dois traços também demonstrou que a distribuição das pontuações foi muito influenciada pelas histórias evolucionárias compartilhas entre as línguas da árvore global. Essas duas dimensões de complexidade gramatical parecem ter uma alta

estabilidade filogenética (cf. Figura 3), o que sugere uma restrição filogenética da fusão e da informatividade.

Um olhar mais atento para as línguas faladas na África Meridional mostra o motivo das distâncias filogenéticas explicarem a variação na pontuação de fusão melhor do que as características demográficas das sociedades (cf. Figura 4A). A maioria das línguas nessa área tem uma pontuação de fusão alta, inclusive a língua com maior pontuação, o soto do sul (1,38): uma língua banta do sul que é usada em uma sociedade com alta exotericidade por > 5,5 milhões de falantes de L1 e > 7 milhões de falantes de L2. Em contraste, o tsonga, outra língua banta do sul, tem pontuação baixa de modo considerável: -0,6. Ela também é falada em um nicho exotérico, mas o grau de exotericidade é mais baixo, considerando que ela tem menos falantes de L2 (> 3 milhões) do que o soto do sul, apesar de ter aproximadamente 1 milhão a mais de falantes de L1. Devido aos ambientes sociolinguísticos, seria possível esperar uma maior semelhança nas pontuações entre essas duas línguas e mais fusão em tsonga. Entretanto, observamos uma diferença menos pronunciada entre a pontuação do tsonga e aquela da sua língua irmã, o xítsua, com uma pontuação média superior (0,7), falada em um nicho mais exotérico por > 6 milhões de falantes de L1 e nenhum falante de L2. Algumas outras línguas com pontuações de fusão muito baixas da África Meridional são línguas de outras famílias linguísticas: taa oriental (família tuu) (-1,43) com 2500 falantes de L1 e o amkoe (família kx'a) (-0,94), língua ameaçada gravemente de extinção, com várias dezenas de falantes de L1. Apesar de serem faladas em sociedades bastante esotéricas, essas línguas têm pontuação de fusão extremamente baixas. Das línguas da África Meridional que foram analisadas, as línguas com menor pontuação (ou com pontuação moderada, como no caso do xítsua) são duas línguas banto próximas ou línguas de famílias linguísticas que não o banto, o que indica que as relações genealógicas e, em alguma medida, o contato entre línguas vizinhas fornece uma base melhor para o entendimento da variação na pontuação de fusão.

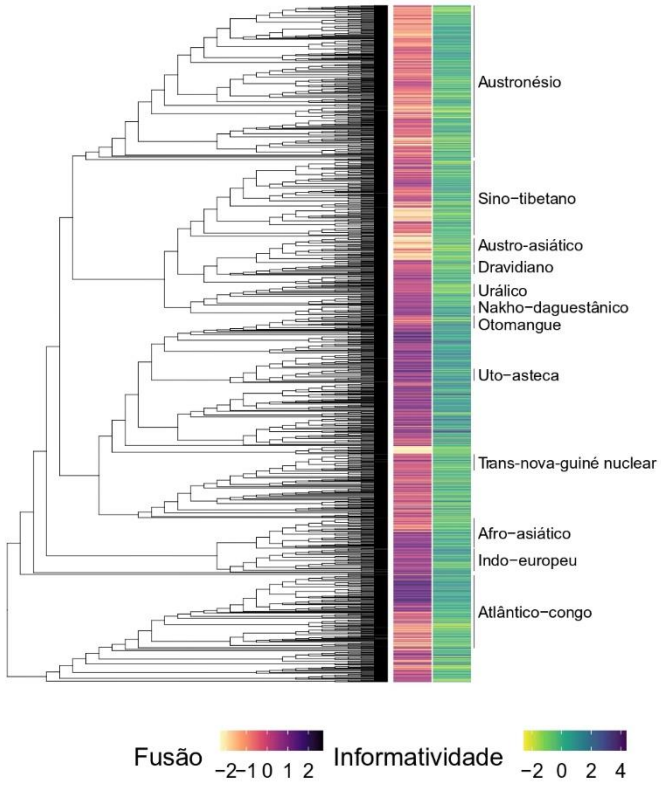

Figura 3: As pontuações de fusão e informatividade na árvore global

Nota: as pontuações com um mínimo de 0 (ausência de todos os traços das métricas) e máximo de 1 (presença de todos os traços das métricas) foram padronizadas com uma média de 0 e variância de 1. Detectamos muitos padrões de línguas próximas com pontuação parecida, o que pode indicar a transmissão confiável da complexidade gramatical de línguas ancestrais às descendentes em vez das adaptações em larga escala da complexidade gramatical em mudanças de fatores sociodemográficos. De forma parecida com a distribuição geográfica, percebemos que as pontuações de fusão seguem um padrão mais definido de agrupamento filogenético em comparação às pontuações de informatividade.

A família urálica fornece o exemplo de um caso em que pode ser desafiador separar os efeitos filogenéticos dos efeitos espaciais, porque a distribuição da pontuação de informatividade pode ser explicada tanto por relações filogenéticas quanto por distâncias geográficas/fenômenos de contato. A maioria das línguas urálicas tem pontuação baixa na métrica de informatividade (cf. Figura 4, B e C). As exceções, com pontuações maiores, são as línguas que pertencem aos ramos como o samoiedo (nganasan, selkup, nenets da tundra e enets da floresta), úgrico (húngaro) e permiano (komi-permyak e komi-zyrian), que se afastaram antes do resto das línguas organizadas nos ramos mari, mordóvico, sámi e fínico (Watanabe, 2010). Os contrastes nas pontuações de informatividade entre essas línguas irmãs não parecem condizer com a hipótese de que uma menor informatividade deveria existir em sociedades

exotéricas. Por exemplo, o vótico, com 25 falantes, tem pontuação mais baixa em informatividade (-1,79) do que o estoniano (-0,8), seu parente mais próximo, com > 1 milhão de falantes. Ainda que as distâncias filogenéticas sirvam como uma explicação para a distribuição das pontuações das famílias, a proximidade geográfica/contato pode ter sido outro fator influente. Por exemplo, o húngaro condiz com as altas pontuações das línguas indo-europeias das proximidades. De forma parecida, o nenets da tundra e o komi-permyak podem ter pontuações altas de informatividade por causa do bilinguismo dos falantes na Rússia. Ao contrário das altas proporções de falantes adultos de L2, o bilinguismo infantil é associado com uma alta na complexidade gramatical em vez de sua perda (Trudgill 2011).

Uma das diferenças chaves entre a análise apresentada aqui e em outros trabalhos é que nós usamos métodos filogenéticos espaciais para modelar explicitamente os efeitos da não independência genealógica e geográfica. Assim, conseguimos não só abordar duas fontes de não independência entre línguas em nossas amostras interlinguísticas (Bromham *et al.* 2022; Mace *et al.* 1994), mas também quantificar e comparar a importância relativa dos efeitos filogenéticos e espaciais. A maioria dos estudos anteriores ignoram a dependência entre línguas pertencentes à mesma família ou localizadas na mesma área ao tratar as famílias e as áreas como efeitos aleatórios ou usar amostras de línguas vindas de locais e famílias diferentes em uma tentativa de excluir as línguas que são não independentes (Eff 2004; Bromham 2017). Nossos resultados mostram com clareza que essa diferença na metodologia é realmente importante (Figura 2). Apenas algumas variáveis, como o número de falantes de L1 e a veicularidade, junto ao *status* oficial e uso na educação, para fusão, fazem aparecer efeitos positivos fracos ou moderados depois da aplicação do controle para genealogia e geografia, ao passo que outros efeitos parecem influentes somente em modelos sem esses efeitos aleatórios e desaparecem quando a não independência das línguas passa a ser controlada.

Para examinar as diferenças entre os nossos resultados e os achados descritos em trabalhos anteriores, reanalisamos os dados de complexidade morfológica usados por Lupyan e Dale (2010)[11]. Modelamos a relação entre as pontuações da complexidade

[11] Os dados foram conseguidos através de G. Lupyan por comunicação pessoal em 2 de junho de 2023 e disponibilizados no repositório digital Zenodo (https://doi.org/10.5281/zenodo.10420654), na pasta de dados nomeada complexity_data_WALS.csv.

morfológica e o número de falantes de L1. Os resultados revelaram uma correlação negativa, mas extremamente fraca, entre a complexidade morfológica e o número de falantes de L1 no modelo com efeitos filogenéticos espaciais (-0,02). Nenhuma correlação foi encontrada no modelo sem esses efeitos. Para entender se esses achados das reanálises são sensíveis ao critério de corte para a cobertura mínima dos traços — as pontuações de complexidade morfológica em Lupyan e Dale (2010) são calculados se pelo menos 10% de traços da métrica estão disponíveis no WALS para uma língua —, aumentamos o corte ao mínimo de 35% de traços disponíveis. Não foi possível achar qualquer efeito do número de falantes de L1 em complexidade morfológica com o aumento do corte enquanto controlamos os efeitos filogenéticos espaciais (cf. Tabela S4). Com esses resultados, esperamos que, na modelagem da pontuação de fusão (calculada com o uso de dados de outra base de dados) junto ao número de falantes de L1 como o único preditor (sem o controle da genealogia e geografia), não encontramos evidência para a correlação entre essas variáveis. Isso sugere que a diferença fundamental entre os nossos resultados e os achados de pesquisas anteriores está provavelmente nos dados usados para o cálculo das pontuações de complexidade gramatical e a proporção das línguas mal descritas na amostra. Os resultados anteriores talvez sejam o fruto da escassez de dados no WALS e a aplicação de um critério de corte baixo para a cobertura mínima dos traços.

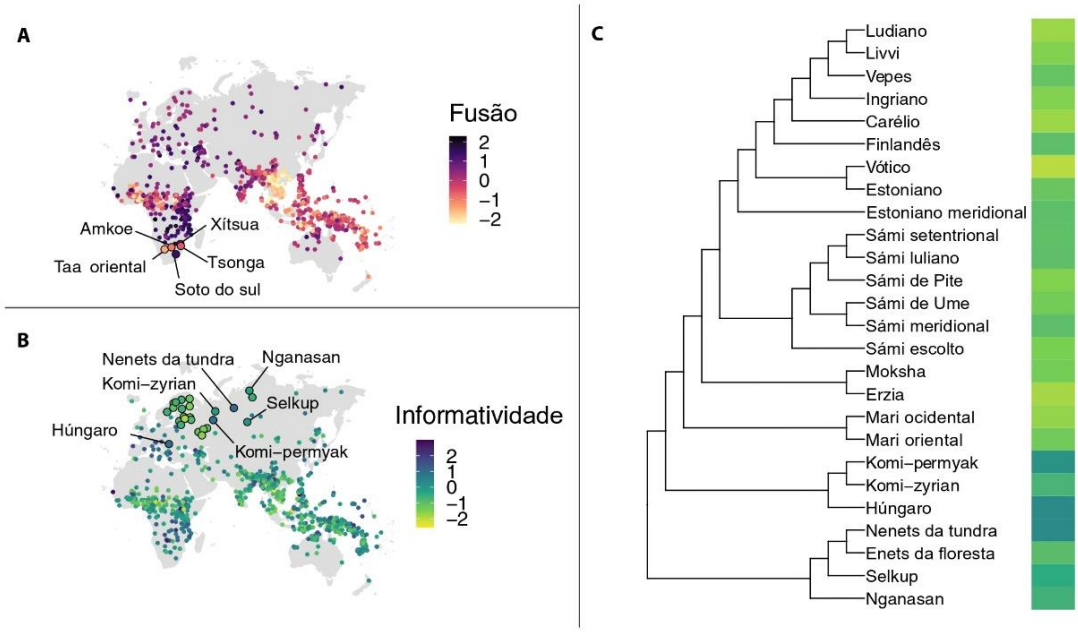

Figura 4: A distribuição global das pontuações de fusão e informatividade e a distribuição das pontuações de fusão no subgrupo da árvore global

Nota: os painéis mostram a distribuição de pontuações de fusão na África Meridional (**A**) e pontuações de informatividade na Eurásia (**B**) com um foco nas línguas urálicas e a filogenia das línguas urálicas inclusas na árvore global (**C**). As pontuações com um mínimo de 0 (ausência de todos os traços da métrica) e um máximo de 1 (presença de todos os traços da métrica) foram padronizadas com uma média de 0 e variância de 1. A diferença nas pontuações de fusão entre duas línguas banto — tsonga (baixa) e soto do sul (alta) — podem ser explicadas pela perspectiva do parentesco filogenético, tendo em vista a semelhança que o tsonga guarda com as pontuações de outros casos atípicos de baixa pontuação na África Meridional vindos de outras famílias linguísticas além da banta e, em menor extensão, com a língua irmã e vizinha xítsua (A). As pontuações de informatividade mais altas nas línguas urálicas revelam um padrão visível de agrupamento filogenético (C) e encontram-se no samoiedo (nganasan, selkup, nenets da tundra e enets da floresta), úgrico (húngaro) e permiano (komi-permyak e komi-zyrian), que se distanciaram mais cedo do resto das línguas. As pontuações mais altas de informatividade nas línguas urálicas pode ser atribuída também ao contato com as línguas indo-europeias (B): o húngaro é rodeado por línguas europeias com pontuações altas e os falantes de nenets da tundra e do komi-permyak são normalmente bilíngues, também falando russo.

Outra vantagem de nossa abordagem é o uso de duas métricas de complexidade em uma amostra global abrangente. As pesquisas empíricas anteriores focaram em um domínio gramatical, como o número de casos (Bentz; Winter 2013) ou a síntese verbal (Sinnemäki; Di Garbo 2018), ou incluíram uma grande variedade de traços relativos a procedimentos diferentes de codificação e interpretação de complexidade — cf. os estudos baseados no WALS (Dryer; Haspelmath 2013), como Lupyan e Dale (2010), Koplenig (2019) e Bentz *et al.* (2016). Além disso, seguimos uma abordagem sistemática focada em delinear a fusão e a informatividade, o que permite tomar decisões baseadas em princípios sobre a escolha de traços do Grambank para cada

métrica e evitar a inclusão de traços que não se alinham com nenhuma das interpretações descritas. Pesquisas futuras poderiam explorar a possibilidade de outras dimensões de complexidade gramatical serem sensíveis à influência de fatores sociodemográficos. Um caminho promissor para pesquisas futuras seria medir o grau de desvio do princípio "um significado para uma forma" das línguas (Miestamo 2008) em uma escala interlinguística.

Assim, as descobertas anteriores positivas podem ser frutos de amostras pequenas e não representativas, métricas que englobam características gramaticais que se enquadram em dimensões de complexidade distintas ou métodos sem controle suficiente da genealogia e geografia. Ao superar essas limitações anteriores e usar uma amostra ampla com um poder estatístico melhor, não achamos evidências que as sociedades de estranhos falam línguas menos complexas gramaticalmente e que os fatores sociodemográficos usados neste estudo são fortes agentes de fusão e informatividade. Pesquisas futuras poderiam se basear na perspectiva metodológica adotada neste artigo e examinar o impacto de um grupo de variáveis sócio-históricas com mais nuances. O banco de dados global Grambank poderia ser usado para estudar outras questões ambiciosas sobre a diversidade e a evolução linguística.

Embora outras estruturas linguísticas, por exemplo, o léxico (Wray; Grace 2007; Raviv; Meyer; Lev-Ari 2019a; 2019b) ou traços gramaticais individuais, como a marcação de caso (Bentz; Winter 2013; Sinnemäki 2020), talvez se adaptem a fatores sociodemográficos em mudança, não achamos evidências de que as duas dimensões de complexidade gramatical medidas aqui são sensíveis às pressões sociolinguísticas na maneira como foram concebidas como hipótese. Concluímos que a herança filogenética e o empréstimo entre vizinhas próximas explicam a maior parte da distribuição da complexidade gramatical entre as línguas do mundo em relação a essas duas variáveis. Essa descoberta sugere que, mesmo se o tamanho populacional de falantes tiver um papel na diminuição da complexidade das línguas, o poder de seleção é baixo. O tamanho populacional pode mudar de forma rápida e imprevisível por eventos externos (guerras, doenças e migrações), sem falar no crescimento populacional (Greenhill *et al.* 2018). Essas mudanças talvez não deixem vestígios perceptíveis na complexidade gramatical por duas razões: a estabilidade da natureza filogenética dessas variáveis complexas restringe a taxa em que esses traços se adaptam

ao ambiente sociolinguístico ou a labilidade do tamanho populacional significa que a adaptação é mais lenta que as pressões de seleção do novo regime seletivo.

É importante que os trabalhos futuros explorem os efeitos em potencial de outros fatores sociodemográficos com um maior refinamento do que as variáveis demográficas que estão disponíveis atualmente. A atenção deveria ser investida naqueles fatores que mudam em um ritmo relativamente lento e, assim, proporcionam pressões de seleção relativamente duradouras. Novos estudos deveriam considerar com cuidado a interação entre a genealogia e a geografia quando modelarem a adaptação das línguas ao ambiente e explicar o motivo de alguns traços serem mais sensíveis do que outros às pressões sociolinguísticas.

Materiais e métodos

Banco de dados

Métricas

Nesta pesquisa, o foco está em duas interpretações da complexidade gramatical delineadas para melhor entender as forças que as modelam:

1) o grau de fusão — quanto uma língua depende de marcadores de fusão fonológica (Bickel; Nichols 2007);
2) o número de distinções gramaticais explícitas e obrigatórias que não são costumeiramente marcadas em todas as línguas.

Esses fenômenos linguísticos são estimados por duas métricas que são baseadas em traços disponíveis no Grambank v1.0 (https://doi.org/10.5281/zenodo.7740140). A primeira métrica mede a fusão. Ela é responsável por marcadores gramaticais de fusão fonológica. As pontuações aumentam à medida que existem mais marcadores de fusão fonológica. Para cada traço do Grambank relacionado aos marcadores de fusão, uma língua pode receber 1 "ponto de fusão" se for codificada como "presente" no banco de dados. Se essa língua não usar um marcador de fusão (ela é codificada como "ausente") para um determinado traço, então ela recebe 0 pontos de fusão. Em seguida, calculamos a média de todas essas pontuações de fusão por língua para definir a pontuação.

Por exemplo, para a marcação de plural em substantivos, uma língua como o inglês, que produz as formas no plural com marcadores fonológicos de fusão (*-s*), recebe 1 ponto de fusão para esse traço. Por outro lado, línguas como o rapanui e o maori, que não têm marcadores de fusão fonológica com o objetivo de marcar o plural nominal, ou o vietnamita, sem marcação de plural nos substantivos, recebe 0 para esse traço de fusão. Em geral, essa métrica se destina a indicar com sistematicidade os marcadores gramaticais de fusão fonológica e pode ser comparada a outras métricas que registram o grau de complexidade morfológica (Lupyan; Dale, 2010), a complexidade do inventário (Nichols; Bentz 2018) e o sintetismo (Szmrecsanyi; Kortmann 2009).

Não incluímos os traços associados com a derivação (GB047, GB048 e GB049) ou a marcação morfossintática (GB146). Além disso, excluímos os traços relacionados ao caso morfológico em pronomes (GB071 e GB073), uma vez que, em línguas onde a marcação de caso é presente, pronomes, em oposição aos substantivos, são particularmente propensos ao supletismo, o que representa um exemplo de morfologia não linear em vez de aditiva.

A métrica de fusão foi construída a partir dos dados do Grambank, que, por sua vez, depende da leitura de descrições gramaticais e a comunicação com especialistas. Reconhecemos que autores diferentes podem ter outras abordagens sobre o que eles definem como “fusão” e “independência fonológica”. Entretanto, apesar dessas diferenças em potencial, as pontuações de fusão resultantes estão em concordância com o modo que trabalhos anteriores classificam as línguas do mundo em relação à complexidade morfológica (ver Materiais e métodos para a comparação da pontuação de fusão e a pontuação e os pareceres sobre a complexidade morfológica na literatura existente).

A segunda métrica consiste de traços que contribuem para explicitar a marcação obrigatória de distinções gramaticais e semânticas que não são comumente marcadas em todas as línguas. Para isso, excluímos os domínios gramaticais tipicamente evidentes relacionados à negação, posse, construções comparativas, perguntas polares e marcação de argumentos A e P (com a ajuda da ordem das palavras, casos em palavras nominais ou índices em verbos). As línguas recebem pontuações pela presença de traços de informatividade. Marcadores fonológicos de fusão e independentes têm uma contribuição igual para a pontuação final de informatividade

quando eles indicam a mesma função gramatical. Isso significa que, se uma língua tem uma marcação obrigatória de plural com marcadores fonológicos de fusão e/ou marcadores fonológicos independentes, ela receberá 1 ponto de informatividade. A partir do exemplo anterior, o inglês, o maori e o rapanui receberiam 1, enquanto o vietnamita teria 0. Essa métrica revela o grau de marcação gramatical nas línguas e segue uma ideia parecida com a medição da gramaticalidade (Szmrecsanyi; Kortmann, 2009), com a ressalva de que excluímos alguns domínios que são geralmente marcados de modo explícito e obrigatório na maioria das línguas (negação, perguntas polares etc.).

Em seguida, as pontuações das duas métricas para cada função gramatical foram somadas e o valor médio foi obtido para cada língua, para que o mínimo possível para uma língua que não tem traços relacionados à fusão ou à informatividade seja 0 e o valor máximo possível seja 1 se uma língua tem todos os traços pertinentes relacionados à fusão ou à informatividade. Na realidade, não achamos línguas que atingem as pontuações máximas em ambas as métricas, mas várias línguas (quase) se aproximam do mínimo de 0. Por exemplo, uma das línguas com menor pontuação (0) em fusão é o hu (austro-asiático), sem nenhum traço de fusão, enquanto a pontuação em informatividade mais baixa, 0,1, foi obtida pelo jukun takum (atlântico-congo). O tariana (aruaque) atinge a maior pontuação nas duas métricas: 0,7 em fusão e 0,66 em informatividade. Logo após, as pontuações de 0 a 1 foram padronizadas para a média de 0 e variância de 1.

Medimos a fusão e informatividade apenas para aquelas línguas que são bem descritas no Grambank (Skirgård *et al.* 2023b) e removemos as línguas com mais do que 25% de valores faltantes em todos os traços do Grambank. De 2430 línguas no conjunto de dados, computamos a pontuação de fusão e de informatividade de 1291 línguas. Dessa forma, as pontuações resultantes são representações robustas das dimensões de complexidade pretendidas.

Apesar do foco em marcadores fonológicos de fusão em vez de inflexões morfológicas, nossa métrica de fusão ainda é comparável com o que outros estudos mediram como complexidade morfológica. Por exemplo, de acordo com a métrica baseada em uma variedade de traços do WALS, o turco teve uma classificação extremamente alta (0,775) e o vietnamita (austro-asiático) foi a língua com a menor pontuação (0,141) (Bentz *et al.* 2016). Em nossos dados, não quantificamos as

pontuações de fusão para o vietnamita, mas observamos uma divisão entre o turco e algumas línguas austro-asiáticas que não têm todos os traços de fusão: o turco recebeu 0,44 e a língua no limite máximo, o tariana (aruaque), chegou em uma pontuação alta de 0,7, enquanto a pontuação de fusão do thavung (do mesmo ramo viético que o vietnamita), do hu, do prai e do rumai palaung é igual a 0. De forma semelhante, nossa métrica apreende os contrastes entre línguas do ramo kiranti e kuki-chin, como sugerido por DeLancey (2015). Reivindica-se que o camling (sino-tibetano) apresenta uma complexidade morfológica extrema e pontua 0,42 em nossa métrica, enquanto que se afirma que o mara chin (sino-tibetano) perdeu complexidade, o que se reflete em sua pontuação mais baixa: 0,12 (DeLancey 2015). Além disso, os padrões geográficos na distribuição da fusão estão alinhados com a literatura tipológica e os estudos de complexidade. Um *hotspot* de destaque para a baixa fusão se localiza no sul do continente asiático, o que é esperado devido ao perfil tipológico das línguas dessa área (Enfield 2005), e outro se situa no oeste africano, o que está alinhado com a proposta do cinturão de baixa complexidade (Bentz 2016).

Variáveis sociodemográficas

A fusão e a informatividade são previstas com base nas seguintes variáveis demográficas e sociais: número de falantes de L1 e veicularidade (juntamente com a proporção de falantes de L2 na amostra de 120 línguas) (Eberhard; Simons; Fennig 2021) e número de línguas vizinhas, o *status* da língua (oficial/não oficial) e o uso na educação (língua de educação/não é língua de educação), que são obtidos das informações adicionais disponíveis em Bromham *et al.* (2022), a parir de dados coletados pelos autores ou recuperados originalmente do World Language Mapping System v16 e v17 (WLMS, http://worldgeodatasets.com) e de Leclerc (2019).

Fizemos a transformação logarítmica dos números brutos de falantes L1 com uma base 10 e, então, padronizamos essa variável e o número de línguas vizinhas para obter uma média de 0 e variância de 1. A padronização é feita para eliminar os efeitos potenciais de casos atípicos extremos nos resultados da modelagem filogenética espacial. O número de falantes de L1 transformado em logaritmo (mas não padronizado) foi usado para implementar o termo de interação entre essa variável e a veicularidade (ou a proporção de L2 na amostra de 120 línguas).

O número de línguas vizinhas foi calculado a partir do número de intersecções em um círculo de 10000 km² de uma determinada língua e os polígonos de outras línguas (Bromham *et al.* 2022). Os polígonos e as localizações de ponto único foram coletados do WLMS v16 e v17; as localizações de ponto único foram usadas para aproximar áreas linguísticas usando projeções de Voronoi quando o WLMS não forneceu polígonos (Bromham *et al.* 2022).

O *status* de uma língua é uma variável binária: ele pode ser não oficial ou oficial em nível nacional, e a língua é usada ou não é usada na educação. As línguas foram codificadas como oficiais se fossem formalmente reconhecidas assim ou se fossem tratadas como as principais línguas de educação, comércio, meios de comunicação e governo em países que não reconhecem formalmente qualquer língua como oficial, como a Austrália (Bromham *et al.* 2022). Nestes casos, essas línguas também foram codificadas como línguas de educação se nenhuma outra língua era utilizada como língua de educação no território (Bromham *et al.* 2022). Todas as línguas minoritárias que não foram reconhecidas como oficiais, mas foram usadas como meio de instrução na educação, de acordo com o *L'aménagement linguistique dans le monde* (Leclerc 2019), também foram designadas como línguas de educação (ver Bromham *et al.* (2022) e os materiais adicionais para mais detalhes sobre a codificação das variáveis sociodemográficas utilizadas).

Reconhecemos que a modelagem de traços gramaticais com preditores sociodemográficos flutuantes representa desafios. Observamos que (i) os dados do Grambank e as fontes de variáveis sociodemográficas são típicas das populações contemporâneas e (ii) as variáveis sociodemográficas são particularmente propensas a mudar mais rapidamente (Bentz *et al.* 2015; Sinnemäki 2009). Pesquisas anteriores interpretaram de forma diacrônica a correlação entre estruturas gramaticais sincrônicas e sociodemográficas. No entanto, não está claro quanto tempo é necessário para que as mudanças no ambiente sociolinguístico desencadeiem as mudanças nas estruturas gramaticais. Além disso, o uso de variáveis demográficas comuns, como o tamanho da população, a proporção de falantes de L2 e o número de línguas vizinhas, como representantes de exotericidade, tem sido questionado em diversos estudos (Sinnemäki; Di Garbo 2018; Verkerk; Di Garbo 2022; Di Garbo; Kashima 2021; Di Garbo; Verkerk 2022). Embora essas limitações sejam inerentes a todas as pesquisas interlinguísticas sobre a ligação entre estruturas gramaticais e exotericidade, nosso

estudo supera as limitações anteriores relativas ao tamanho da amostra, ao controle de fontes não independentes e à escolha de traços linguísticos.

Informações filogenéticas e geográficas

Nós usamos o Glottolog 4.5 (Hammarström *et al.* 2021) para verificar as informações sobre as localizações e a família linguística das línguas em nossa amostra. A base de dados AUTOTYP (Bickel *et al.* 2022) forneceu as informações sobre a distribuição das línguas em 24 áreas linguísticas.

Para modelar as mudanças de evolução dos traços linguísticos no tempo e controlar a ancestralidade comum, mapeamos as pontuações de complexidade gramatical e os valores de variáveis sociodemográficas na filogenia EDGE global (Bouckaert *et al.* 2022). Essa superárvore global integra as informações de classificação linguística do Glottolog e as filogenias publicadas, assim como as respectivas localizações.

Modelagem filogenética espacial

Adotamos uma técnica de modelagem filogenética espacial proposta inicialmente por Dinnage, Skeels e Cardillo (2020). Essa abordagem bayesiana usa a Aproximação de Laplace Aninhada e Integrada (INLA, do inglês *integrated nested Laplace approximation*) (Rue; Martino; Chopin 2009; Martins *et al.* 2013) para estimar a distribuição posterior conjunta dos parâmetros do modelo e é implementada pelo pacote R-INLA (R Core Team 2022; Rue; Martino; Chopin 2009). Além de avaliar os coeficientes de efeitos fixos, a modelagem filogenética espacial nos permitiu calcular a influência relativa dos efeitos aleatórios — aqui, relações espaciais (distâncias geográficas) e filogenéticas — na variável resposta.

Construção de efeitos aleatórios

Com o objetivo de incorporar as relações filogenéticas e espaciais como efeitos aleatórios nos modelos, passamos por várias etapas para representar essas relações na forma de matrizes de precisão, conforme exigido pelo INLA. Tornamos essas matrizes

comparáveis ao padronizá-las com uma variância de 1. A seguir, explicamos como esse processo é feito.

Construímos uma matriz filogenética de variância-covariância que quantifica o comprimento dos ramos compartilhados entre as línguas na árvore global com base na suposição de um modelo browniano de movimento da evolução e com o uso da função "vcv.phylo" no software ape (Paradis; Schliep 2019). A árvore serviu para a construção de uma matriz de precisão filogenética padronizada com a aplicação da função "inverseA" no pacote MCMCglmm (Hadfield 2010).

Seguimos um conjunto semelhante de etapas para calcular e padronizar as matrizes espaciais. Primeiro, calculamos uma matriz de variância-covariância com a função de covariância espacial de Matérn implementada na função "varcov.spatial" no pacote geoR (Ribeiro *et al.* 2020). Segundo, padronizamos a matriz de variância-covariância pela variância e a invertemos para criar uma matriz de precisão. Por último, a matriz de variância-covariância padronizada foi transformada na matriz de precisão. Estimamos duas matrizes espaciais para dois conjuntos de parâmetros: (i) $\phi = 1{,}25$ e $\kappa = 1$ ("conjunto local") e (ii) $\phi = 17$ e $\kappa = 1$ ("conjunto regional") (cf. Claessens, Kyritsis e Atkinson (2020) e Skirgård *et al.* (2023a) para outros exemplos de como esse tipo de controle de autocorrelação espacial é implementado). Esses parâmetros da matriz de covariância espacial foram escolhidos para diferenciar duas suposições: de acordo com os parâmetros correspondentes ao conjunto local, a difusão de pontuações em métricas semelhantes entre línguas não é provável em distâncias superiores a 1000 km, enquanto, com os parâmetros do conjunto regional, a difusão pode ocorrer ao longo de vários milhares de quilômetros (cf. a Figura S1 para mais detalhes). O ajuste dos modelos com cada uma dessas matrizes nos permitiu comparar as suposições para saber qual é a difusão das pontuações das métricas que corresponde aos nossos dados: as línguas são mais propensas a ter pontuações semelhantes apenas em nível local ao longo de centenas de quilômetros ou a difusão é provável em regiões maiores, como continentes ou grandes áreas linguísticas que se estendem por milhares de quilômetros? Para comparação, introduzimos também um terceiro controle para autocorrelação espacial: efeitos aleatórios de 24 áreas da base de dados AUTOTYP (Bickel *et al.* 2022), em que cada área é considerada independente da outra, por meio da qual negligenciamos as distâncias geográficas entre as áreas. Entende-se que esses três efeitos espaciais diferentes estão em competição entre si e representam diferentes

formas de operacionalizar a influência do contato entre línguas vizinhas. Os modelos que incorporam os efeitos filogenéticos e versões locais de efeitos espaciais predizem melhor a fusão e a informatividade (cf. as Tabelas 1 e 2).

Teste de sensibilidade

Usamos prioris de complexidade penalizada (PC, do inglês *penalized complexity*) (Simpson *et al.* 2017) para a precisão da verossimilhança e os efeitos filogenéticos e espaciais. As matrizes de precisão são padronizadas para ter uma variância de 1. Nos resultados do texto principal, os prioris de PC são definidos de modo que 10% da densidade de probabilidade do priori em relação ao SD da verossimilhança ou dos efeitos aleatórios esteja acima de 1. Como teste de sensibilidade, fizemos a variação da densidade de probabilidade em 1, 10, 50 e 99%, mas isso não afeta nossas conclusões (cf. Tabela S5).

Medição do sinal filogenético

Estimamos o lambda (λ) de Pagel (Pagel 1999) para medir o sinal filogenético de fusão e informatividade na árvore EDGE global usando o pacote R, phytools (Revell 2012). Os valores de λ normalmente variam de 0 a 1. O λ = 1 implica um sinal filogenético alto, o que significa que as pontuações evoluem de uma maneira esperada de acordo com um modelo de movimento browniano. Por outro lado, o λ = 0 sugere que não há nenhum sinal filogenético e indica que a distribuição das pontuações evoluiu de modo independente das relações filogenéticas entre as línguas na filogenia.

Materiais suplementares

O arquivo em PDF inclui o seguinte: Figura S1 e Tabelas S1 a S5[12].

[12] Disponíveis em: https://www.science.org/doi/suppl/10.1126/sciadv.adf7704/suppl_file/sciadv.adf7704_sm.pdf.

Referências

AĬKHENVALD, Aleksandra Yurevna. *A grammar of Tariana, from northwest Amazonia*. Cambridge: Cambridge University Press, 2003.

BAECHLER, Raffaela. Diachronic complexification and isolation. *Yearbook of the Poznan Linguistic Meeting*, v. 1, p. 1-28, 2015

BENTZ, Christian. The Low-Complexity-Belt: Evidence for large-scale language contact in human prehistory. *In*: ROBERTS, Sean G. *et al.* (ed.). *Proceedings of the 11th International Conference (EVOLANG11) on The Evolution of Language*, New Orleans, LA, USA, v. 21, 2016.

BENTZ, Christian; RUZSICS, Tatyana; KOPLENIG, Alexander; SAMARDŽIĆ, Tanja. A comparison between morphological complexity measures: typological data vs. language corpora. *In*: BRUNATO, Dominique; DELL'ORLETTA, Felice; VENTURI, Giulia; FRANÇOIS, Thomas; BLACHE, Philippe (ed). *Proceedings of the workshop on computational linguistics for linguistic complexity (cl4lc)*. Osaka: The COLING 2016 Organizing Committee, 2016. p. 142-153.

BENTZ, Christian; VERKERK, Annemarie; KIELA, Douwe; HILL, Felix; BUTTERY, Paula. Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PloS one*, v. 10, n. 6, p. e0128254, 2015.

BENTZ, Christian; WINTER, Bodo. Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change*, v. 3, n. 1, p. 1-27, 2013.

BICKEL, Balthasar; NICHOLS, Johanna. Inflectional morphology. *Language typology and syntactic description*, v. 3, n. 2, p. 169-240, 2007.

BICKEL, Balthasar; NICHOLS, Johanna; ZAKHARKO, Taras; WITZLACK-MAKAREVICH, Alena; HILDEBRANDT, Kristine; RIEßLER, Michael; BIERKANDT, Lennart; ZÚÑIGA, Fernando; LOWE, John B. The AUTOTYP database (v1. 0.1). *Zenodo*, 2022. Disponível em: https://zenodo.org/record/6793367

BOUCKAERT, Remco; REDDING, David; SHEEHAN, Oliver; KYRITSIS, Thanos; GRAY, Russell; JONES, Kate E.; ATKINSON, Quentin. Global language diversification is linked to socio-ecology and threat status. *SocArXiv Papers*, 2022.

BROMHAM, Lindell *et al.* Global predictors of language endangerment and the future of linguistic diversity. *Nature ecology & evolution*, v. 6, n. 2, p. 163-173, 2022.

BROMHAM, Lindell. Curiously the same: swapping tools between linguistics and evolutionary biology. *Biology & Philosophy*, v. 32, p. 855-886, 2017.

CHEN, Sihan; GIL, David; GAPONOV, Sergey; REIFEGERSTE, Jana; YUDITHA, Tessa; TATARINOVA, Tatiana; PROGOVAC, Ljiljana; BENITEZ-BURRACO, Antonio. Linguistic and memory correlates of societal variation: A quantitative analysis. *PsyArViv Preprints*, 2023.

CLAESSENS, Scott; KYRITSIS, Thanos; ATKINSON, Quentin D. Revised analysis shows relational mobility predicts sacrificial behavior in Footbridge but not Switch or Loop trolley problems. *Proceedings of the National Academy of Sciences*, v. 117, n. 24, p. 13203-13204, 2020.

CLAHSEN, Harald *et al*. Morphological structure in native and nonnative language processing. *Language learning*, v. 60, n. 1, p. 21-43, 2010.

DAHL, Östen. *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: John Benjamins, 2004.

DALE, Rick; LUPYAN, Gary. Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in complex systems*, v. 15, n. 03n04, p. 1150017, 2012.

DELANCEY, Scott. The Historical Dynamics of Morphological Complexity in Trans-Himalayan. *Linguistic Discovery*, v. 13, n. 2, 2015.

DI GARBO, Francesca; Kashima, Eri; Napoleão de Souza, Ricardo; Sinnemäki, Kaius. Concepts and methods for integrating language typology and sociolinguistics. *In*: BALLARÈ, Silvia; INGLESE, Guglielmo (ed.). *Tipologia e Sociolinguistica: Verso un approccio integrato allo studio della variazione:* Atti del Workshop della Società Linguistica Italiana 20 settembre 2020, v. 5, p. 143-176. 2021.

DI GARBO, Francesca; VERKERK, Annemarie. A typology of northwestern Bantu gender systems. *Linguistics*, v. 60, n. 4, p. 1169-1239, 2022.

DINNAGE, Russell; SKEELS, Alexander; CARDILLO, Marcel. Spatiophylogenetic modelling of extinction risk reveals evolutionary distinctiveness and brief flowering period as threats in a hotspot plant genus. *Proceedings of the Royal Society B*, v. 287, n. 1926, p. 20192817, 2020.

DRYER, Matthew. S.; HASPELMATH, Martin. (ed). *The world atlas of language structures online*. München: Max Planck Digital Library, 2013.

EBERHARD, David M.; SIMONS, Gary F.; FENNIG, Charles D. (ed.). Ethnologue: languages of the world. *Ethnologue Global Dataset, 24rd Edition*. Dallas, Texas: SIL International, 2021. Disponível em: http://www.ethnologue.com

EFF, E. Anthon. Does Mr. Galton still have a problem? Autocorrelation in the standard cross-cultural sample. *World Cultures*, v. 15, n. 2, p. 153-170, 2004.

ENFIELD, Nick J. Areal linguistics and mainland Southeast Asia. *Annu. Rev. Anthropol.*, v. 34, p. 181-206, 2005.

GELMAN, Andrew; HWANG, Jessica; VEHTARI, Aki. Understanding predictive information criteria for Bayesian models. *Statistics and computing*, v. 24, n. 6, p. 997-1016, 2014.

GIVÓN, Talmy. Context as other minds: the pragmatics of sociality. *Cognition and Communication*. Amsterdam: John Benjamins, 2005.

GIVÓN, Talmy; YOUNG, Phil. Cooperation and interpersonal manipulation in the society of intimates. *In*: SHIBATANI, Masayoshi (ed.). *The grammar of causation and interpersonal manipulation*. Amsterdam: John Benjamins, 2002. p. 23-56.

GREENHILL, Simon J. *et al*. Population size and the rate of language evolution: a test across Indo-European, Austronesian, and Bantu languages. *Frontiers in psychology*, v. 9, p. 576, 2018.

HADFIELD, Jarrod D. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of statistical software*, v. 33, p. 1-22, 2010.

HAMMARSTRÖM, Harald; FORKEL, Robert; HASPELMATH, Martin; BANK, Sebastian. glottolog/glottolog: Glottolog database 4.4. *Zenodo*, 2021. [Conjunto de dados].

HARLOW, Ray. Māori [*Languages of the World/Materials 20*]. Munich and Newcastle: LINCOM Europa, 1996.

IGARTUA, Iván. Loss of grammatical gender and language contact. *Diachronica*, v. 36, n. 2, p. 181-221, 2019.

KOPLENIG, Alexander. Language structure is influenced by the number of speakers but seemingly not by the proportion of non-native speakers. *Royal Society open science*, v. 6, n. 2, p. 181274, 2019.

KUSTERS, Wouter *et al*. Complexity in linguistic theory, language learning and language change. *In*: MIESTAMO, Matti *et al*. (ed.). *Language Complexity*. Amsterdam: John Benjamins, 2008.

KUSTERS, Wouter. *Linguistic complexity*. Utrecht: LOT — Netherlands Graduate School of Linguistics, 2003.

LECLERC, Leclerc. Index alphabétique complet de tous les pays, états, territoires ou régions (souverains ou non) de ce site. *L'aménagement linguistique dans le monde*, 2019. Disponível em: https://www.axl.cefan.ulaval.ca/monde/index_alphabetique.htm

LEVINSON, Stephen C.; GRAY, Russell D. Tools from evolutionary biology shed new light on the diversification of languages. *Trends in cognitive sciences*, v. 16, n. 3, p. 167-173, 2012.

LUPYAN, Gary; DALE, Rick. Language structure is partly determined by social structure. *PloS one*, v. 5, n. 1, p. e8559, 2010.

LUPYAN, Gary; DALE, Rick. Why are there different languages? The role of adaptation in linguistic diversity. *Trends in cognitive sciences*, v. 20, n. 9, p. 649-660, 2016.

MACE, Ruth *et al*. The comparative method in anthropology [and comments and reply]. *Current anthropology*, v. 35, n. 5, p. 549-564, 1994.

MAITZ, Péter; NÉMETH, Attila. Language contact and morphosyntactic complexity: Evidence from German. *Journal of Germanic Linguistics*, v. 26, n. 1, p. 1-29, 2014.

MARTINS, Thiago G.; SIMPSON, Daniel; LINDGREN, Finn; RUE, Håvard. Bayesian computing with INLA: new features. *Computational Statistics & Data Analysis*, v. 67, p. 68-83, 2013.

MCWHORTER, John. *Language interrupted: Signs of non-native acquisition in standard language grammars*. Oxford: Oxford University Press, 2007.

MIESTAMO, Matti *et al*. Grammatical complexity in a cross-linguistic perspective. *In*: MIESTAMO, Matti; SINNEMÄKI, Kaius; KARLSSON, Fred (ed.). *Language Complexity*. Amsterdam: John Benjamins, 2008.

NICHOLS, Johanna; BENTZ, Christian. Morphological complexity of languages reflects the settlement history of the Americas. *New Perspectives on the Peopling of the Americas*, 2019.

NICHOLS, Johanna; BICKEL, Balthasar; DRYER, Matthew; HASPELMATH, Martin. The World Atlas of Language Structures Online (v2020.3). *Zenodo*, 2013. [Conjunto de dados]. Disponível em: https://doi.org/10.5281/zenodo.7385533

NIKOLAEVA, Irina. *A grammar of Tundra Nenets*. Berlin/Boston: De Gruyter Mouton, 2014.

PAGEL, Mark. Inferring the historical patterns of biological evolution. *Nature*, v. 401, n. 6756, p. 877-884, 1999.

PARADIS, Emmanuel; SCHLIEP, Klaus. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, v. 35, n. 3, p. 526-528, 2019.

R CORE TEAM. Developement Core. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, 2022. Disponível em: http://www.r-project.org/

RAVIV, Limor; MEYER, Antje; LEV-ARI, Shiri. Compositional structure can emerge without generational transmission. *Cognition*, v. 182, p. 151-164, 2019.

RAVIV, Limor; MEYER, Antje; LEV-ARI, Shiri. Larger communities create more systematic languages. *Proceedings of the Royal Society B*, v. 286, n. 1907, p. 20191262, 2019.

REVELL, Liam J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in ecology and evolution*, n. 2, p. 217-223, 2012.

RIBEIRO JUNIOR, Paulo J.; DIGGLE, Peter J. Analysis of geostatistical data. *The geoR package, version, 2006*. p. 1-6, 2020.

RUE, Håvard; MARTINO, Sara; CHOPIN, Nicolas. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B:* Statistical Methodology, v. 71, n. 2, p. 319-392, 2009.

SIMPSON, Daniel; RUE, Håvard; RIEBLER, Andrea; MARTINS, Thiago G., SØRBYE, Sigrunn H. Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.* v. 32, n. 1, p. 1–28, 2017.

SINNEMÄKI, Kaius. Complexity in core argument marking and population size. In: *Language complexity as an evolving variable*. Oxford: Oxford University Press, 2009. p. 126-140.

SINNEMÄKI, Kaius. Linguistic system and sociolinguistic environment as competing factors in linguistic variation: A typological approach. *Journal of Historical Sociolinguistics*, v. 6, n. 2, p. 20191010, 2020.

SINNEMÄKI, Kaius; DI GARBO, Francesca. Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: A typological study of verbal and nominal complexity. *Frontiers in psychology*, v. 9, p. 1141, 2018.

SKIRGÅRD, Hedvig *et al.* Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, v. 9, n. 16, p. eadg6175, 2023a.

SKIRGÅRD, Hedvig; HAYNIE, Hannah J.; BLASI, Damián E.; PASSMORE, Sam; CHIRA, Angela; MAURITS, Luke; DINNAGE, Russell; FORKEL, Robert; GREENHILL, Simon J.; ENGLISCH, Johannes. grambank/grambank-analysed: grambank-analysed v1.0 (v1.0). *Zenodo*, 2023b. [Software]. Disponível em: https://doi.org/10.5281/zenodo.7740822

SZMRECSANYI, Benedikt; KORTMANN, Bernd. Between simplification and complexification: Non-standard varieties of English around the world. *In*: SAMPSON, Geoffrey; GIL, David; TRUDGILL, Peter (ed.). *Language Complexity as an Evolving Variable*. Oxford: Oxford University, 2009. p. 64-79.

THURSTON, William R. How exoteric languages build a lexicon: Esoterogeny in West New Britain. *VICAL*, v. 1, p. 555-579, 1989.

THURSTON, William R. *Processes of change in the languages of north-western New Britain*. Canberra: The Australian National University, 1987.

THURSTON, William R. Sociolinguistic typology and other factors effecting change in northwestern New Britain, Papua New Guinea. *In*: DUTTON, Tom. *Culture*

change, language change: Case studies from Melanesia. Canberra: Australian National University, 1992.

TRUDGILL, Peter. *Sociolinguistic typology:* Social determinants of linguistic complexity. Oxford: Oxford University Press, 2011.

VERKERK, Annemarie; DI GARBO, Francesca. Sociogeographic correlates of typological variation in northwestern Bantu gender systems. *Language Dynamics and Change*, v. 12, n. 2, p. 155-223, 2022.

WATANABE, Sumio. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, v. 11, n. 12, 2010.

WRAY, Alison; GRACE, George W. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, v. 117, n. 3, p. 543-578, 2007.

ZEIJLSTRA, Hedde. Negation in natural language: On the form and meaning of negative elements. *Language and Linguistics Compass*, v. 1, n. 5, p. 498-518, 2007.

Agradecimentos: ficamos gratos pelo auxílio na aplicação técnica do INLA a R. Dinnage. Agradecemos pela ajuda com o design gráfico a H.-G. Sell, em particular, com a Figura 4. O D.E.B. foi apoiado por uma Bolsa Branco Weiss e uma Bolsa Harvard Data Science. Agradecemos ao público do 55º Encontro Anual da Societas Linguistica Europaea e da Conferência Conjunta sobre Evolução da Linguagem pelos seus comentários. **Financiamento**: este trabalho foi apoiado pelo Departamento de Evolução Linguística e Cultural (para O.S., R.D.G., S.J.G. e H.S.), pela Bolsa Branco Weiss (para D.E.B.) e pela Bolsa Harvard Data Science (para D.E.B.). **Contribuições dos autores**: Conceitualização: O.S., H.S., S.M.M. e H.J.H. Planejamento das métricas: H.S. e H.J.H. Revisão das métricas: O.S. e S.M.M. Metodologia: O.S., H.S., S.P., H.J.H., S.J.G. e D.E.B. Software: O.S., S.P., H.S. e S.J.G. Análise formal e investigação: O.S. Visualização: O.S., H.S., S.P. e S.J.G. Supervisão: H.S., S.M.M., D.E.B., S.J.G., R.D.G. e V.G. Redação — versão original: O.S., S.M.M., D.E.B., H.S. e R.D.G. Redação — revisão e edição: todos os autores. **Conflito de interesses**: os autores declaram não ter interesses conflitantes. **Disponibilidade dos dados e materiais**: todos os dados, códigos e materiais estão disponíveis no texto principal, nos Materiais suplementares e em arquivos digitais pelo Zenodo

(https://doi.org/10.5281/zenodo.10420654), exceto o conjunto de dados do Ethnologue em sua forma original. As variáveis do Ethnologue (número padronizado de falantes de L1, número padronizado e com transformação logarítmica de falantes de L1, número padronizado de todos os usuários das línguas, número padronizado e com transformação logarítmica de todos os usuários das línguas e proporção de falantes de L2) foram disponibilizados publicamente através de um acordo concluído para transferência dos materiais. O acesso a essas variáveis permite ajustar todos os modelos nas análises, exceto o modelo que inclui a interação entre o número de falantes de L1 com transformação logarítmica e a veicularidade (ou proporção de falantes de L2 em uma subamostra de 120 línguas). O acordo só permite disponibilizar publicamente as variáveis que não podem ser transformadas de volta a sua forma original. Como o número de falantes de L1 com transformação logarítmica poderia ser transformado de volta, um conjunto especial de *scripts* dentro do repositório foi criado para que os usuários sem acesso ao Ethnologue possam executar todos os modelos, exceto o mencionado. Os leitores interessados podem acessar o conjunto de dados original do Ethnologue assinando o nível "*Essentials*" em www.ethnologue.com/pricing/, o que custaria US$ 40/mês por usuário com cobrança anual ou US$ 199 por cada mês. O conjunto de dados do Grambank v1.0 está disponibilizado em um arquivo da internet no Zenodo (https://doi.org/10.5281/zenodo.7740140).