UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE FÍSICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA

MAELYSON ROLIM FONSECA DOS SANTOS

HEURÍSTICA E LEIS DE ESCALA PARA UM PROBLEMA COMPLEXO: A DISTRIBUIÇÃO DA DIVERSIDADE LINGUÍSTICA NA TERRA**.**

Recife
2021

MAELYSON ROLIM FONSECA DOS SANTOS

HEURÍSTICA E LEIS DE ESCALA PARA UM PROBLEMA COMPLEXO: A DISTRIBUIÇÃO DA DIVERSIDADE LINGUÍSTICA NA TERRA.

Tese apresentada ao Programa de Pós-Graduação em Física da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de doutor em Física. Área de Concentração: Física Teórica e Computacional.

Orientador: Prof. Dr. Marcelo Andrade de Filgueiras Gomes

Recife

2021

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

S237h Santos, Maelyson Rolim Fonseca dos
Heurística e leis de escala para um problema complexo: a distribuição da diversidade linguística na terra / Maelyson Rolim Fonseca dos Santos. – 2021.
117 f.: il., fig., tab.

Orientador: Marcelo Andrade de Filgueiras Gomes.
Tese (Doutorado) – Universidade Federal de Pernambuco. CCEN, Física, Recife, 2021.
Inclui referências e apêndice.

1. Física teórica e computacional. 2. Leis de escala. 3. Sistemas complexos. I. Gomes, Marcelo Andrade de Filgueiras (orientador). II. Título.

530.1 CDD (23. ed.) UFPE- CCEN 2021 - 151

MAELYSON ROLIM FONSECA DOS SANTOS

HEURÍSTICA E LEIS DE ESCALA PARA UM PROBLEMA COMPLEXO: A DISTRIBUIÇÃO DA DIVERSIDADE LINGUÍSTICA NA TERRA

Tese apresentada ao Programa de Pós-Graduação em Física da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Física.

Aprovada em: 26/08/2021.

BANCA EXAMINADORA

_____

Prof. Marcelo Andrade de Filgueiras Gomes
Orientador
Universidade Federal de Pernambuco

_____

Profa. Azadeh Mohammadi
Examinadora Interna
Universidade Federal de Pernambuco

_____

Prof. Paulo Roberto de Araujo Campos
Examinador Interno
Universidade Federal de Pernambuco

_____

Profa. Viviane Moraes de Oliveira
Examinadora Externa
Universidade Federal Rural de Pernambuco

_____

Prof. Tsang Ing Ren
Examinador Externo
Universidade Federal de Pernambuco

Para Bruna, os dias todos.

AGRADECIMENTOS

Esta tese é resultado da hábil orientação, do talento incomparável e da generosa compreensão do Professor Marcelo Gomes. Desde a minha graduação sou inspirado por seu trabalho e foi uma honra pesquisar no Laboratório de Simulação Analógica. Muito obrigado Professor. Senti muito quando a pandemia interrompeu as conversas que tínhamos em seu gabinete rodeados por sua incrível biblioteca.

Sou muito grato a Alexsandra Melo por salvar-me nas/das atividades burocráticas sempre com excelência, simpatia e muita gentileza.

Obrigado Dona Fagna (mainha) e Seu Mauricio (painho) por me ensinarem a amar o conhecimento, estimularem meus sonhos e me acolherem após cada queda. Eu amo vocês!

Tias Corrinha, Cristina e Liduina, tios Eraldo, Fagno e Sérgio, não consigo mensurar quão fundamentais vocês são em minha vida. Obrigado.

Obrigado Dona Romana, Seu Giko e Laís por me acolherem em sua família.

Aos meus irmãos Cássio e Larissa, meu muito obrigado pela companhia nas muitas – e incompreensíveis – jornadas da vida.

Obrigado Doutora Débora pelo partilhar da mesa e tão boas conversas.

Queridos Pedro, Tati e Maria, a vida faz mais sentido porque estamos juntos. Obrigado por sempre estarem.

Obrigado Hemilly. Foi muito importante contar com tua amizade e tuas orações. (Ainda espero o podcast.)

Obrigado Leopoldo, Raquel, Alice e Victor pela amizade para além dos agradáveis almoços e jantares. Leo, uma vez mais repito que “as palavras convencem, o exemplo arrasta” :D

Obrigado Andrew, Guilherme, Leopoldo (de novo!), Marcelo, Pedro (de novo!) e Wando pela irmandade.

Sou grato a Francisco Carol por ser um exemplar amigo e companheiro de laboratório, discussões, chamadas, *playlists*, jogos e Caminho.

Obrigado Abel Jr. pela constante e gentil disposição para conversas aleatórias e orientações estatísticas.

Agradeço aos bons companheiros que tive durante a jornada no Departamento de Física da UFPE enquanto compartilhávamos a sala de aula, o laboratório ou os estudos para os EGDs. Obrigado Adson, Daniel, Mário, Milton, Tiago Araújo, Hugo, Renata Hora,

Tawan, Fillipe Cesar, Ricardo Batista, Alyson José e Victor Hugo.

Sem uma ligação de Thiago Sobral eu não chegaria ao doutorado. Obrigado meu velho.

Muito obrigado à comunidade que se reúne como Mangue por me ensinarem o idioma da Fé, da Esperança e do Amor.

Obrigado Milton Nascimento e Lô Borges por *Clube da Esquina*, Jóhann Jóhannsson por *The Theory of Everything* e Of Monsters and Men por *My Head Is an Animal*.

Valeu Paul Zaloom e Mark Rits!

Agradeço a CAPES, CNPq e FACEPE por financiarem minha pesquisa e o Laboratório de Simulação Analógica.

Agradeço ao Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco pela licença que me permitiu finalizar a pesquisa e a escrita desta tese.

Ao longo deste doutorado fui diariamente acolhido, incentivado e amado por Bruna Peralva. Obrigado Bruna por escolher compartilhar comigo o teu momento na história e o teu lugar no Universo. As palavras que não consigo escrever são todas tuas.

Soli Deo Gloria

"Houve um tempo em que todos os habitantes do mundo falavam a mesma língua e usavam as mesmas palavras" (BÍBLIA. . . , 2016).

RESUMO

O surgimento da linguagem, ocorrido antes do êxodo africano, é possivelmente a transição mais significativa na evolução dos hominídeos. No último século tem sido observada a crescente extinção de idiomas, representando não apenas uma redução no repertório de visões de mundo, mas também uma redução do conhecimento biológico, ecológico, geográfico e tecnológico. Nas últimas décadas, a expansão dos bancos de dados sobre idiomas foi acompanhada por um correspondente número de estudos estatísticos, propostas de modelos e simulações computacionais da dinâmica da linguagem. Nesta tese, fazendo uso de ferramentas provenientes da física estatística de sistemas complexos, são investigados padrões da atual distribuição linguística na Terra. Utilizando registros relativos a mais de sete mil idiomas, são apresentadas leis de escala alométricas entre a diversidade linguística e os tamanhos geográfico, demográfico e econômico dos países. Em seguida é discutido um modelo fractal que justifica o expoente $z = 1/3$ da lei de escala observada entre a diversidade linguística e a área. Em outro capítulo, um modelo heurístico simples, que emerge de um tipo de estrutura variacional, sugere, a partir de um princípio de maximização, que a relação entre a diversidade linguística e a área dos países deve ser caracterizada por um expoente algo maior. Este último modelo reproduz a lei de escala observada hoje na Terra para os cento e quarenta e sete maiores países. Ulteriormente, a relação entre diversidade linguística e área é examinada através de um modelo termodinâmico de campo médio com forças entrópicas e de auto-exclusão. É exposto que este modelo concorda com os dados empíricos que mostram a diminuição da diversidade linguística com o aumento da latitude e fornece ainda uma base para entender cenários futuros de perda da diversidade linguística. A partir de processos de classificação e ordenamento, são apresentadas funções hiperbólicas associadas à parâmetros linguísticos, econômicos, geográficos e demográficos dos países. As distribuições acumuladas de idiomas bem como de grupos étnicos em função da população são apresentadas e são investigadas as leis de escala emergentes da classificação das famílias linguísticas segundo o número de idiomas e o número de falantes. Por fim, é analisada a distribuição de tamanhos de idiomas das catorze maiores famílias linguísticas contemporâneas.

Palavras-chave: diversidade linguística; leis de escala; sistemas complexos.

ABSTRACT

The emergence of language, which occurred before the African exodus, is possibly the most significant transition in the evolution of hominids. The last century has witnessed a growing extinction of languages, representing not only a reduction in the repertoire of worldviews, but also a reduction in biological, ecological, geographic, and technological knowledge. In recent decades, the growth of language databases has been accompanied by an increasing number of statistical studies, model proposals, and computer simulations of language dynamics. In this work, using tools from statistical physics of complex systems, are investigated patterns of Earth's current distribution of languages. Using data on more than $7,000$ languages, allometric scaling laws are presented between linguistic diversity and the geographic, demographic, and economic size of countries. Then a fractal model is discussed that justifies the exponent $z = 1/3$ observed in the scaling law between linguistic diversity and area. In another chapter, a simple heuristic model, which emerges from a type of variational framework, suggests, from a maximization principle, that the relation between linguistic diversity and country area should be characterized by an exponent somewhat larger. This latter model reproduces the scaling law observed on Earth today for the one hundred and forty-seven largest countries. Later, the relation of linguistic diversity to area is examined using a mean-field thermodynamic model with entropic and self-avoidance forces. It is exposed that this model agrees with empirical data showing decreasing linguistic diversity with increasing latitude and providing a basis for understanding future scenarios of loss of linguistic diversity. From classification and ordering processes, hyperbolic scaling laws associated with linguistic, economic, geographic, and demographic parameters of the countries are presented. The cumulative distributions of languages as well as ethnic groups as a function of the population are presented and the scaling laws emerging from the classification of language families according to the number of languages and the number of speakers are investigated. Finally, the languages size distribution for the largest fourteen contemporary language families is analyzed.

Keywords: complex systems, linguistic diversity; scaling.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

"Porque tudo isto são palavras, e só palavras,
fora das palavras não há nada,"

José Saramago

O surgimento da linguagem, ocorrido antes do êxodo africano, é possivelmente a transição mais significativa na evolução dos hominídeos. A expressão da linguagem em termos de diferentes idiomas tem sido objeto de investigação e amplo interesse desde a antiguidade como exemplificado no famoso experimento de Psamético narrado por Heródoto em sua *História*, no épico sumério *Enmercar e o Senhor de Arata* e no relato do *Pentateuco* sobre a Torre de Babel. A diversidade linguística, além de um dos traços culturais mais difundidos e duradouros, guarda similaridades com alguns padrões da biodiversidade, como por exemplo pelo aumento da diversidade linguística em direção ao equador (gradiente latitudinal) e pela difusão de idiomas em maiores áreas nas latitudes mais altas (regra de Rapoport). Tais similaridades, bem como outras relações, entre a diversidade linguística, cultural e biológica têm sido estudadas ao longo das últimas décadas por pesquisadores de diversos campos. Nesta tese utilizando ferramentas provenientes da física estatística de sistemas complexos estudamos alguns padrões da atual distribuição linguística na Terra. Em particular, estaremos interessados na relação entre a diversidade linguística, quantificada pelo número total de idiomas, e a área.

1.1 Bancos de dados

O *Ethnologue* (SIMONS; FENNIG, 2017) é uma das poucas listagens extensivas dos idiomas do mundo. Além da sua característica quase única e do seu amplo uso na pesquisa nas últimas décadas, um fator determinante para a adoção dele se deve ao fato de que as informações apresentadas nele têm sido atualizadas continuamente desde a primeira edição produzida em 1951. Uma pertinente revisão deste banco de dados, ainda que realizada para edições anteriores àquela adotada em nosso trabalho, foi realizada por Hammarström (HAMMARSTRÖM, 2015). Os dados utilizados nesta tese foram obtidos a partir da vigésima edição do *Ethnologue* publicada em versão digital em 2017. No entanto, ao longo dos últimos três anos o acesso gratuito ao *Ethnologue* foi sendo dificultado e chegou a ser totalmente bloqueado em 2020 (MATACIC, 2020).

O número de idiomas vivos é uma questão fundamental cuja resposta se faz necessária destrinchar. Na apresentação do *Ethnologue* se lê que

> because languages are constantly changing and demonstrate significant internal variation, the total number of living languages in the world cannot be known precisely. Therefore, that number changes as knowledge of the world's languages improves.[1]

Isto pode ser observado no fato da décima terceira edição do *Ethnologue*, utilizada no primeiro trabalho sobre relações de escalas para a diversidade de idiomas (GOMES et al., 1999), listar 6700 idiomas enquanto a vigésima edição, aqui utilizada, lista um total de 7099 idiomas vivos. A edição que utilizamos também contém informações relativas a 360 idiomas extintos recentemente, ou seja, que ficaram fora de uso desde a sua primeira edição.

Como uma nota alarmante, lembramos que a crescente extinção de idiomas, representa não apenas uma redução no repertório de visões de mundo, mas também uma redução do conhecimento biológico, ecológico, geográfico e tecnológico. Diante das mudanças climáticas e seus riscos à sobrevivência humana é crucial observar a conexão entre a biodiversidade e a diversidade cultural e linguística humana (WILDER et al., 2016; FRAINER et al., 2020).

Para dirimir casos como aqueles em que o mesmo idioma pode ser chamado por nomes diferentes e casos em que idiomas diferentes recebem a mesma nomenclatura, os dados apresentados pelo *Ethnologue* contém um código identificador exclusivo de três letras para cada idioma em um padrão nomeado ISO 639-3[2]. Assim é possível distinguir, por exemplo, o idioma Amba, que pertence a família Nigero-Congolesa e cujo código identificador é rwm, do idioma Amba, pertencente à família Austronesiana e cujo código identificador é utp.

Quanto a vitalidade dos idiomas, o *Ethnologue* resume o estado de cada um deles segundo a Escala Graduada Expandida de Interrupção Intergeracional (EGIDS na sigla em inglês) (LEWIS; SIMONS, 2010). Esta escala é composta por treze níveis, com cada número mais alto na escala representando um maior nível de interrupção na transmissão intergeracional do idioma. A Tabela 1 fornece definições resumidas dos treze níveis da escala e a Figura 1 apresenta o número de idiomas em cada nível. Desta figura observamos que mais da metade dos idiomas listados no *Ethnologue* são categorizados como escritos ou vigorosos e que apenas um conjunto muito pequeno de idiomas tem caráter internacional.

O *Ethnologue* adota o padrão ISO 3166[3] para determinar quais entidades geopolíticas devem ser listadas como países. Dessa forma exibe registros linguísticos para 236 países. Este valor, maior que os 193 membros da Organização das Nações Unidas, é justificado por

[1] "Como os idiomas estão mudando constantemente e demonstram uma variação interna significativa, o número total de idiomas vivos no mundo não pode ser conhecido com precisão. Portanto, esse número muda conforme o conhecimento das línguas do mundo melhora.'

[2] <https://iso639-3.sil.org/> (acesso em 14/07/2021).

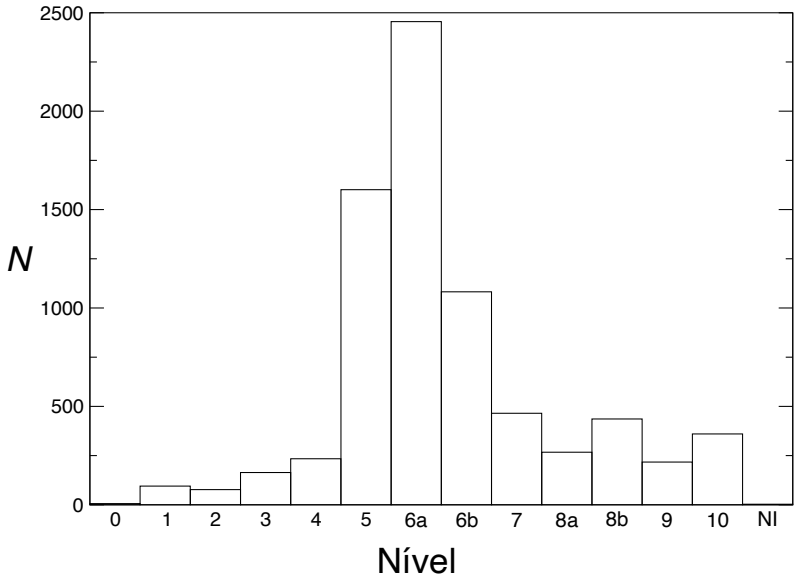
[3] <https://www.iso.org/iso-3166-country-codes.html> (acesso em 14/07/2021).

Tabela 1 – Escala Graduada Expandida de Interrupção Intergeracional para idiomas.

Nível	**Etiqueta**	**Descrição**
0	Internacional	A língua é usada internacionalmente para uma ampla gama de funções.
1	Nacional	O idioma é utilizado na educação, trabalho, meios de comunicação, e governo a nível nacional.
2	Regional	A língua é utilizada para meios de comunicação e serviços governamentais locais e regionais.
3	Comércio	A língua é usada para o trabalho local e regional, tanto por membros da comunidade como por pessoas de fora.
4	Educacional	A alfabetização na língua está sendo transmitida através de um sistema de educação pública.
5	Escrito	O idioma é utilizado por via oral, por todas as gerações e é utilizado informalmente na forma escrita em algumas partes da comunidade.
6a	Vigoroso	O idioma é utilizado por via oral por todas as gerações, e está sendo aprendido pelas crianças como sua primeira língua.
6b	Ameaçado	O idioma é utilizado por via oral, por todas as gerações, mas apenas alguns membros da geração fértil o transmitem aos filhos.
7	Em perigo	Os membros da geração fértil conhecem suficientemente bem a língua para usá-la entre si, mas nenhum deles está transmitindo-o aos seus filho.
8a	Moribundo	Os únicos falantes ativos da língua ainda vivos são membros da geração dos avós.
8b	Quase extinto	Os únicos falantes ativos da língua ainda vivos são membros da geração dos avós ou dos bisavós, que têm pouca oportunidade de usar a língua.
9	Simbólico	A língua serve como lembrete da identidade ou herança cultural para uma comunidade étnica. Ninguém tem mais do que uma proficiência simbólica em termos de usar a língua.
10	Extinto	Ninguém mantém mais um sentimento de identidade étnica associada com a respectiva língua, mesmo para fins simbólicos.

Fonte: *Em defesa das línguas minoritárias do Brasil* (EBERHARD, 2013).

Figura 1 – Número N de idiomas em cada nível da Escala Graduada Expandida de Interrupção Intergeracional (EGIDS). A categoria NI contém três idiomas sem informações disponíveis. Para detalhes ver Tabela 1. Dados extraídos da vigésima edição do *Ethnologue* publicada em 2017.



Fonte: Autor (2021).

considerar, por exemplo, departamentos ultramarinos franceses, como Guiana Francesa, Guadalupe e Ilha de Reunião como entidades geopolíticas independentes. As informações relativas às populações dos países também foram obtidas do *Ethnologue*.

Informações sobre a geolocalização dos idiomas foram obtidas do *Glottolog*, uma crescente base de dados gratuita sobre línguas (HAMMARSTRÖM et al., 2020). Os registros relativos às áreas dos países foram obtidos dos Indicadores de Desenvolvimento Mundial do Banco Mundial (WORLD BANK, 2017b). Este também foi o banco de dados utilizado para obtenção de dados do Produto Interno Bruto (PIB) relativo ao ano de 2016 (WORLD BANK, 2017a). Entretanto tais informações econômicas não estavam disponíveis para todo o supracitado conjunto de 236 países, mas sim para 194 países[4]. São estes 194 países que constituem o conjunto que será analisado a partir do Capítulo 3. Os dados de temperatura média anual foram obtidos do *World Bank's Climate Change Knowledge Portal* (WORLD BANK, 2020) limitados, no entanto, a 186 países[5]

[4] Os quarenta e dois países cujo valor de $PIB-2016$ não estava disponível nos Indicadores de Desenvolvimento Mundial do Banco Mundial são Andorra, Anguila, Aruba, Bermudas, Coreia do Norte, Curaçao, Eritreia, Gibraltar, Guadalupe, Guernsey, Guiana Francesa, Ilha de Reunião, Ilha de São Bartolomeu, Ilha de São Martinho, Ilha Norfolk, Ilhas Cayman, Ilhas Cook, Ilhas Feroe, Ilhas Pitcairn, Ilhas Virgens Americanas, Jersey, Líbia, Martinica, Mayotte, Mônaco, Montserrat, Niue, Nova Caledônia, Países Baixos Caribenhos, Polinésia Francesa, Porto Rico, Saint-Pierre e Miquelon, San Marino, São Martinho, Síria, Taiwan, Território do Oceano Índico Britânico, Tokelau, Turks e Caicos, Vaticano, Venezuela e Wallis e Futuna.

[5] Os oito países cujo valor da temperatura média anual não estava disponível no *World Bank's Climate Change Knowledge Portal* são Guam, Hong Kong, Ilha de Man, Ilhas Virgens Britânicas, Macau, Nauru,

1.2 Sobre a estrutura desta tese

Explorando registros de mais de sete mil idiomas obtidos do *Ethnologue*, como discutido na próxima seção, apresentaremos relações alométricas para a diversidade linguística dos países, ou seja, a relação entre o tamanho geográfico, demográfico e econômico dos países e o total de idiomas falados em seus territórios. Este número de idiomas e sua ampla variação de tamanhos, discutido Capítulo 6, permite descrições por leis de escala, ferramenta introduzida no estudo de fenômenos físicos desde *Duas Novas Ciências* (GALILEI, s/d), a última obra galileliana, compreendida como o apanhado do conjunto de trabalhos e pensamentos do cientista pisano. Naquelas páginas, enquanto especulava as bases da ciência moderna, Galileu discutia questões relativas à resistência dos materiais usando hipóteses de escala. Contemporaneamente, leis de escala aparecem na modelagem de vários fenômenos complexos que transcendem os limites da física (WEST; BROWN; ENQUIST, 1997; BROCK, 1999; BETTENCOURT et al., 2007; ZHANG; YU, 2010). De fato, conforme aponta Geoffrey B. West "To varying degrees, fractality, scale invariance, and self-similarity are ubiquitous across nature from galaxies and clouds to your cells, your brain, the Internet, companies, and cities."[6] (WEST, 2017).

No Capítulo 2 discutimos brevemente alguns aspectos da linguagem e dos sistemas complexos e introduzimos os conceitos de fractais e leis de escala. Concomitantemente apresentamos e revisamos excertos da produção científica associada ao estudo da distribuição da diversidade linguística na Terra.

No Capítulo 3 apresentamos leis de escala alométricas entre a diversidade linguística e os tamanhos geográfico, demográfico e econômico dos países. Ainda neste capítulo, um modelo fractal que justifica o expoente da lei de escala observada entre a diversidade linguística e a área é introduzido.

O Capítulo 4 expõe um modelo heurístico simples, que emerge de um tipo de estrutura variacional, com o objetivo de justificar a relação entre diversidade linguística e área a partir de um princípio de maximização.

No Capítulo 5 é apresentado um modelo termodinâmico de campo médio com forças entrópicas e de auto-exclusão a partir do qual a relação entre diversidade linguística e área atualmente observada na Terra pode ser entendida. É discutido como o modelo concorda com os dados empíricos que mostram a diminuição da diversidade linguística com o aumento da latitude e, em seguida, é pontuado como o modelo fornece uma base para entender cenários futuros de perda de diversidade linguística.

O Capítulo 6 apresenta a partir de processos de classificação e ordenamento, funções

Palestina e Samoa Americana.

[6] "Em vários graus, fractalidade, invariância de escala e auto-similaridade são onipresentes na natureza, desde galáxias e nuvens até suas células, seu cérebro, a Internet, empresas e cidades."

hiperbólicas associadas à parâmetros linguísticos, econômicos, geográficos e demográficos dos países. As distribuições acumuladas de idiomas bem como de grupos étnicos em função da população são apresentadas e são investigadas as leis de escala emergentes da classificação das famílias linguísticas segundo o número de idiomas e o número de falantes. Por fim, é analisada a distribuição de tamanhos de idiomas das catorze maiores famílias linguísticas contemporâneas.

Por fim, no Capítulo 7 apresentamos as considerações finais deste trabalho, resumimos nossas conclusões e apontamos possibilidades de pesquisas futuras.

2 ANTECEDENTES

"Porque a natureza é uma obra em
aberto que nos cabe aceitar e potenciar."

Valter Hugo Mãe

Aqui introduziremos conceitos que serão utilizados nos próximos quatro capítulos. Para este fim, o presente capítulo segue uma trajetória convergente iniciada com uma breve discussão acerca do amplo tema da linguagem. Em seguida são discutidos alguns aspectos dos sistemas complexos para então serem apresentadas duas importantes ferramentas para nossa abordagem: a geometria fractal e as leis de escala. A penúltima seção apresenta leis de escala associadas aos estudos da linguagem. Por fim discutimos alguns aspectos matemáticos e estatísticos das leis de potência. Ao longo das seis seções apontamos, dentro da vasta literatura, textos que consideramos adequados para uma introdução ao problema complexo que é a distribuição da diversidade linguística na Terra.

2.1 Linguagem

No trecho final da Primeira Jornada do *Diálogo sobre os dois máximos sistemas do mundo*, o personagem Sagredo reflete sobre a perspicácia do engenho humano. Dentre as muitas invenções que o espantam, Sagredo declara que

> acima de todas as invenções estupendas, que superioridade de espírito foi a daquele homem que imaginou encontrar um modo de comunicar seus pensamentos mais recônditos a qualquer outra pessoa, ainda que distante por um intervalo muito grande de lugar e de tempo! Falar com aqueles que estão nas Índias, falar com aqueles que ainda não nasceram, nem existirão senão daqui a mil ou dez mil anos! E com quanta facilidade, com a junção de vinte pequenos caracteres sobre um papel! (GALILEI, 2011).

Inegavelmente é espantoso que a junção de pequenos caracteres sobre um papel nos permitam ler, quatro séculos depois, um dos textos fundamentais para o desenvolvimento da ciência moderna, mesmo sabendo que o homem imaginado por Sagredo não existiu.

Tal possibilidade se faz concreta devido a faculdade da linguagem que "é um traço nos seres humanos anatômicos modernos, que deve ter aparecido antes do êxodo africano" (BERWICK; CHOMSKY, 2017). Lembremos que o aparecimento do *Homo Sapiens* é estimado em cerca de quinhentos a trezentos mil anos enquanto o *Homo Sapiens Sapiens* emergiu nos últimos cem mil anos (CAVALLI-SFORZA; CAVALLI-SFORZA, 2002). É

interessante observar que, apesar do aparecimento da linguagem ser compreendido como uma transição significativa na evolução dos hominídeos (SMITH; SZATHMARY, 1997), a evolução da linguagem continua sendo um dos grandes mistérios de nossa espécie (HAUSER et al., 2014).

Muito embora a linguagem seja considerada a coisa mais notável sobre nosso eu moderno (TATTERSALL, 2012) é comum o impasse com relação a sua definição. Recentemente, Noam Chomsky escreveu que "2500 anos de estudos intensos e produtivos não conseguiram chegar a uma resposta clara sobre o que é a linguagem." (CHOMSKY, 2018). Além disso,

> às vezes, o termo linguagem é usado para se referir à linguagem humana; às vezes, é usado para se referir a qualquer sistema simbólico ou a qualquer modo de comunicação ou representação como quando se fala da linguagem das abelhas, ou de linguagens de programação, ou ainda de linguagem das estrelas e assim por diante (BERWICK; CHOMSKY, 2017).

Berwick e Chomsky apontam que entre as muitas questões intrigantes sobre a linguagem, duas se destacam. Primeiro, por que ela existe? Em segundo lugar, por que existem tantas línguas? (BERWICK; CHOMSKY, 2017). Como veremos, a diversa distribuição linguística da Terra apresenta alguns interessantes padrões que serão investigados ao longo dessa tese. Para tal fim faremos uso de ferramentas provenientes da física estatística de sistemas complexos que serão discutidas a partir da próxima seção.

2.2 Sistemas complexos

Em um primeiro olhar, idiomas podem parecer não guardar muitas similaridades com outros fenômenos cotidianos como o trânsito, as movimentações financeiras ou com aqueles que, enquanto vivemos em grandes aglomerados urbanos, assistimos com temor como as propagações de incêndios florestais e epidemias. No entanto, em todos esses casos estamos diante de sistemas onde múltiplos agentes, interagindo entre si e com o meio externo, resultam na emergência de fenômenos coletivos. Em *How Nature Works*, Per Bak aponta que "We see complex phenomena around us so often that we rake them for granted without looking for further explanation."[1] (BAK, 1996).

As definições de um sistema complexo, ou de complexidade, são tão diversas quanto tais fenômenos (LADYMAN; LAMBERT; WIESNER, 2013). Enquanto Per Bak define, de modo sintático, como complexos os sistemas com grande variabilidade (BAK, 1996) e Parisi afirma que um sistema é complexo se seu comportamento depende crucialmente dos detalhes do sistema (PARISI, 1999), Mitchell define como complexo

> a system in which large networks of components with no central control and simple rules of operation give rise to complex collective behavior,

[1] "Vemos fenômenos complexos ao nosso redor com tanta frequência que os desconsideramos, sem buscar mais explicações."

> sophisticated information processing, and adaptation via learning or evolution.[2] (MITCHELL, 2009).

Para fins desta tese, podemos compreender como complexo um sistema com muitos componentes que interagem entre si de modo que o comportamento emergente coletivo desses componentes é mais do que a soma de seus comportamentos individuais. Essa definição pode ser ainda mais abreviada ao se fazer uso do lema *More is different*[3] proveniente do título de um artigo de P. W. Anderson publicado em 1972 (ANDERSON, 1972). Anderson, além de questionar a generalidade do ponto de vista reducionista, discutiu como a natureza seria organizada de forma hierárquica de modo que certas propriedades emergentes surgiriam ao se ascender entre os níveis de hierarquia. Esse clássico artigo ajuda a compreender que não é apenas nos campos interdisciplinares que sistemas complexos são observados. Muitos dos fenômenos estudados pela física da matéria condensada e física estatística também são exemplos de sistemas complexos.

Enquanto alguns autores sugerem que a ciência dos sistemas complexos é uma nova disciplina (THURNER; KLIMEK; HANEL, 2018), Pietronero considera que o estudo da complexidade é menos uma nova disciplina científica e mais uma mudança de perspectiva na *forma mentis* dos cientistas (PIETRONERO, 2008). De toda forma, é crucial que tenhamos em mente que "Nature can produce complex structures even in simple situations, and can obey simple laws even in complex situations."[4] (GOLDENFELD; KADANOFF, 1999).

Do mesmo modo como a palavra latina que originou o verbete *complexo* carrega a ideia de entrelaçamento, há um tecido de muitas técnicas que são utilizadas no estudo de sistemas complexos. Fractais e leis de escala são algumas das mais importantes destas ferramentas e as duas próximas seções têm como objetivo apresentá-las.

2.3 Fractais

Durante a década de 70 do século passado, um editor francês considerou que o título *Objets physiques de dimension fractionnaire*[5] não era uma boa escolha para o livro que Benoit B. Mandelbrot estava escrevendo. Em sua autobiografia, Mandelbrot relata que, após a intervenção do editor, buscou uma palavra que carregasse a ideia de uma pedra quebrada, algo irregular e fragmentado (MANDELBROT, 2012). Os estudos de latim feitos por aquele autor durante a juventude o levaram até o adjetivo latino *fractus*

[2] "Um sistema em que grandes redes de componentes sem controle central e regras simples de operação dão origem à um comportamento coletivo complexo, ao processamento sofisticado de informações e à adaptação por meio de aprendizagem ou evolução."

[3] Mais é diferente.

[4] "A natureza pode produzir estruturas complexas mesmo em situações simples e pode obedecer a leis simples mesmo em situações complexas."

[5] Objetos concretos de dimensão fracionária.

cujo significado atendia completamente suas expectativas de descrever algo quebrado. O livro foi então nomeado *Les objets fractals*[6] (MANDELBROT, 1975) e o termo fractal se tornou um verbete presente no estudo dos mais diversos fenômenos. É digno de nota que muitos objetos identificados atualmente como fractais foram amplamente estudados no final do século 19 e no início do século 20 sobretudo por matemáticos interessados no problema das curvas contínuas porém não diferenciáveis (SCHROEDER, 2009).

Alguns anos antes da publicação do livro em que apresentou o termo fractal, Mandelbrot publicou um artigo onde argumentava que estruturas geográficas que carregavam a característica de serem autosimilares poderiam ser descritas por uma quantidade $\mathscr{D}$ que possuiria propriedades de dimensão embora pudesse assumir um valor fracionário (MANDELBROT, 1967). É conveniente lembrar que a dimensão de um conjunto pode ser compreendida como uma grandeza que fornece tanto uma descrição de quanto espaço o conjunto preenche quanto informações sobre as propriedades geométricas de tal conjunto (FALCONER, 2014). Dessa forma a dimensão fractal, nas palavras de Mandelbrot, seria "uma medida do grau de irregularidade e de fragmentação" de um objeto (MANDELBROT, 1998).

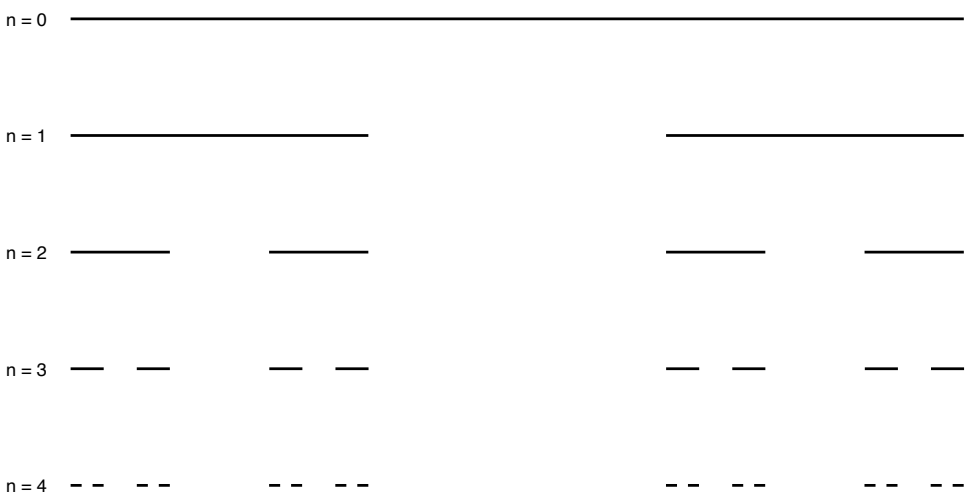
Embora ainda não tenhamos definido autosimilaridade, podemos compreender este conceito a partir de alguns processos iterativos simples. Iniciemos na etapa $n = 0$ com um segmento de reta disposto horizontalmente conforme a Figura 2. Na etapa $n = 1$ removemos o terço central desse segmento de reta. Agora, de posse de dois segmentos de reta, na etapa $n = 2$ removemos os terços centrais destes segmentos. Prosseguimos sucessivamente removendo a terça parte central de cada novo segmento de reta. Quando $n \to \infty$, a coleção de infinitos segmentos gerados por este processo iterativo formará o que é conhecido como Conjunto de Cantor.

Como segundo exemplo, tomemos, em $n = 0$, um triângulo equilátero conforme a Figura 3. O processo gerador deste conjunto consiste na retirada de um triângulo equilátero invertido de altura igual a metade da altura do triângulo original, conforme observamos nas etapas $n = 1$, $n = 2$ e $n = 3$ da Figura 3. Quando $n \to \infty$, a figura obtida será o Triângulo (ou Gaxeta) de Sierpiński.

É interessante observar na Figura 2 que os dois ramos que surgem em $n = 1$ são cópias do segmento de reta da etapa $n = 0$ reescalonadas por um fator de 1/3. Dessa forma, é possível afirmar que o Conjunto de Cantor é autosimilar. De igual modo, observando a Figura 3, podemos compreender que o Triângulo de Sierpiński também possui a característica de ser autosimilar tendo 1/2 como fator de reescalonamento. Estes dois exemplos, construídos por processos iterativos geométricos, nos permitem compreender que há uma relação entre o número de cópias e o fator de reescolanamento em um fractal.

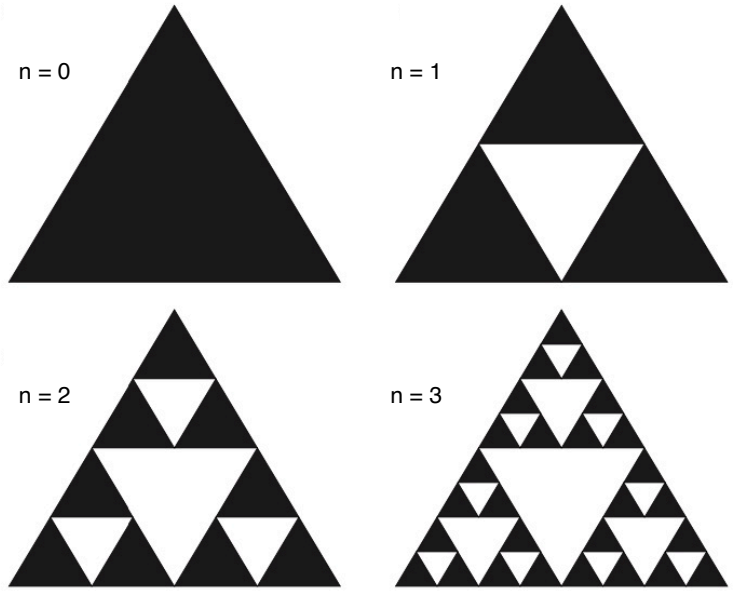
[6] Os objetos fractais.

Figura 2 – Primeiras etapas da construção do Conjunto de Cantor.



Fonte: Autor (2021).

Figura 3 – Primeiras etapas da construção do Triângulo de Sierpiński.



Fonte: Autor (2021).

De modo geral, temos que

$$Nr^{D_S} = 1 \tag{2.1}$$

com D_S definido como a dimensão de autosimilaridade para o caso onde N é o número de cópias iguais obtidas a partir do todo segundo o fator de reescalonamento r. Podemos reescrever a Equação 2.1 de modo a obter

$$D_S = \frac{\log N}{\log(1/r)}. \tag{2.2}$$

Para o caso do Conjunto de Cantor, onde $N = 2$ e $r = 1/3$, temos

$$D_S = \frac{\log 2}{\log 3} \approx 0,6309. \tag{2.3}$$

Enquanto para o Triângulo de Sierpiński, onde $N = 3$ e $r = 1/2$, temos

$$D_S = \frac{\log 3}{\log 2} \approx 1,585. \tag{2.4}$$

Sendo $0,6309$ maior que a dimensão topológica de um ponto ($D_T = 0$) e $1,585$ maior que a dimensão topológica de uma reta ($D_T = 1$), podemos, de modo geral, compreender um fractal como um objeto cuja dimensão de autosimilaridade é maior que sua dimensão topológica. Estes dois exemplos podem induzir à conclusão equivocada de que sistemas autosimilares têm necessariamente dimensão fractal não inteira. Um exemplo de objeto fractal de dimensão inteira é a versão tridimensional da Gaxeta de Sierpiński cuja dimensão fractal é igual a dois.

Embora outras definições de dimensão ampliem e aprofundem a caracterização de fractais, aqui, conforme lista Falconer (FALCONER, 2014), compreenderemos um conjunto $\mathscr{F}$ como fractal quando as seguintes propriedades forem observadas:

(i) $\mathscr{F}$ possuir estrutura fina, isto é, detalhes em escalas arbitrariamente pequenas;

(ii) $\mathscr{F}$ for, tanto local quanto globalmente, muito irregular de modo que a linguagem geométrica tradicional não o descreva adequadamente;

(iii) $\mathscr{F}$ for de alguma forma autosimilar, mesmo que de modo aproximado ou estatístico;

(iv) A dimensão fractal de $\mathscr{F}$ for, de modo geral, maior do que sua dimensão topológica;

(v) $\mathscr{F}$ for definido de uma forma muito simples, possivelmente de modo recursivo.

2.4 Leis de escala

Essencialmente, todos os problemas que serão discutidos nesta tese poderão ser descritos em termos de distribuições do tipo leis de escala. Tais distribuições podem ser compreendidas como a expressão matemática da autosimilaridade presente em muitos sistemas complexos. Por lei de escala entendemos a existência de uma função não linear do tipo lei de potência de modo que

$$f(x) = Ax^{\alpha} \tag{2.5}$$

com A e α constantes e $\alpha \neq 0$. Notemos que se tomarmos $x \to \lambda x$, com λ constante, teremos

$$f(\lambda x) = A\lambda^{\alpha}x^{\alpha} = \lambda^{\alpha} f(x). \tag{2.6}$$

Ou seja, reescalar x por uma constante multiplicativa leva a uma reescala de f por outra constante. Uma discussão sobre aspectos matemáticos e estatísticos das leis de potência será apresentada na Seção 2.6. Aqui optaremos por não discutir A, apresentaremos a Equação 2.5 como

$$y \sim x^{\alpha} \tag{2.7}$$

e diremos que y escala com x com um expoente α. Se tomarmos, por exemplo, x como uma medida do tamanho de um sistema, a expressão 2.7 nos informa como este sistema responde quando seu tamanho muda. Faremos também a escolha de apresentar em escala duplo-logarítmica todos os gráficos associados às leis de escala aqui discutidas. Dessa forma, as funções do tipo 2.5 serão visualizadas como segmentos de retas nos gráficos discutidos nos capítulos subsequentes.

A busca pelas raízes do uso de hipóteses de escala nos leva à origem da ciência moderna, mais especificamente até a obra galileliana *Duas Novas Ciências* (GALILEI, s/d). Ali, pela voz de Salviati, Galileu aponta a impossibilidade física de aumentar o tamanho das estruturas "não somente na arte, mas também na natureza (...) até dimensões enormes" e justifica essa conclusão com argumentos de escala e resistência de materiais. Duas décadas atrás, Peterson propôs que o papel crucial que as leis de escala têm no último livro escrito por Galileu se deve a duas palestras acerca da geometria do inferno de Dante proferidas pelo jovem Galileu quase cinquenta anos antes na Academia Florentina e que teriam permanecido na mente do autor até o fim de sua vida (PETERSON, 2002).

A partir da segunda metade do século 19, tanto nas ciências naturais quanto nas ciências sociais, muitas distribuições do tipo lei de escala foram observadas. Algumas dessas observações empíricas se tornaram leis epônimas dentre as quais aquela proposta por Vilfredo Pareto é certamente uma das mais conhecidas (CIRILLO, 1978). Outra lei epônima, e que guarda uma relação com a lei de Pareto, é a Lei de Zipf que será discutida na Seção 2.5. Ainda da primeira metade do século 20 é possível destacar os trabalhos de Yule sobre o tamanho de uma família de plantas e duas famílias de besouros (YULE, 1925) e o trabalho de Lotka sobre publicações científicas (LOTKA, 1926).

Foi também naquela época que um importante uso de leis de escala na biologia emergiu: o estudo de como a taxa metabólica, mensurada pela potência basal P, varia com o tamanho do corpo de animais, mensurado pela massa M (SCHMIDT-NIELSEN, 1984). No final do século 19, Rubner propôs que $P \sim M^{2/3}$ (RUBNER, 1883), no entanto, cinco décadas depois, Kleiber publicou os resultados que apontaram que $P \sim M^{3/4}$ (KLEIBER, 1932). A fractalidade e a invariância de tamanho do ramo final do sistema vascular compõem junto com a minimização da energia o tripé do mecanismo apresentado em 1997 por West, Brown e Enquist para explicar tanto a lei de Kleiber quanto outras relações alométricas do tipo $Y \sim M^{b}$, com b sendo um múltiplo de 1/4 (WEST; BROWN; ENQUIST, 1997).

Um recorrente uso do conceito de escala em estudos ecológicos é a relação espécies-área (SAR na sigla em inglês) (ROSENZWEIG, 1995; STORCH; MARQUET; BROWN, 2007). Arrhenius, um século atrás, propôs uma SAR do tipo lei de potência (ARRHENIUS, 1921). Sucintamente temos

$$S \sim A^z \tag{2.8}$$

onde S é o número de espécies em uma área A. Diante de cenários de perda de biodiversidade, a relação espécies-área pode ser utilizada para estimar a extinção de espécies devido à perda de habitat. Um equivalente linguístico da lei de escala 2.8 desempenhará um papel crucial nos três capítulos seguintes.

É importante observar que, para além do estudo de sistemas complexos, transições de fase e fenômenos críticos, equações como a lei da gravitação universal de Newton e a lei de Stefan-Boltzmann são exemplos clássicos de leis de escala. Contemporaneamente, a observação de leis de escala se difundiu entre diversas áreas. Distribuições desse tipo são reportadas desde o estudo de cidades até investigações sobre número de acessos à páginas na rede mundial de computadores passando por estudos de geomorfologia e ciências cognitivas. Geoffrey West fornece no livro *Scale* uma introdução acessível ao tema (WEST, 2017) enquanto discussões mais aprofundadas são apresentadas em revisões publicadas nas duas últimas décadas (NEWMAN, 2005; CLAUSET; SHALIZI; NEWMAN, 2009; PINTO; LOPES; MACHADO, 2012).

2.5 Leis de escala em problemas envolvendo aspectos linguísticos

George Kingsley Zipf é o responsável pela lei epônima mais conhecida dos estudos de leis de escala e linguística. É digno de nota, entretanto, observar que as distribuições zipfianas transcendem os estudos linguísticos (NEWMAN, 2005). Assim como é necessário registrar que distribuições de frequência similares àquelas reportadas por Zipf foram observadas de modo independente por Auerbach (AUERBACH, 1913) e Estoup (ESTOUP, 1916) duas décadas antes da publicação de *The Psycho-Biology of Language* em 1935.

Analisando peças escritas em latim por Plauto, amostras de discursos em chinês da região de Pequim e trechos de jornais em inglês, Zipf observou que à medida que o número de ocorrências aumentava, o número de palavras diferentes que possuíam esse número de ocorrências diminuía (ZIPF, 1965). Ao construir gráficos duplo-logarítmicos a partir dos dados tabelados, Zipf reportou que sendo a o número de palavras de uma determinada frequência b, então

$$ab^2 = k \tag{2.9}$$

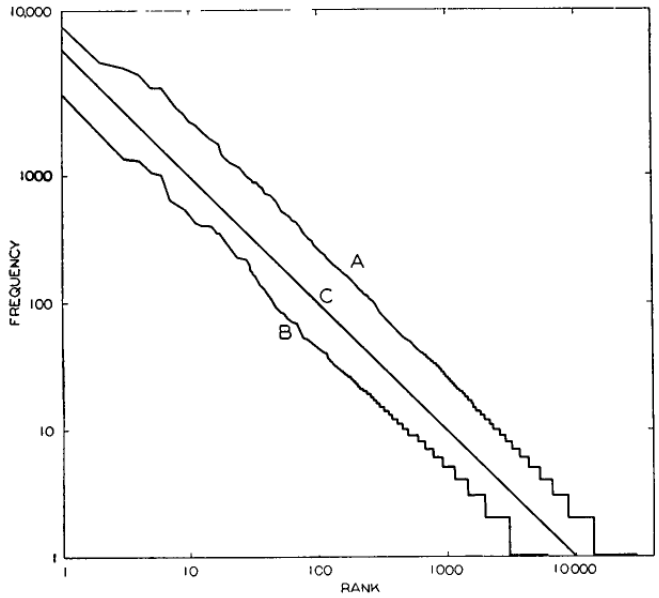
onde k seria uma constante para maioria das diferentes palavras do vocabulário em uso. Na década seguinte, Zipf apresentou uma formulação diferente dessa relação (ZIPF, 1949). Em *Human Behaviour and the Principle of Least Effort*, ele mostrou que assumindo como

f a frequência de uma palavra e como r a classificação desta palavra na tabela descrescente de frequências, então

$$r \times f = C \tag{2.10}$$

com C igual a uma constante (Figura 4).

Figura 4 – Frequência f como função da classificação r para palavras extraídas (A) do livro Ulisses, (B) de uma coleção de textos de jornais norte-americanos e (C) a função $f \sim r^{-1}$.



Fonte: *Human Behavior and the Principle of Least Effort* (ZIPF, 1949).

Observando as duas relações reportadas por Zipf notamos que $b = f$. Tomando como n o número de palavras de uma determinada frequência, podemos escrever a Equação 2.9 como

$$n \sim f^{-\alpha} \tag{2.11}$$

e a Equação 2.10 como

$$f \sim r^{-\beta}. \tag{2.12}$$

Como a classificação r de uma palavra com frequência f pode ser entendida como o número de palavras que ocorrem pelo menos f vezes, podemos escrever

$$r = \sum_{f'=f}^{\infty} n(f') \approx \int_{f}^{\infty} n(f')df'. \tag{2.13}$$

Com as Equações 2.11 e 2.12, temos

$$f^{-\frac{1}{\beta}} \approx \int_{f}^{\infty} f'^{-\alpha} df' \tag{2.14}$$

e assim

$$\beta = \frac{1}{\alpha - 1}. \tag{2.15}$$

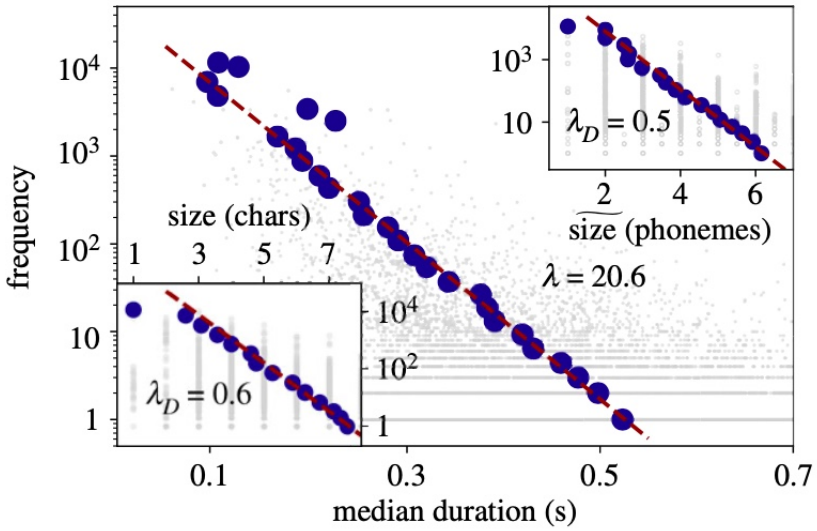
De modo que $\beta = 1$ quando $\alpha = 2$, ou seja, os expoentes reportados por Zipf em 1949 e 1935.

Zipf, também reportou que quanto maior a frequência de uma palavra menor seria o número de caracteres constituintes dela (ZIPF, 1965). Essa tendência das unidades linguísticas mais abundantes serem mais curtas, embora não tenha tido uma formulação quantitativa por Zipf, recebe o nome de Lei da brevidade ou Lei da abreviatura de Zipf. Recentemente, Torre et al. (TORRE et al., 2019) propuseram a seguinte formulação para a Lei da brevidade

$$f \sim D^{-\lambda_D l} \tag{2.16}$$

onde f é a frequência do elemento linguístico (palavra ou fonema, por exemplo), l é o tamanho do elemento linguístico (medido, por exemplo, em tempo médio de duração, número de fonemas ou número de caracteres) e D é o tamanho do alfabeto do nível linguístico de interesse (número de diferentes caracteres ou fonemas do idioma, por exemplo). A Figura 5 apresenta um exemplo desta lei obtida a partir de um banco de dados de registros de fala coloquial.

Figura 5 – Frequência de cada palavra (pontos cinza) do banco de dados *Buckeye* em função do seu tempo médio de duração (em segundos). Os círculos azuis são o resultado da aplicação de *binning* logarítmico às frequências. O painel superior direito mostra a mesma relação, mas considerando o número médio de fonemas por palavra, enquanto o painel inferior esquerdo representa o número de caracteres por palavra.



Fonte: *On the physical origin of linguistic laws and lognormality in speech* (TORRE et al., 2019).

No início da década de 50 do século passado, Claude Shannon usou a distribuição de frequência reportada por Zipf para estimar a entropia do inglês (SHANNON, 1951). Na mesma década, outra lei de escala observada em textos foi reportada por Gustav Herdan

(HERDAN, 1958). Dado os desenvolvimentos posteriores, a lei que conecta o vocabulário V, ou seja, o número de palavras distintas, com o número total N de palavras em um texto ficou conhecida como Lei de Heaps (HEAPS, 1978) e é escrita como

$$V \sim N^{\lambda} \tag{2.17}$$

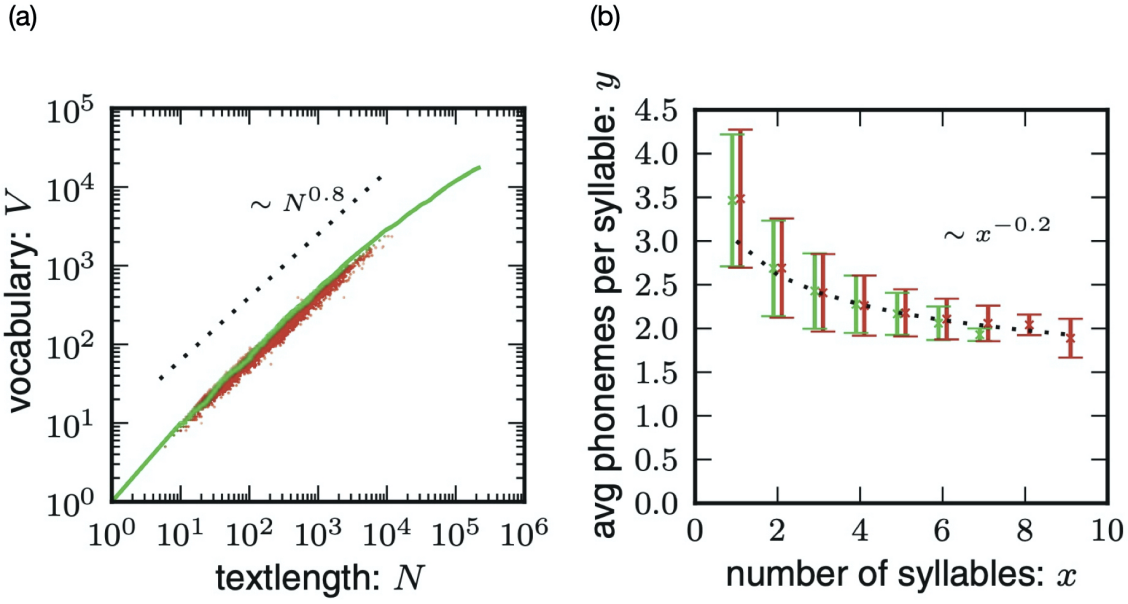
onde λ é nomeado expoente de Heaps. Para um exemplo, ver a Figura 6 (a).

Uma outra lei de escala recorrente nos estudos linguísticos é baseada originalmente nas observações de Paul Menzerath sobre fonemas e comumente exposta segundo o lema "quanto maior o todo, menores são as suas partes". A Lei de Menzerath-Altmann (ALTMANN, 1980) pode ser escrita como

$$y \sim x^{b} exp(-cx) \tag{2.18}$$

onde x é um construto linguístico (número de sílabas, por exemplo) e y é o tamanho de seus constituintes (fonemas por sílaba, por exemplo). Para um exemplo, ver a Figura 6 (b). Junto com as quatro leis supracitadas, Eduardo G. Altmann e Martin Gerlach listaram outras leis estatísticas associadas aos estudos linguísticos (ALTMANN; GERLACH, 2016). Embora os estudos de tais leis de escala tenham se concentrado predominantemente na expressão escrita da linguagem, regimes de escala similares são reportados para linguagem falada (BIAN et al., 2016; TORRE et al., 2019).

Figura 6 – (a) Vocabulário V como função do tamanho N do texto e (b) número médio de fonemas por sílaba y como função do número de sílabas x. Dados obtidos do livro Moby Dick (pontos verdes) e da Wikipedia em inglês (pontos vermelhos). As linhas pontilhadas são a lei de Heaps (a) e Menzerath-Altmann (b).



Fonte: *Statistical laws in linguistics* (ALTMANN; GERLACH, 2016).

Leis de escala foram também encontradas no estudo da distribuição de idiomas vivos e reportadas originalmente duas décadas atrás por Gomes et al. (GOMES et al., 1999). Similar a supracitada *SAR* (Seção 2.4), uma relação de escala entre a diversidade linguística D e a área A do país foi encontrada de modo que

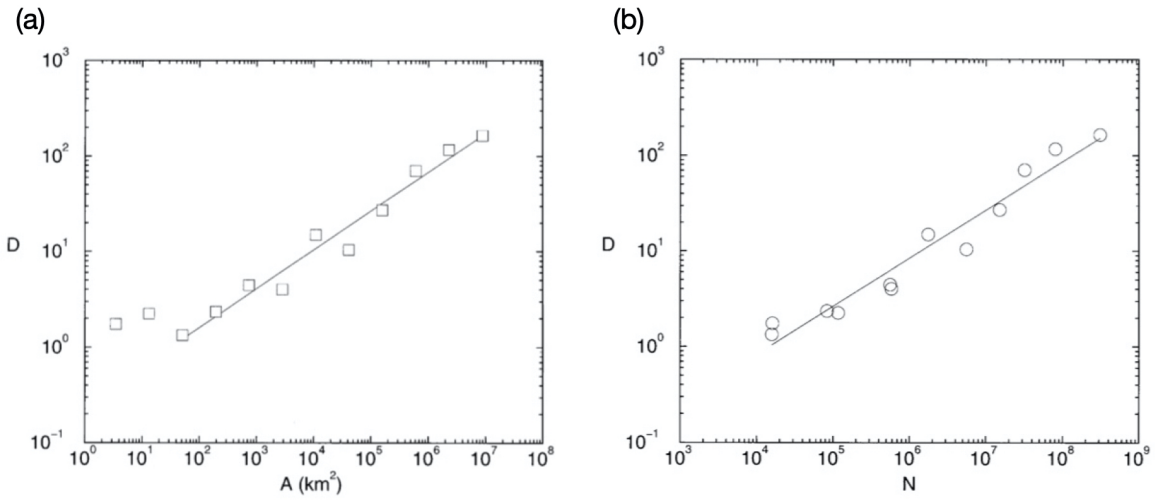
$$D \sim A^z \tag{2.19}$$

com $z = 0,41 \pm 0,03$ conforme a Figura 7 (a). Já ao estudar a relação entre a diversidade linguística D e a população N do país foi observado que

$$D \sim N^\nu \tag{2.20}$$

com $\nu = 0,50 \pm 0,04$ conforme a Figura 7 (b). Essa análise será revisitada e os resultados serão atualizados no Capítulo 3.

Figura 7 – (a) Diversidade linguística média D em função da área A do país. A linha reta é o melhor ajuste cuja inclinação fornece o expoente $z = 0,41 \pm 0,03$. (b) Diversidade linguística média D em função da população N do país. A linha reta é o melhor ajuste cuja inclinação fornece o expoente $\nu = 0,50 \pm 0,04$.



Fonte: *Scaling relations for diversity of languages* (GOMES et al., 1999).

Quando investigada a distribuição da diversidade linguística entre os vários países, foi observado um regime de dupla lei de escala. Sendo $N(> D)$ o número de países com uma diversidade linguística maior que D, demonstrou-se, conforme Figura 8 (a), que

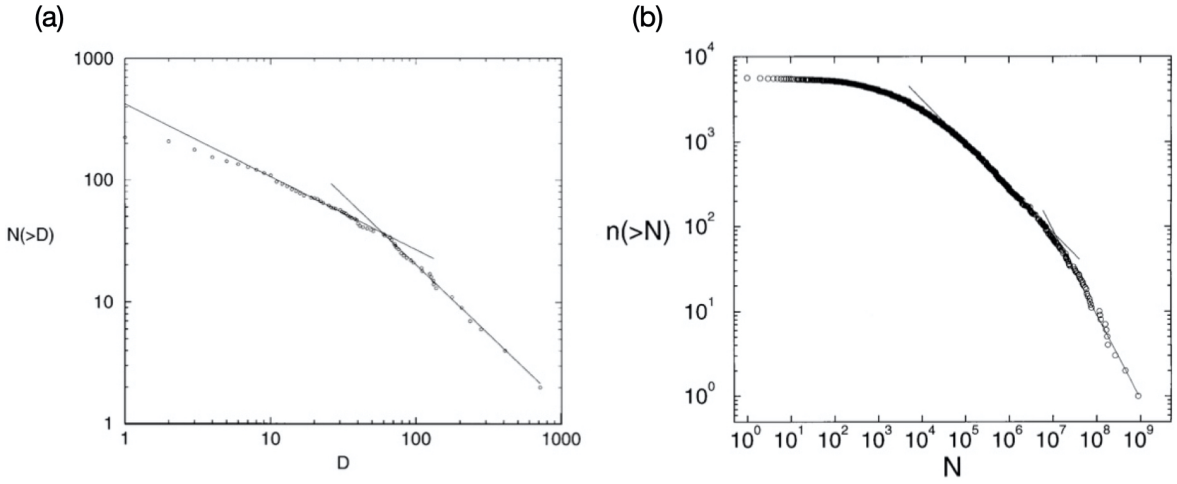
$$N(> D) \sim D^{-B} \tag{2.21}$$

com $B = 0,6$ para $6 < D < 60$ e $B = 1,1$ para $60 < D < 700$. A distribuição de tamanhos dos idiomas, ou população linguística, também apresentou um regime de dupla lei de escala. Adotando como $n(> N)$ o número de idiomas com uma população maior que N foi observado que

$$n(> N) \sim N^{-\tau} \tag{2.22}$$

com $\tau = 0,5$ para $5 \times 10^4 < N < 6 \times 10^6$ e $\tau = 1,0$ para $2 \times 10^7 < N < 1 \times 10^9$ como visto na Figura 8 (b). O Capítulo 6 apresentará uma análise atualizada dessas grandezas.

Figura 8 – (a) Número de países $N(> D)$ com diversidade linguística maior que D em função de D. As linhas retas fornecem os expoentes $B = 0,6$ para $6 < D < 60$ e $B = 1,1$ para $60 < D < 700$. (b) Número de idiomas $n(> N)$ com população maior que N em função de N. As linhas retas fornecem os expoentes $\tau = 0,5$ para $5 \times 10^4 < N < 6 \times 10^6$ e $\tau = 1,0$ para $2 \times 10^7 < N < 1 \times 10^9$.



Fonte: *Scaling relations for diversity of languages* (GOMES et al., 1999).

2.6 Breves apontamentos sobre leis de potência

A Lei de Zipf (Equação 2.9) pode ser generalizada e escrita como

$$p(x) = Cx^{-\alpha} \tag{2.23}$$

com $p(x)$ sendo a frequência da variável x e C uma constante. Quando x se estende por muitas décadas, se mostra útil fazer uma binarização logarítmica dos dados antes da construção dos gráficos. Dessa forma cada caixa é criada de modo que sua largura seja um múltiplo fixo e positivo da largura da caixa anterior. Tomando por a esse múltiplo, temos que a n-enésima caixa se estende de $x_{n-1} = x_{min}a^{n-1}$ até $x_n = x_{min}a^n$ e o número esperado de amostras contidas nesse intervalo é dado por

$$\int_{x_{n-1}}^{x_n} p(x)dx = C\int_{x_{n-1}}^{x_n} x^{-\alpha}dx = C\frac{(a^{\alpha-1}-1)}{\alpha-1}(x_{min}a^n)^{1-\alpha}. \tag{2.24}$$

Podemos também fazer o estudo da distribuição acumulada $P(x)$, ou seja, a probabilidade $P(x)$ de que x tenha um valor maior ou igual a x:

$$P(x) = \int_x^\infty p(x')dx'. \tag{2.25}$$

Dada a Equação 2.23, temos

$$P(x) = C \int_x^\infty x'^{-\alpha} dx' = \frac{C}{\alpha - 1} x^{-(\alpha-1)}. \tag{2.26}$$

Portanto observamos que a distribuição acumulada também é descrita como uma lei de potência tendo neste caso expoente $\alpha - 1$. Ademais, a distribuição acumulada apresenta a vantagem de poder ser construída sem a necessidade da binarização discutida acima. Já o valor da constante C pode ser obtido segundo a normalização

$$\int p(x)dx = 1. \tag{2.27}$$

Assumindo que a distribuição 2.23 é válida a partir x_{min}, temos

$$1 = C \int_{x_{min}}^\infty x^{-\alpha} dx = C \frac{1}{1-\alpha} \left(x^{-\alpha+1} \right)_{x_{min}}^\infty . \tag{2.28}$$

Com $\alpha > 1$, de modo que o resultado da integral não divirja, temos

$$C = (\alpha - 1) x_{min}^{\alpha-1}. \tag{2.29}$$

Assim, podemos escrever a Equação 2.23 como

$$p(x) = \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-\alpha}. \tag{2.30}$$

2.6.1 Leis de potência para variáveis discretas

Embora grande parte da discussão realizada anteriormente também seja válida para quantidades discretas, é necessário definir a distribuição da lei de potência de uma maneira ligeiramente diferente. Nesse caso, definimos que a probabilidade p_k de medir o valor discreto k obedece

$$p_k = C' k^{-\beta}, \tag{2.31}$$

para um expoente constante β. Com o caso $k = 0$ descartado para evitar que essa distribuição divirja.

O valor da constante C' pode ser obtido de acordo com a normalização

$$\sum_{k=1}^{\infty} p_k = 1, \tag{2.32}$$

ou seja,

$$1 = C' \sum_{k=1}^{\infty} k^{-\beta}. \tag{2.33}$$

A soma do lado direito da equação pode ser reconhecida como a função zeta de Riemann (EDWARDS, 2001), tal que

$$\zeta(\beta) = \sum_{k=1}^{\infty} k^{-\beta}. \tag{2.34}$$

Portanto,

$$C' = \frac{1}{\zeta(\beta)} \tag{2.35}$$

e assim podemos escrever a Equação 2.31 como

$$p_k = \frac{k^{-\beta}}{\zeta(\beta)}. \tag{2.36}$$

Além da Equação 2.31, existem formas alternativas de descrever leis de potência para variáveis discretas como, por exemplo, aquelas que fazem uso de funções Beta (NEWMAN, 2005).

A discussão acima se aplica às quantidades analisadas a partir do próximo capítulo, como diversidade linguística, populações por país e por idioma, áreas e Produto Interno Bruto que são, como veremos, grandezas discretas e, em geral, inteiras.

3 PODE A DIVERSIDADE LINGUÍSTICA CONDICIONAR ASPECTOS ECONÔMICOS?

"Kolik jazyků umíš,
tolikrát jsi člověkem."[1]

Provérbio checo

Neste capítulo, uma análise estatística detalhada mostra que, além da bem estabelecida relação diversidade linguística-área (GOMES et al., 1999), há uma notável relação de escala entre o tamanho da economia de um país, medido pelo Produto Interno Bruto (PIB), e o número de idiomas falado em seu território. Entre essas duas análises, é discutido como a diversidade linguística está relacionada com o tamanho da população de cada país. Mais exatamente são apresentadas leis de escala alométricas entre a diversidade linguística e o (*i*) tamanho geográfico quantificado pela área, (*ii*) tamanho demográfico quantificado pela população e (*iii*) tamanho econômico quantificado pelo PIB dos países. Ademais, introduzimos um modelo fractal que justifica o expoente $z = 1/3$ da lei de escala observada entre a diversidade linguística e a área. Um subgrupo dos resultados aqui discutidos foi publicado no artigo *"Revisiting scaling relations for linguistic diversity"* (SANTOS; GOMES, 2019) disponível no Apêndice A.

3.1 Diversidade e área territorial

Charles Darwin escreveu em *The descent of man, and selection in relation to sex* que "The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel."[2] (DARWIN, 1871). De fato, tanto a dinâmica quanto a distribuição espacial dos idiomas têm paralelos com aquelas apresentadas por espécies biológicas (MUFWENE, 2001; WHITFIELD, 2008; PAGEL, 2009; BURNSIDE et al., 2012; UPADHYAY; HASNAIN, 2017; HUA et al., 2019). A coocorrência de diversidade linguística e biológica nas mesmas áreas geográficas parece sugerir que algumas regras de organização sejam compartilhadas (MOORE et al., 2002; PAGEL; MACE, 2004; GORENFLO et al., 2012; CARDILLO; BROMHAM; GREENHILL,

[1] "Quanto mais línguas você fala, mais vezes você é humano."

[2] "A formação de diferentes línguas e de espécies distintas, e as provas de que ambas se desenvolveram por meio de um processo gradual, são curiosamente paralelas."

2015), por exemplo a relação número de espécies (S) - área (A), $S \sim A^z$, (ROSENZWEIG, 1995). Idiomas e espécies também têm similar correlação entre o risco de extinção e a distribuição geográfica (SUTHERLAND, 2003). Relações tão próximas entre a diversidade biológica e a diversidade cultural, onde os idiomas são apenas uma parte desse conjunto mais amplo, resultaram no desenvolvimento do conceito de diversidade biocultural (LOH; HARMON, 2005; MAFFI, 2005). Convém observar que, paralelamente ao aparecimento do termo diversidade linguística em diversos trabalhos ao longo das últimas décadas, surgiram diferentes modos de quantificar essa diversidade.

Utilizamos os dados de diversidade linguística da vigésima edição do *Ethnologue* (SIMONS; FENNIG, 2017), publicada em 2017, que lista mais de 7000 idiomas falados em 236 países. Ou seja, aqui atualizamos a análise realizada por Gomes e colaboradores (GOMES et al., 1999), após 20 anos, intervalo de tempo em que não deixaram de ocorrer algumas alterações no *Ethnologue* (para detalhes ver a Seção 1.1). O trabalho realizado no final do século passado utilizou a décima terceira edição *Ethnologue* que trazia um pouco mais de quatrocentos idiomas a menos do que o número total de idiomas listados na edição utilizada na presente análise.

O *Ethnologue* fornece para cada país a diversidade linguística total correspondente D, que é definida como o número de idiomas vivos utilizados como primeiro idioma. Segundo essa definição, o país com maior diversidade linguística é a Papua Nova Guiné, com 840 idiomas, seguida pela Indonésia e Nigéria, com 709 e 527 idiomas, respectivamente. A Nova Guiné, especificamente, foi nos últimos anos objeto de vários estudos (GORENFLO et al., 2012; TURVEY; PETTORELLI, 2014). O fato desses países não terem grandes valores de áreas não altera significativamente a análise estatística apresentada a seguir que envolve um conjunto muito maior de países.

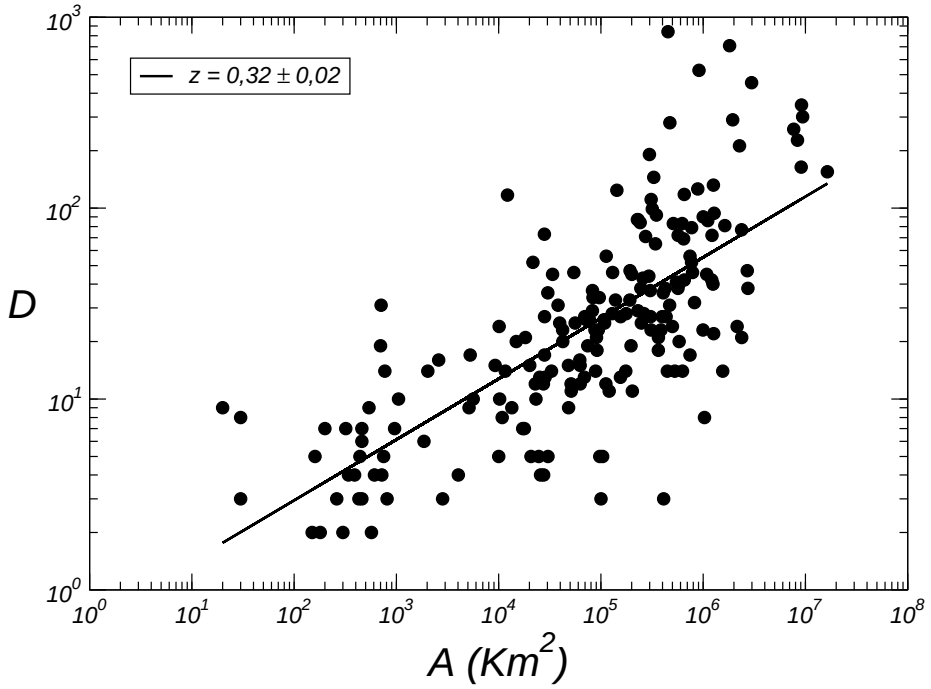
Inicialmente, analisamos a relação entre a diversidade linguística D e a área terrestre A (em quilômetros quadrados) dos países. Na Figura 9 apresentamos em escala duplo-logarítmica a distribuição $D \times A$. Embora seja um subgrupo do total de 236 países listados pelo *Ethnologue*, o conjunto de países analisados aqui é responsável por mais de 98% da população mundial e mais de 97% da área superficial terrestre. Como frequentemente observado com espécies biológicas, a relação entre diversidade linguística D e área territorial A, como mostrado na Figura 9, é bem descrita pela lei de escala

$$D \sim A^z \tag{3.1}$$

onde o expoente $z = 0,32 \pm 0,02$. Esse expoente é um pouco menor do que o apresentado por Gomes et al. (GOMES et al., 1999), $z = 0,41 \pm 0,03$, mas igual dentro da incerteza estatística ao valor relatado por Loh e Harmon (LOH; HARMON, 2005).

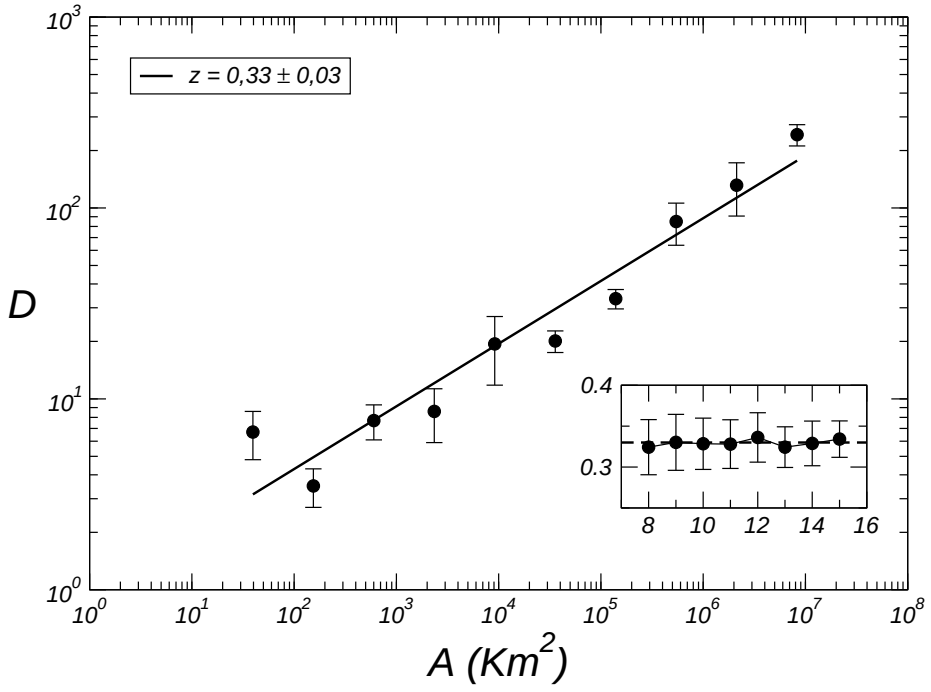
Uma linha de tendência mais clara surge após a categorização dos países em dez grupos de acordo com a área e com o subsequente cálculo da diversidade média de idiomas

Figura 9 – Diversidade linguística D em função da área territorial A do país para os dados não categorizados ($n = 194$). A linha contínua é o melhor ajuste para $D \sim A^z$ e sua inclinação fornece o expoente $z = 0,32 \pm 0,02$ ($r = 0,73$).



Fonte: Autor (2021).

Figura 10 – Diversidade linguística D em função da área A para os dados agrupados. A linha contínua é o melhor ajuste para $D \sim A^z$ e sua inclinação fornece o expoente $z = 0,33 \pm 0,03$ ($r = 0,97$). Inserção: valor do expoente z em função do número de categorias.



Fonte: Autor (2021).

vivos em cada grupo conforme apresentado na Figura 10. Após essa separação, o expoente obtido foi $z = 0,33 \pm 0,03$ em uma relação de escala que se estende por mais de cinco décadas. Ou seja, a lei de escala é robusta com relação a esse tipo de agrupamento dos dados. Uma possível dependência entre o expoente z e a quantidade de categorias utilizadas foi analisada. Ao variar o número de categorias entre oito e quinze, a inserção na Figura 10 aponta que o valor do expoente z também é independente da quantidade de categorias. A concordância entre o resultado sem categorização (Figura 9) e os resultados obtidos a partir dos dados categorizados (Figura 10), atesta, consequentemente, a robustez da relação de escala entre a diversidade linguística e a área territorial dos países. O valor obtido para o expoente z lembra aqueles observados em ecologia para a relação entre diversidade de espécies e área (HOBOHM, 2003; DESMET; COWLING, 2004).

De posse do valor do expoente z é possível investigar a densidade de idiomas vivos ρ_D, definida como

$$\rho_D = \frac{D}{A}. \tag{3.2}$$

De acordo com a dependência supracitada (Equação 3.1), temos

$$\rho_D \sim A^{-z'} \tag{3.3}$$

onde $z' = 1 - z$ de modo que $z' = 0,67$. Um valor igual de z', dentro da incerteza, também pode ser obtido diretamente a partir dos dados agrupados conforme apresentado na Figura 11. Tais resultados levam à constatação de que áreas maiores têm proporcionalmente menor densidade de idiomas, ou seja, um efeito de rarefação no número de idiomas falados à medida que a área de amostragem cresce, característico da simetria de dilatação de um sistema fractal.

3.2 Diversidade e população

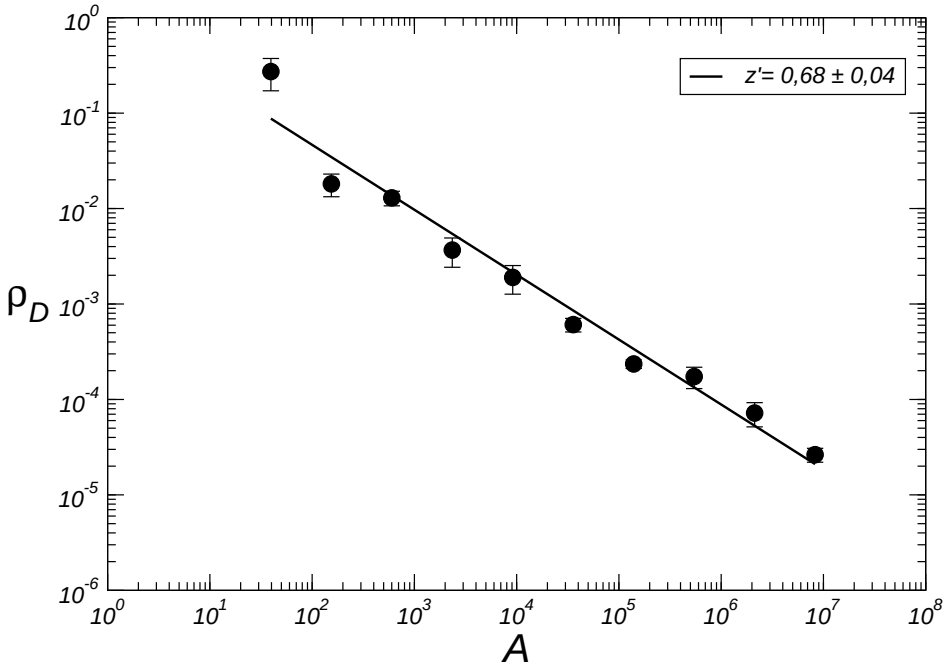
A população foi adotada como parâmetro demográfico a ser investigado em relação à diversidade linguística. Adotando a estratégia apresentada na seção anterior, a relação entre diversidade linguística D e população N, como mostrado na Figura 12, também é bem descrita por uma dependência de escala do tipo de lei de potência

$$D \sim N^{\nu}. \tag{3.4}$$

Relações semelhantes entre diversidade de tamanhos de fragmentos e tamanho da população foram relatadas em fenômenos de fragmentação (COUTINHO; GOMES; ADHIKARI, 1992).

O ajuste apresentado na Figura 12 fornece $\nu = 0,39 \pm 0,03$. Adotando uma categorização que separe os 194 países em 10 grupos segundo suas populações, é possível observar uma linha de tendência mais clara. A Figura 13 tem como melhor ajuste à Equação

Figura 11 – Densidade de idiomas vivos ρ_D em função da área do país para os dados agrupados. A linha contínua é o melhor ajuste e sua inclinação fornece o expoente $z' = 0,68 \pm 0,04$ ($r = 0,98$).



Fonte: Autor (2021).

3.4 a curva que fornece $\nu = 0,41 \pm 0,03$. Uma vez mais, conforme inserção da Figura 13, este valor de expoente não está condicionado ao número de categorias utilizadas na separação dos países, embora as flutuações no expoente ν sejam um pouco maiores do que as observadas no expoente z. A extrapolação do referido ajuste permite descobrir qual o tamanho de população associado à existência de um idioma individualmente. Segundo essa análise, uma população da ordem de 500 pessoas definiria o conjunto mínimo relacionado à existência de um idioma.

De acordo com as Equações 3.1 e 3.4, temos que a população cresce com área do país segundo

$$N \sim A^{z/\nu}. \tag{3.5}$$

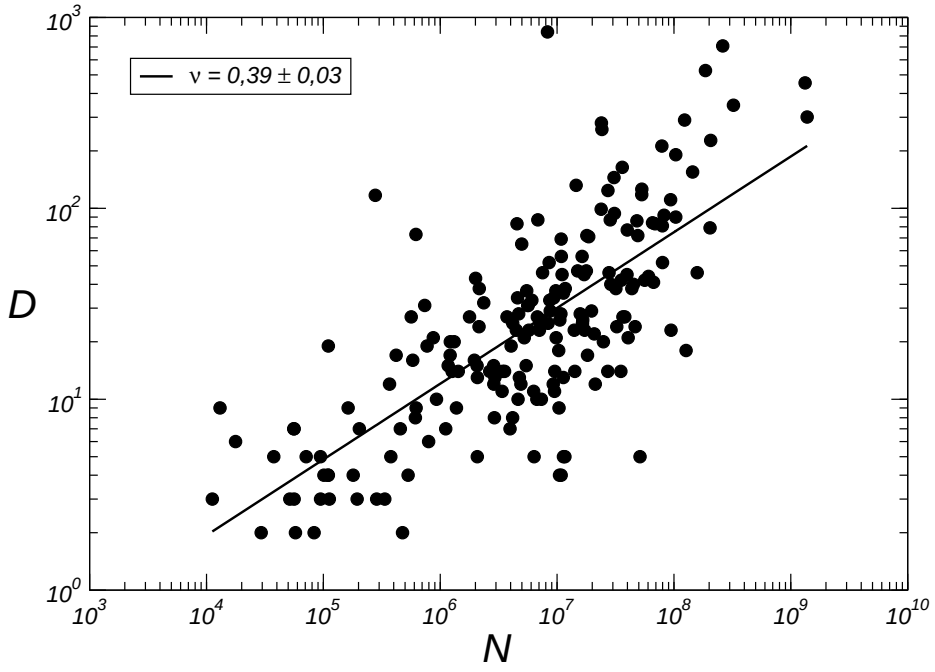
Essa relação sugere que os grupos populacionais seguem uma distribuição fractal. De acordo com os valores dos expoentes z e ν obtidos anteriormente, temos $z/\nu = 0,80 \pm 0,09$. Este valor é igual dentro da incerteza estatística ao obtido diretamente a partir dos dados de área e população dos países (não mostrado) e aquele reportado por Zhang e Yu (ZHANG; YU, 2010). Sendo a densidade populacional da Terra ρ_N, definida como

$$\rho_N = \frac{N}{A}, \tag{3.6}$$

constata-se que sobre a superfície terrestre a densidade populacional não é constante mas segue a relação de escala

$$\rho_N \sim A^{-0,20}, \tag{3.7}$$

Figura 12 – Diversidade linguística em função da população do país para os dados não categorizados ($n = 194$). A linha contínua é o melhor ajuste e sua inclinação fornece o expoente $\nu = 0,39 \pm 0,03$ ($r = 0,73$).



Fonte: Autor (2021).

ou seja, obedece também a uma distribuição hiperbólica, como encontrado para a densidade de idiomas na Equação 3.3. Paralelamente, podemos observar que sendo $\langle R^2 \rangle$ o raio quadrático médio associado à área segundo $A \sim \langle R^2 \rangle$ e tomando a relação 3.5, temos

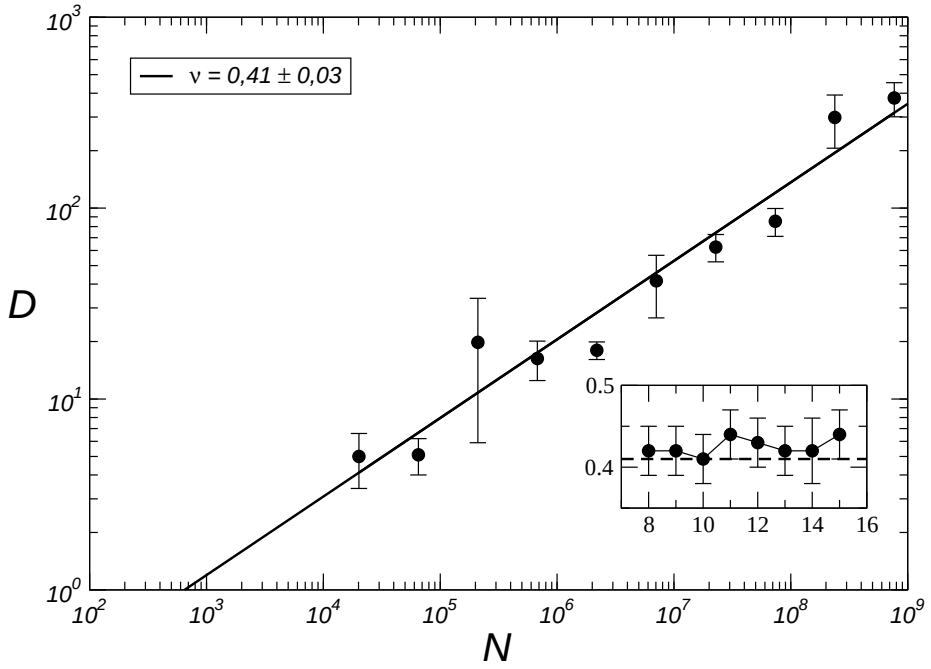
$$N \sim \mathcal{R}^{\delta}, \tag{3.8}$$

onde $\delta = 2z/\nu = 1,6 \pm 0,2$ e $\mathcal{R} = \langle R^2 \rangle^{1/2}$. Este valor de δ é igual à dimensão fractal do esqueleto (*backbone*) do aglomerado de percolação em duas dimensões, $D_B = 1,60 \pm 0,05$ (HERRMANN; HONG; STANLEY, 1984). Este resultado aponta que a população humana está distribuída sobre a superfície terrestre em um conjunto fractal de dimensão $\delta = 1,6$. De igual modo, quando observada a relação 3.1 temos que diversidade linguística escala com $\mathcal{R}$ de acordo com

$$D \sim \mathcal{R}^{d'} \tag{3.9}$$

com $d' = 2z = 0,66 \pm 0,06$. Portanto, os idiomas estão distribuídos sobre a superfície da Terra em um conjunto fractal Δ de dimensão $2/3$, dentro de incertezas estatísticas $\leq 10\%$. Este valor do expoente d' ou, equivalentemente, do expoente z, pode ser entendido a partir de um modelo fractal simples, como explicado na próxima seção.

Figura 13 – Diversidade linguística em função da população para os dados categorizados. A linha tracejada é o melhor ajuste e cuja inclinação fornece o expoente $\nu = 0,41 \pm 0,03$ ($r = 0,97$). Inserção: valor do expoente ν em função do número de categorias.



Fonte: Autor (2021).

3.3 Um modelo fractal justificando o expoente z

Aqui, um modelo fractal simples (SANTOS; GOMES, 2021b) com significado biológico é introduzido para explicar a lei de escala observada entre a diversidade linguística dentro de uma área geográfica discutida na Seção 3.1. Supõe-se que o modelo possa ser relevante para explicar alguns aspectos das curvas de diversidade de espécies em estudos ecológicos (DURRETT; LEVIN, 1996; PLOTKIN et al., 2000). Para tal discussão convém recordar que, conforme discutido na seção anterior, o conjunto de idiomas se distribui em um conjunto fractal Δ de dimensão $d' = 2z = 0,66 \pm 0,06$.

Nosso modelo que justifica a lei de escala 3.1, baseia-se em três hipóteses simples e razoáveis que levam a uma entidade matemática muito estudada e complexa referente à estrutura geométrica, Σ, dos deslocamentos de populações migrantes primitivas espalhadas pela Terra por milênios, em particular após a revolução agrícola (RENFREW, 1989). Ou seja, Σ aqui é um objeto matemático que denota o tipo de trajetória seguida pelos migrantes. As trajetórias Σ incorporam simultaneamente (i) a busca por tentativa e erro de bons lugares para estabelecer um assentamento e, eventualmente, desenvolver

alguma plantação de subsistência e (ii) o cuidado de evitar o confronto com populações já estabelecidas, mantendo-se o mais distante possível. (iii) Além disso, durante essas incursões, os migrantes tendem a evitar, também, as regiões já visitadas.

Nossa escolha da estrutura matemática para levar em conta as propriedades listadas acima, a saber: (i) algum tipo de aleatoriedade, (ii) algum tipo de auto-exclusão e (iii) algum tipo de memória, é a caminhada aleatória auto-excludente ($SARW$ na sigla em inglês) em uma superfície bidimensional. A caminhada aleatória auto-excludente é uma estrutura com dimensão topológica igual à unidade e uma dimensão fractal que reflete a relação de escala assintótica entre o número de passos ($N \gg 1$) e distância ponta-a-ponta (R) da trajetória,

$$N \sim R^{d''}, \tag{3.10}$$

com sua dimensão fractal assumindo o valor $d'' \approx 1,33$ (DE GENNES, 1979; LI; MADRAS; SOKAL, 1995). Portanto, temos dois conjuntos fractais: Δ, representando a distribuição espacial da diversidade linguística, e Σ, $SARWs$, representando as trajetórias de pessoas migrantes ao longo dos milênios. A seguir, passamos a discutir as consequências das estruturas desses dois conjuntos fractais no problema da distribuição da diversidade linguística na Terra.

Matematicamente, dois conjuntos fractais Δ e Σ, respectivamente de dimensões d' e d'', mergulhados em um espaço d-dimensional não têm interseção se

$$d' + d'' \leq d \tag{3.11}$$

(LOVEJOY; SCHERTZER; LADOY, 1986; FALCONER, 2014). Esta regra geral de não intersecção de conjuntos é a mesma que diz, por exemplo, que duas linhas retas ($d' = d'' = 1$) dispostas aleatoriamente no plano ($d = 2$) têm uma interseção marginal de um único ponto, uma vez que o caso de linhas exatamente paralelas tem probabilidade de ocorrência nula. Portanto, podemos reformular nosso argumento dizendo que a distribuição de idiomas (dimensão d') em uma superfície bidimensional ($d = 2$) garantiria melhor sua sobrevivência contra agentes distribuídos em um conjunto com a dimensão d'', no mesmo espaço, se

$$d' \leq 2 - d'', \tag{3.12}$$

ou

$$z \leq \frac{2 - d''}{2}. \tag{3.13}$$

Essa condição dimensional simples leva a uma estabilidade razoável do sistema de idiomas falado por populações estabelecidas, *vis-à-vis* a distribuição de novas populações que chegam. No presente caso, mostrado na Figura 10 , obtemos a condição

$$d' + d'' = 0,66 + 1,33 = 1,99 \leq d = 2, \tag{3.14}$$

dentro de uma incerteza de $0,06$. Portanto, existe uma relação de complementaridade entre a estrutura fractal da distribuição de idiomas e a distribuição fractal associada às trajetórias das populações migrantes. Ou seja, podemos compreender a relação de escala atualmente observada para a diversidade linguística $D \sim A^z$ como uma relíquia de tempos primordiais em que populações humanas viajavam a pé por longas distâncias em busca de um bom local para viver e desenvolver sua agricultura, perfazendo caminhadas que se aproximavam de *SARWs* em duas dimensões. Para aqueles que considerarem este modelo "excessivamente simples", devemos lembrar que o *SARW*, particularmente em $d = 2$, é um dos objetos matemáticos mais complexos já imaginados. Assim, o modelo só aparenta ser simples pelo fato de incorporar um módulo que, em si, é excessivamente complexo. Em outras palavras, não fazemos nenhuma simplificação ao modelar a deambulação dos humanos ao longo de milênios por *SARWs*; pelo contrário, estas são, intrinsecamente, altamente complexas.

Além disso, conjecturamos que esse tipo de argumento pode ser útil em algumas situações de interesse biológico. Como exemplo, considere uma fauna com um predador que geralmente se move usando trilhas tortuosas cuja estrutura fractal se aproxima de um *SARW* ($d'' = 1,33$); neste caso, as várias espécies de presas são protegidas caso estejam distribuídas em um conjunto fractal cuja dimensionalidade é menor que $2 - 1,33 \leq 0,67$, ou seja, que o expoente z na *SAR* correspondente satisfaça $z \leq 0,335$. Quanto menor o expoente z, mais protegida a presa estará do ataque desses predadores. Para manter o equilíbrio na dinâmica predador-presa, no entanto, é possível permitir alguma intercessão entre esses dois conjuntos, de modo que valores um pouco maiores em comparação com $z = 0,33$ sejam possíveis. Como um exemplo adicional, para predadores que executam passeios aleatórios ($d'' = 2$), a condição limite prediz $d' = 0$, isto é, uma dependência logarítmica $S \sim LnA$ entre o número de espécies-presas e a área. Do fato de que todas as trajetórias têm dimensão topológica 1, segue-se que $d'' \geq 1$ e, daí, $d' \leq 2 - d'' \leq 1$ ou $z \leq 0,5$, como pode ser observado (veja (DURRETT; LEVIN, 1996; PLOTKIN et al., 2000; TRIANTIS; GUILHAUMON; WHITTAKER, 2012)).

3.4 Diversidade e Produto Interno Bruto

Abordaremos agora uma possível relação entre a diversidade linguística e o tamanho econômico de um país. Convém observar que com a proliferação de bancos de dados digitais, uma grande dose de ceticismo e crítica deve ser exercida ao examinar o crescente número de estudos que buscam relações quantitativas entre variáveis culturais, demográficas, econômicas e geográficas. Se isso não for feito, correlações positivas espúrias podem ser encontradas. Dentro dessa estrutura, Roberts e Winters apontaram o risco de se enfatizar demais essas correlações (ROBERTS; WINTERS, 2013). Para ilustrar, esses autores mostram que uma correlação positiva surge, por exemplo, entre a diversidade linguística

e o número percentual de mortes anuais no trânsito. Eles observam que, infelizmente, alguns desses estudos estão recebendo atenção da mídia sem um amplo entendimento da complexidade da questão e existe o risco de estudos mal controlados afetarem as políticas públicas. Apesar do risco inerente de entender mal a relação entre variáveis associadas à fenômenos complexos, algumas importantes relações socioeconômicas de escala são amplamente conhecidas (BROCK, 1999; GABAIX, 2009) e têm uma base teórica mais satisfatória, como exemplo, a lei de Pareto, que afirma que a distribuição de renda é muito desigual, com a população mais rica controlando a maior parte dos recursos (CIRILLO, 1978).

Por outro lado, olhando para um cenário muito distante, no início das concepções teóricas que definem o fundamento do mundo ocidental, pode ser oportuno recordar as palavras de abertura na Metafísica de Aristóteles "Todos os homens, por natureza, desejam conhecer." (ARISTÓTELES, 1969). Esse é um tipo de afirmação que significa que o fato de viver deve ser colocado como uma espécie de conhecimento. É tentador complementar esse pensamento amplamente conhecido com uma ideia de origem incerta, embora frequentemente associada ao filósofo romeno Emil Cioran, de que "On n'habite pas un pays, on habite une langue."[3] (CIORAN, 1987). Se houver alguma plausibilidade nessas duas declarações, será razoável concluir que um mundo com grande diversidade linguística representa um mundo com mais conhecimento. Por outro lado, a perda de diversidade linguística representa não apenas uma redução no repertório de visões de mundo, mas também uma redução nas informações biológicas, ecológicas, geográficas e tecnológicas, para citar apenas alguns aspectos (ROMAINE, 2007; AUSTIN; SALLABANK, 2011; WILDER et al., 2016).

Tendo investigado as leis de escala entre a diversidade linguística e o tamanho geográfico (área) e o tamanho demográfico (população) e considerando que a atividade econômica geralmente privilegia o conhecimento, somos levados a indagar se a diversidade linguística pode estar associada a algum indicador econômico básico. Anteriormente Pool (POOL, 1972) e Nettle (NETTLE, 2000) discutiram a relação entre diversidade linguística e um parâmetro econômico. É importante destacar as diferenças metodológicas e de conjunto de dados destes trabalhos em relação a análise que apresentaremos abaixo. Tanto Pool quanto Nettle utilizaram o Produto Interno Bruto *per capita* como parâmetro econômico enquanto nós adotamos o Produto Interno Bruto (PIB) que, apesar de possíveis ambiguidades metodológicas, tem sido, durante décadas, o indicador mais amplamente utilizado para medir a atividade econômica de qualquer país (WEIL; SHARMA, 2013). Como diversidade linguística, Pool e Nettle utilizaram o percentual da população que tinha como primeiro idioma aquele que era o mais falado no país, embora Nettle também tenha analisado o número de idiomas por milhão de pessoas. Quanto ao número de países,

[3] "Não se habita um país, habita-se uma língua"

Pool analisou 133 enquanto Nettle excluiu os países com menos de 10000 km^2, o que reduziu o conjunto de dados para apenas 107 países. Nossa análise inicial não possui essa restrição de área e, portanto, compreende quase o dobro do número de países. Ademais, não há reivindicações de leis de escala nesses artigos; os ajustes utilizados por esses autores sugerem uma dependência do tipo exponencial. Durante os anos recentes diversas investigações interdisciplinares seguiram a busca pelas conexões entre a diversidade de idiomas e elementos econômicos (GINSBURGH; WEBER, 2011). Ainda assim, de acordo com Desmet, Ortuño-Ortíz e Wacziarg "the relation between linguistic diversity and the level of economic development has been somewhat understudied."[4] (DESMET; ORTUÑO-ORTÍN; WACZIARG, 2016).

Os dados do PIB, relativos a 2016, estão em unidades de milhões de dólares e foram obtidos dos Indicadores de Desenvolvimento Mundial do Banco Mundial (informações detalhadas sobre a construção do banco de dados utilizado neste trabalho estão disponíveis na Seção 1.1). A razão de termos usado uma base de 194 países nas duas seções anteriores decorreu da limitação dos dados do $PIB - 2016$ que estavam disponíveis apenas para este conjunto menor de países quando comparados com aqueles relativos à diversidade linguística.

De maneira semelhante à abordagem discutida nas duas seções anteriores, propomos que a diversidade D escala com o PIB segundo

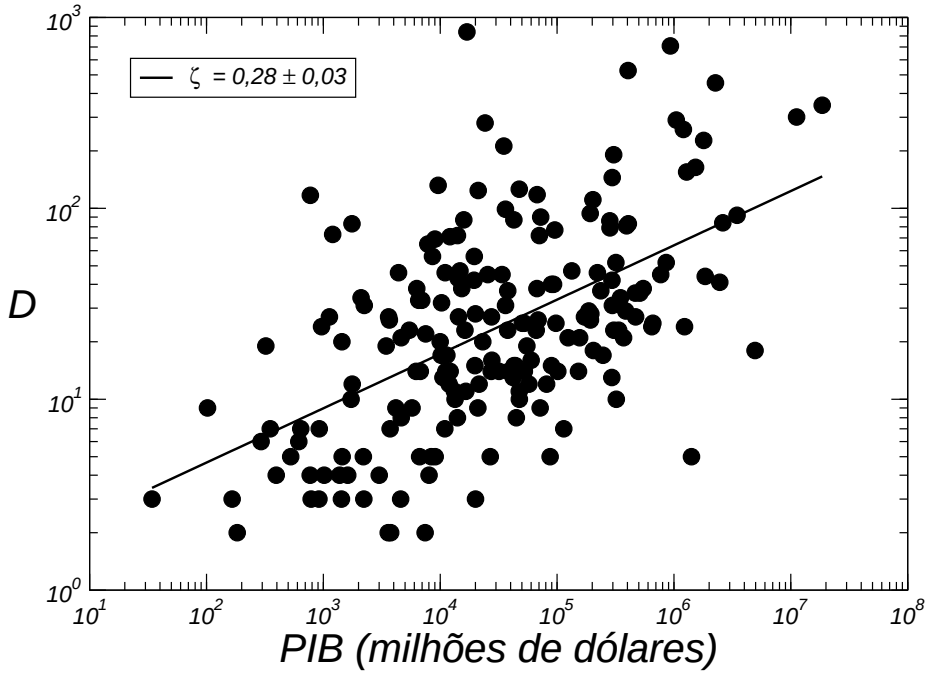
$$D \sim (PIB)^{\zeta}, \tag{3.15}$$

onde o expoente $\zeta = 0,28 \pm 0,03$ é obtido a partir do ajuste apresentado na Figura 14, para os dados não categorizados. Ao dividirmos os países em 10 categorias, desta vez de acordo com o PIB e, em seguida, calcularmos a diversidade média de idiomas vivos em cada categoria (Figura 15), observa-se um aumento significativo do coeficiente de correlação, de $0,56$ para $0,96$, embora o expoente permaneça invariante dentro das incertezas, $\zeta = 0,31 \pm 0,03$, em conformidade com o expoente z da Equação 3.1. Portanto, existe uma relação de escala robusta entre diversidade linguística e PIB ao longo de mais de cinco décadas. Aqui também analisamos a dependência entre o valor do expoente ζ e o número de categorias utilizadas na separação dos países. Na inserção da Figura 15, apresentando os resultados da variação do número de categorias entre 8 e 15, vemos que o expoente ζ é robusto e independente da categorização. Parece evidente que as maiores economias estão estatisticamente relacionadas a uma maior diversidade linguística.

A causa dessa lei de escala pode ser simplesmente, diriam muitos, uma consequência de que o PIB tende a aumentar com a área. Mas, um acoplamento mais direto e sinérgico entre diversidade linguística e atividade econômica e, portanto, no PIB pode estar presente. Convém apontar que a presença de diversidade linguística antecipa o multilinguismo e

[4] "A relação entre diversidade linguística e o nível de desenvolvimento econômico foi pouco estudada."

Figura 14 – Diversidade linguística em função do Produto Interno Bruto para os dados não categorizados ($n = 194$). A linha tracejada é o melhor ajuste e cuja inclinação fornece o expoente $\zeta = 0,28 \pm 0,03$ ($r = 0,56$).

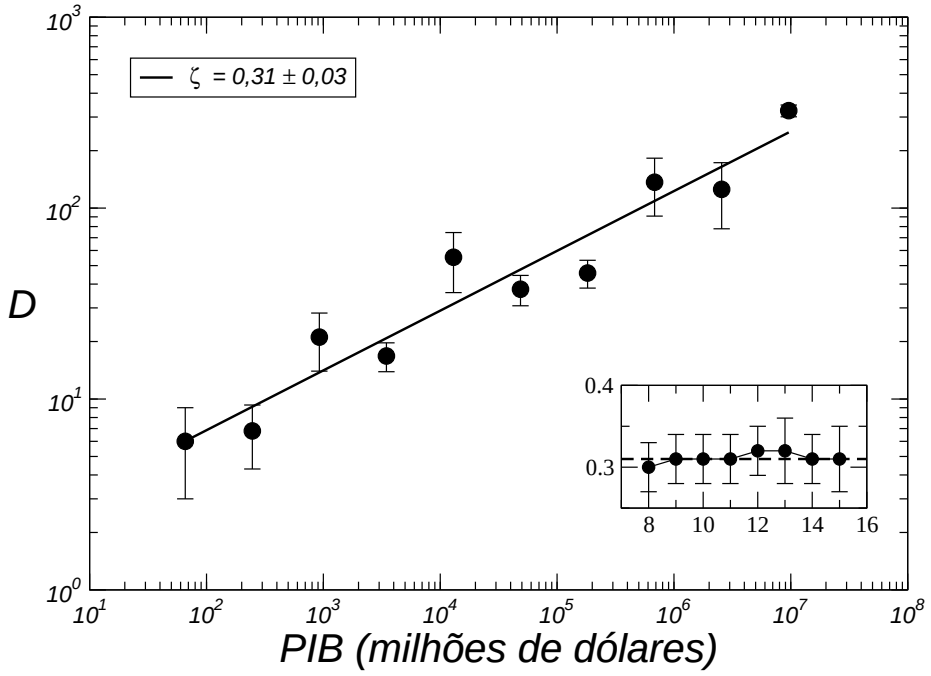


Fonte: Autor (2021).

que este tem sido associado à uma variedade de benefícios econômicos para indivíduos e comunidades (GRIN; SFREDDO; VAILLANCOURT, 2013; HOGAN-BRUN, 2017).

Muitos tendem a pensar que maior diversidade linguística leva a mais "desacerto" ou "confusão" no seio da força de trabalho. Isso se reflete, por exemplo, no uso de expressões no mercado de empregos, que associam desempenhos positivos, às "equipes que falam a mesma língua". Mas essa percepção não pode decorrer de que os modelos de gestão exigem a priori equipes homogêneas? O "senso comum" da homogeneidade linguística tem base científica, ou se trata de preconceito? Poderíamos indagar, por exemplo: em hipotéticos cenários pandêmicos graves e sucessivos, quais dentre as populações dos países com PIBs semelhantes, mas com diversidades linguísticas (e étnicas) diferentes teriam maior possibilidade de sobrevivência no longo prazo? Em prazos mais curtos e mirando experiências passadas, a história das colonizações tem mostrado que tende a ser constante e quase inevitável que grupos poderosos ou majoritários se apropriem do conhecimento de minorias (que contribuem para a diversidade linguística), aumentando seu desenvolvimento e poder econômico, ao mesmo tempo em que segregam essas mesmas minorias, mascarando assim a dependência intrínseca do PIB com a diversidade linguística, ao mesmo tempo em

Figura 15 – Diversidade linguística em função do Produto Interno Bruto para os dados categorizados. A linha tracejada é o melhor ajuste e cuja inclinação fornece o expoente $\zeta = 0,31 \pm 0,03$ ($r = 0,96$). Inserção: valor do expoente ζ em função do número de categorias.



Fonte: Autor (2021).

que fortalecem o mito das equipes que falam a mesma língua.

Uma possível relação entre diversidade linguística e o PIB *per capita* foi adicionalmente investigada. Os dados sobre esse parâmetro econômico também foram obtidos dos Indicadores de Desenvolvimento Mundial do Banco Mundial. No entanto, uma análise global como a apresentada na Figura 14 não aponta para a existência de uma relação de escala entre a diversidade linguística e o PIB *per capita*. Esse achado é semelhante àquele reportado anteriormente (SUTHERLAND, 2003). Com o objetivo de investigar uma possível relação entre a diversidade linguística e o grau de concentração de renda dos países, procedemos uma análise semelhante à descrita ao longo desse capítulo adotando como parâmetro econômico o Índice de Gini. Entretanto não foi encontrada uma correlação entre esse parâmetro e a diversidade de idiomas.

4 DERIVANDO A LEI DE ESCALA DIVERSIDADE LINGUÍSTICA-ÁREA A PARTIR DE UM PRINCÍPIO DE MAXIMIZAÇÃO

"Мы вместе пишем книгу времени."[1]

Svetlana Alexijevich

Neste capítulo apresentamos um modelo heurístico simples, que emerge de um tipo de estrutura variacional, com o objetivo de justificar a relação entre diversidade linguística e área a partir de um princípio de maximização. Mostramos que este modelo reproduz a lei de escala robusta para a diversidade linguística *versus* a área observada hoje na Terra para os 147 maiores países caracterizados por possuírem área superior a 18000 km^2. Uma parcela da discussão apresentada neste capítulo foi publicada no artigo *"A heuristic model for the scaling linguistic diversity-area"* (SANTOS; GOMES, 2020) disponível no Apêndice A.

O interesse na origem e na diversificação de idiomas pode ser rastreado pelo menos até o antigo Egito dos Faraós, como comentado, no quinto século anterior a nossa era, por Heródoto no §II, Livro II, da obra História (HERÓDOTO, 2019). Heródoto conta que em um esforço para dirimir a dúvida sobre qual o povo mais antigo no mundo, se egípcios ou frígios, Psamético elaborou um experimento que tinha o idioma como critério a ser avaliado. Nos tempos modernos, 2200 anos depois, a partir do Século XVII e, sobretudo, durante o Século XVIII diversos debates sobre essa questão se desenvolveram. Em 1770, a Academia de Ciências de Berlim elaborou um concurso de ensaios sobre esse tema (SALMON, 1995). Alguns anos antes, mais precisamente em 1740, o matemático e filósofo Pierre-Louis Moreau de Maupertuis, pioneiro do princípio variacional homônimo e dos estudos linguísticos, publicou, *Réflexions philosophiques sur l'origine des langues et la signification des mots* (MAUPERTUIS, 1740). Embora interessado tanto em princípios variacionais quanto na diversidade linguística, Maupertuis não pôde desenvolver um procedimento semelhante ao relatado aqui pela falta de dados sobre a distribuição de idiomas na Terra naquela época.

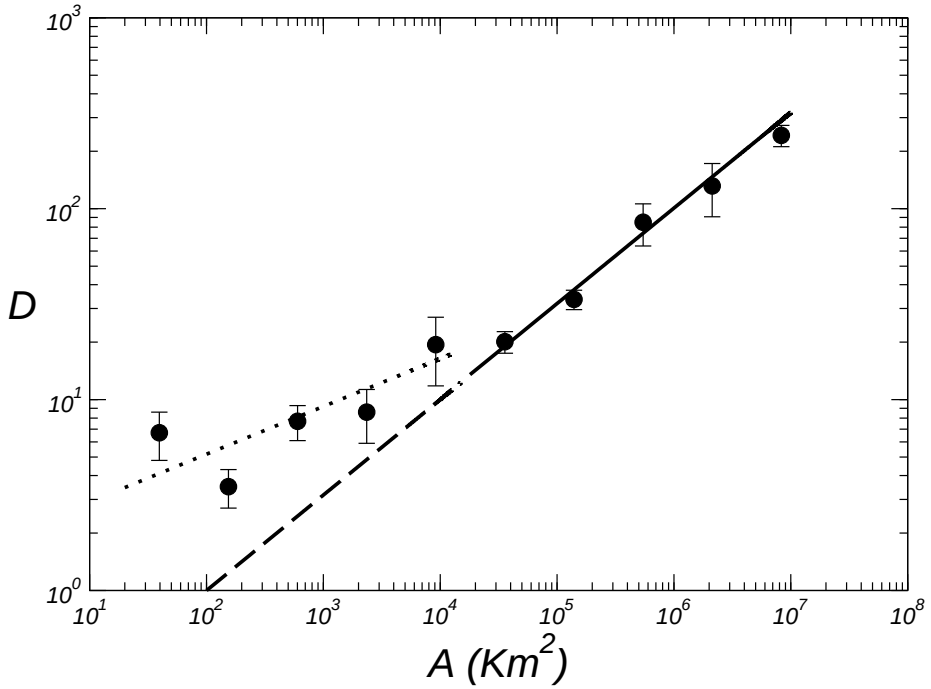
Hoje e nas últimas décadas, a situação é muito diferente, e o tema da origem e da diversificação linguística atraiu atenção de muitos pesquisadores além da comunidade linguística. O "curioso paralelo da evolução das línguas e das espécies" (WHITFIELD, 2008) tem motivado a busca por compreender e modelar a dinâmica linguística tanto de idiomas

[1] "Estamos escrevendo o livro do tempo juntos."

enquanto entidades individuais quanto das famílias e grupos linguísticos (DE OLIVEIRA et al., 2006; DE OLIVEIRA; GOMES; TSANG, 2006; GONG; SHUAI; ZHANG, 2014; GAVIN et al., 2013; GREENHILL et al., 2017; BENTZ et al., 2018).

Conforme visto no capítulo anterior, uma das relações mais básicas que envolvem a diversidade linguística, D, diz respeito à robusta lei de escala que relaciona D com a área A, onde esses idiomas são falados, $D \sim A^z$, ao longo de quase seis décadas de variabilidade na área (GOMES et al., 1999; SANTOS; GOMES, 2019). Dados recentes acerca da distribuição global de idiomas, compilados e apresentados no *Ethnologue* (SIMONS; FENNIG, 2017), levaram ao gráfico exibido na Figura 10 do Capítulo 3. Observando esta figura, pode-se concluir que, independentemente da binarização, $z = 1/3$ dentro de flutuações típicas de 10%, ao longo de mais de cinco ordens de grandeza de área. Entretanto, uma descrição alternativa desta relação de escala pode ser considerada, como mostrado por linhas pontilhadas e contínuas na Figura 16.

Figura 16 – Diversidade linguística D em função da área A, como na Figura 10, adaptado de maneira diferente à linha contínua (pontilhada) que representa a escala $D \sim A^z$ com $z = 1/2$ ($z = 1/4$) para os maiores (menores) 147 (47) países com áreas maiores (menores) que 18000 km^2.



Fonte: Autor (2021).

Portanto conforme mostrado na Figura 16, $z = 1/2$, ao longo de duas décadas e meia de área (linha contínua), em um intervalo muito representativo que considera países pequenos, médios e grandes, com áreas maiores que 18000 km^2, correspondendo a 147

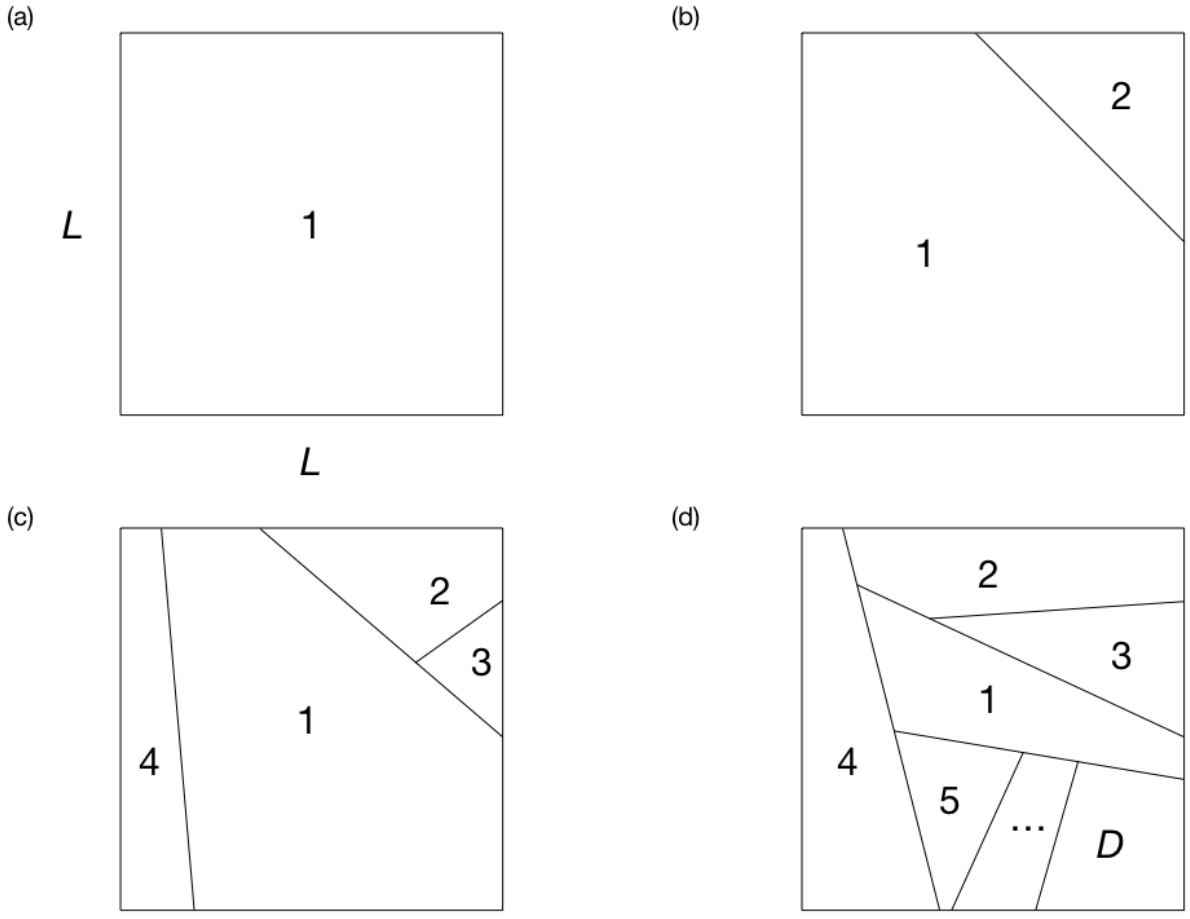
dentre os 194 países com dados disponíveis, restando 47 países pequenos (linha pontilhada), a maioria de natureza insular, pertencentes a um grupo diferente associado com outra relação de escala $D-A$, ao longo de quase três décadas de área, com $z = 1/4$. Aquele grupo de 147 países é bastante representativo na área: a soma de suas áreas corresponde a quase 98% das massas terrestres da Terra. Assim, o foco da discussão a seguir é o exame de uma possível explicação para a lei de da escala $D \sim A^{1/2}$ observada para estados pertencentes, em geral, à grande massa contínua de terras que formam os grandes continentes.

É interessante perguntar se, após um longo período de maturação, a diversidade linguística global observada hoje para uma grande fração de todos os países atuais da Terra evoluiu para uma diversidade linguística máxima possível, compatível com as restrições geográficas, econômicas, políticas e étnico-culturais. Obviamente, qualquer ataque *ab initio* a esse problema é uma tarefa muito complexa, talvez nunca realizável. Nosso objetivo na próxima seção é introduzir um modelo simples de ordem zero que seja razoavelmente realístico e apresente uma resposta a essa pergunta e que possa ser verificado estatisticamente.

4.1 Modelo de maximização da diversidade linguística

Como ponto de partida assumimos que a distribuição de idiomas em um determinado domínio local na superfície da Terra em um passado distante estava distribuída em um conjunto $S = \{a_i\}$ de domínios desconectados cuja área total a_i era associada à um único idioma i do repertório total de idiomas $1, 2, ..., i, ..., D$, onde D é a diversidade linguística total (Figura 17). Cada área a_i representa a soma de todas as áreas de assentamentos ou regiões onde o idioma i é falado. Assim, no início, é suposto que exista uma língua materna única com $a_1 = L^2$ representando a área disponível para todos os seres humanos que falavam tal idioma primordial (Figura 17 (a)). Em decorrência de movimentos migratórios é possível que ocorra separação geográfica desses falantes e que, consequentemente, os idiomas sofram uma especiação alopátrica similar àquela que ocorre com espécies biológicas (SERENO, 1991). Podemos assumir que, num estágio subsequente, a área associada à língua inicial seja separada em duas áreas distintas, cada uma correspondendo a uma língua diferente. (Figura 17 (b)). Esses grupos populacionais portadores desses dois idiomas distintos podem ser divididos e isolados posteriormente para formar novos domínios esparsos (Figura 17 (c)). Tais domínios esparsos podem ser compreendidos como tendo sido separados por regiões despovoadas, como cadeias de montanhas, desertos e cursos de água, que impediram confrontos entre falantes de diferentes idiomas nos tempos antigos. Como tem sido discutido, de fato, as variáveis ambientais e acidentes geológicos desempenham um papel essencial no surgimento da diversidade linguística (AXELSEN; MANRUBIA, 2014; BENTZ et al., 2018). Esses espaços efetivamente desocupados são esquematicamente simbolizados na Figura 17 pela rede de segmentos retilíneos contínuos que separam os vários domínios de

Figura 17 – Ilustração esquemática da dinâmica da evolução das áreas a_i onde os idiomas são falados, para uma área total disponível L^2 aproximadamente constante ou, equivalentemente, uma população aproximadamente constante.



Fonte: Autor (2021).

áreas $a_i s$. Com o tempo, diferentes grupos podem entrar em colaboração ou em guerra, e os idiomas podem entrar em expansão, fusão, extinção ou preservar as respectivas identidades com graus variáveis de trocas linguísticas. Em nosso modelo, supõe-se que a dinâmica desses complexos processos históricos e étnico-culturais ao longo de milênios possa levar ao mosaico da diversidade linguística exibida hoje em nosso planeta.

Os domínios linguísticos introduzidos anteriormente são, portanto, considerados disjuntos e submetidos à regra da soma

$$A = \sum_i a_i, \tag{4.1}$$

onde A é a área total povoada na Terra, que, por sua vez, é uma pequena fração da superfície total disponível, não coberta por águas na Terra. Dentro da estrutura apresentada nesta seção, a condição para alcançar a máxima diversidade no passado é resolvida com o *ansatz*

$$a_i = \alpha.i, \tag{4.2}$$

onde α é uma constante com dimensão de área. Nesses termos, a regra da soma, que representa a soma dos termos da área, é uma progressão aritmética das áreas α, 2α, 3α, 4α, ..., $D\alpha$, tal que

$$A = \alpha\ (1 + 2 + ... + D) = \alpha\frac{D(D+1)}{2}, \tag{4.3}$$

que é equivalente à

$$D \approx \left(\frac{2}{\alpha}\right)^{1/2} A^{1/2}, \tag{4.4}$$

para $D \gg 1$. Ou seja, a lei de escala resultante encontrada neste processo de maximização é exatamente aquela incorporada no ajuste contínuo da Figura 16.

A partir da linha tracejada extrapolada na Figura 16, observamos que $D \approx 1$ para $A \approx 100\ km^2$ portanto, podemos estimar $\alpha = 200\ km^2$ ou $20000\ ha$, uma área associada à região suficiente para garantir a vida de aproximadamente 500 caçadores-coletores (PIMENTEL; PIMENTEL, 2008). Esse mesmo resultado foi obtido anteriormente diretamente dos dados sobre diversidade e população dos países na Seção 3.2. Equivalentemente, isso significa, em termos de nosso modelo, que em algum passado distante o total de áreas ocupadas a_i associadas à um idioma diferente i definiu uma progressão aritmética correspondentemente preenchida por grupos com aproximadamente $500i$ humanos, com $i = 1, 2, 3, ...$. É tentador associar esses grupos de cerca de 500 humanos com um *quanta* de pessoas migrantes. Dessa forma, o *ansatz* anterior é equivalente à quantização das áreas (ou seja, os valores das áreas povoadas são dados por múltiplos inteiros de uma área característica α) ou à quantização do número de grupos de migrantes que falam a mesma língua dentro destas áreas. É interessante notar que

> foi demonstrado que o número de quinhentos indivíduos, característico de uma tribo australiana média, é o valor mínimo para evitar um cruzamento excessivo entre parentes próximos, o que seria prejudicial aos descendentes (CAVALLI-SFORZA; CAVALLI-SFORZA, 2002).

A vantagem deste modelo simples, com poucas e razoáveis suposições, é dupla: (i) fornece a lei de escala observada hoje para a relação entre diversidade linguística e área, válida para todos os países com áreas apreciáveis ($A \geq 18000\ km^2$), ou seja, ao longo de quase três décadas na área e (ii) sugere que o expoente 1/2 pode ser conectado a um princípio de maximização, conforme descrito no início desta seção. Note-se que o expoente 1/2 está presente em vários sistemas físicos, sendo comum em fenômenos importantes sujeitos à invariancia de escala, incluindo fenômenos críticos na física da matéria condensada (STANLEY, 1971).

Supomos que exista uma época primordial em que a diversidade linguística já fosse grande ($D \gg 1$) correspondendo a uma área constante para que nosso argumento seja inicialmente aplicável nesse tempo primordial de referência. Com o passar dos séculos, a

população total, os limites das várias áreas povoadas, a diversidade linguística e a área total mudam com uma dinâmica muito complexa, incluindo a mistura de populações que falam diferentes idiomas em uma mesma região, mas a assinatura $D \sim A^{1/2}$ permaneceu estável como uma relíquia dessa época, pelo menos em uma fração apreciável da área habitada atualmente associada à países com áreas superiores a 18000 km^2.

É justo perguntar se o modelo apresentado acima teve a oportunidade de se materializar na Terra. Somos levados a imaginar que em todos os momentos em que a população estava estacionária (ou quase estacionária) e, correspondentemente, a área ocupada estava estacionária (ou quase estacionária), tivemos a oportunidade de esse modelo ser efetivamente ativo. Uma vez alcançada a situação caracterizada pela lei de escala $D \sim A^{1/2}$, a dinâmica poderia ficar travada nesta relação funcional de diversidade linguística máxima até o presente. Embora nos últimos dois séculos a população humana tenha crescido sem precedentes, a taxa de crescimento populacional do *Homo Sapiens* foi baixa na maior parte de sua história (BONGAARTS, 2009). Em paralelo, é interessante notar que, nos últimos anos, foram discutidas as relações entre o desenvolvimento agrícola pré-histórico e o crescimento do tamanho da população. Expansões demográficas dos grupos humanos foram investigadas (BOCQUET-APPEL, 2011) e a Transição Demográfica Neolítica (NDT, na sigla em inglês) tem sido relacionada à hipótese de dispersão agrícola/linguística (BELLWOOD; RENFREW, 2002). Vale ressaltar ainda que os efeitos das expansões populacionais neolíticas são persistentes, como conjecturado acima para a relação $D - A$, e ainda podem ser percebidos nos padrões atuais de diversidade linguística (RENFREW, 1989).

Uma questão final surge: a evolução da distribuição linguística no futuro pode resultar em uma redução no expoente da lei de escala diversidade-área de 1/2 para 1/4 para a região que compreende países com uma área acima de 18000 km^2? Acreditamos que sim, e essa redução na diversidade, proveniente da extinção de idiomas, já amplamente discutida na literatura (ROMAINE, 2007; AUSTIN; SALLABANK, 2011), também foi sugerida anteriormente a partir de simulações computacionais (DE OLIVEIRA; GOMES; TSANG, 2006).

4.2 Migrações na Terra e além

Como vimos, a partir do gráfico $D(A)$ na Figura 16, é necessária uma área de 100 km^2 para manter um único idioma vivo; tal resultado decorre de dados estáticos obtidos após a implantação da agricultura, evidentemente; já que eles se referem às populações assentadas da configuração atual. Por outro lado, o parâmetro $\alpha = 200$ km^2 associa-se, em nosso modelo, à ideia dinâmica de migrações para ocupação de terras agricultáveis. No estado de migração, os grupos de indivíduos que se deslocavam precisavam usar estratégias

de caça e coleta para assegurar a sobrevivência enquanto estavam em marcha, i.e. eram agentes não-assentados. No modo de caça-coleta praticado nessas migrações, 200 km^2 suporta até 500 humanos caçadores-coletores. Se um grupo desse tamanho conseguisse sobreviver razoavelmente incólume numa longa caminhada e aportasse uma nova língua numa área de cerca de 100 km^2, e nela se assentasse, ele possivelmente sobreviveria com uma população quase-estática, com o auxílio da agricultura.

O problema do tamanho mínimo necessário a uma população de migrantes ou colonizadores para viverem auto-suficientemente num local tem grande interesse teórico, tanto relativo ao passado remoto, como é o caso aqui enfocado, quanto ao presente e ao futuro. Neste último caso, ele se coloca, inclusive, para o problema da colonização de outros sistemas solares e também para a ocupação do nosso nunca esquecido vizinho, o planeta Marte. Numa perspectiva ainda mais avançada, à luz das descobertas, nas últimas décadas, de muitos exoplanetas (BUTLER et al., 2006; SING et al., 2016), pode-se pensar nas condições a serem satisfeitas por colonizadores terráqueos de Proxima Centauri b, exoplaneta que está orbitando dentro da zona habitável da anã vermelha Proxima Centauri, a mais próxima estrela do nosso Sol, distante 4,24 anos-luz da Terra. De fato, Marin & Beluffi calcularam em 2018, usando simulações Monte Carlo, que o tamanho de uma tripulação multi-geração capaz de sobreviver geneticamente saudável a uma viagem de 6300 anos a esse sistema planetário deveria ser constituída de, no mínimo, 98 humanos (MARIN; BELUFFI, 2018). Mais recentemente, J.-M. Salotti (SALOTTI, 2020) usando algumas hipóteses de escala, calcularam que o número mínimo de colonos necessários para uma ocupação auto-suficiente de Marte seria o de 110 pessoas, baseando- se, diferentemente de Marin e Beluffi, nos aspectos de engenharia subjacentes ao problema da instalação e sobrevivência nesse planeta.

De qualquer modo, esses números mínimos de 98 ou 110 colonos, partindo de premissas bastante diferentes, são da mesma ordem daquele número mínimo de migrantes encontrado no modelo de ocupação de áreas, por falantes de uma única língua, discutido neste capítulo, dentro do quadro da distribuição da diversidade linguística atual na Terra. Os problemas associados ao(s) idioma(s) que será(serão) falado(s) pelas tripulações nessas missões espaciais são um dos aspectos que terá que ser levado em consideração nos correspondentes planejamentos. Não sabemos qual será o encaminhamento dado quanto à essa(s) escolha(s) mas, evidentemente, é interessante constatar que tipos de estimativas partindo de argumentos muito diferentes levam a números muito próximos quando grupos de pioneiros humanos são envolvidos na ocupação de novos espaços.

Concluímos este capítulo pontuando que de acordo com muitos outros casos em que um procedimento de maximização é responsável pelo surgimento de leis de escala (PASTOR-SATORRAS; WAGENSBERG, 1998; CHEN, 2012), nós introduzimos aqui um *ansatz* aritmético-geométrico simples que leva à relação de escala $D \sim A^{1/2}$

observada para a diversidade linguística, D, encontrada em áreas $A \geq 180000\ km^2$ na Terra. Conjecturamos que o universo dos idiomas passou por uma evolução no passado remoto, em que a diversidade linguística aumentou de maneira complexa, deixando-nos a dependência funcional $D(A)$ atualmente observada como uma assinatura dessa distribuição primordial dos idiomas. Em adição às análises feitas neste capítulo e no anterior, uma terceira abordagem da relação de escala $D - A$, desta vez termodinâmica, usando forças entrópicas e de auto-exclusão, será apresentada no capítulo seguinte.

5 LEI DE ESCALA DIVERSIDADE LINGUÍSTICA-ÁREA DERIVADA DE UMA ANALOGIA TERMODINÂMICA COM FORÇAS ENTRÓPICAS E DE AUTO-EXCLUSÃO

"Die Grenzen meiner Sprache.
bedeuten die Grenzen meiner Welt."[1]

Ludwig Wittgenstein

Neste capítulo apresentamos um modelo termodinâmico de campo médio com forças entrópicas e de auto-exclusão a partir do qual a relação entre diversidade linguística e área atualmente observada na Terra pode ser entendida. Mostramos que esta abordagem confirma os resultados discutidos nos capítulos anteriores, apontando que $z = 1/2$ ocorre quando a energia livre do presente modelo exibe um tipo particular de simetria de troca. Explicamos a dependência $D - A$ para a população da Terra, ou seja, incluindo o excedente dos muitos idiomas falados por pequenas populações (correspondendo a mais de cinco décadas de variabilidade da área) para uma energia livre ligeiramente diferente exibindo uma quebra dessa simetria. Além disso, o modelo concorda com os dados empíricos que mostram a diminuição da diversidade linguística com o aumento da latitude. Ademais, o presente modelo fornece uma base para entender cenários futuros de perda de diversidade linguística.

5.1 Construindo uma energia livre para a distribuição linguística

Uma afirmação recorrente nos estudos linguísticos aponta que um idioma não existe no vácuo. De fato, uma complexa rede de relações existe entre o meio ambiente, a história, os idiomas e seus falantes. Em meio a essas relações, idiomas nascem, alguns crescem, diversos colidem e, de modo cada vez mais recorrente, muitos idiomas morrem. Nas últimas décadas, um número crescente de modelos de competição linguística, ou dinâmica de idiomas, foi desenvolvido (BAGGS; FREEDMAN, 1990; ABRAMS; STROGATZ, 2003;

[1] "Os limites de minha linguagem significam os limites de meu mundo."

DE OLIVEIRA; GOMES; TSANG, 2006; PATRIARCA et al., 2012; ZHOU; SZYMANSKI; GAO, 2020). Conforme comumente descrito por esses modelos e observado em idiomas reais, tal dinâmica pode ter como resultado o domínio de um idioma e a extinção dos idiomas dominados, a coexistência ou ainda a unificação de idiomas.

Similarmente a outros fenômenos biológicos e culturais, podemos compreender a linguagem como estando sujeita a duas interações opostas: por um lado, uma tendência à uniformidade e, por outro, uma tendência à diversidade. Tais interações estão associadas tanto a cada idioma como entidade particular quanto ao conjunto de idiomas existentes em uma dada região. Ferdinand de Saussure considerava que os idiomas se propagavam sujeitos à duas forças que agiriam simultaneamente e em sentidos contrários. Uma força, o "espírito de campanário", teria caráter particularista enquanto a outra, a força de intercurso, criaria as comunicações entre seres humanos (SAUSSURE, 2012). Em 1949, G. K. Zipf levantou a hipótese de que o processo de comunicação seria regido pela disputa entre forças de unificação e de diversificação do vocabulário (ZIPF, 1949). Mais recentemente essa dinâmica de fragmentação linguística *versus* dominância de um idioma foi discutida no âmbito da modelagem física de sistemas linguísticos (STAUFFER; SCHULZE, 2005; DE OLIVEIRA; GOMES; TSANG, 2006). Essa capacidade de um idioma tanto poder absorver características de outro idioma quanto transmitir suas próprias características quando do contato entre indivíduos provenientes de diferentes comunidades linguísticas foi recentemente resumida na analogia que afirma que "languages are as much sponges as viruses"[2] (PATRIARCA; HEINSALU; LEONARD, 2020).

Aqui nós assumimos que a distribuição de idiomas em uma Terra essencialmente bidimensional é governada, dentro de um domínio circular de raio R, por duas forças antagônicas: um tipo de interação de auto-exclusão V_{AE} que introduz uma tendência para diferentes idiomas permanecerem separados, preservando identidades intrínsecas, e um tipo de termo entrópico V_S que leva em consideração uma atração e mistura inevitáveis entre diferentes culturas com trocas de experiências ao longo do espaço e de grandes escalas de tempo, incluindo o interesse humano recíproco entre diferentes comunidades linguísticas.

Assumimos que essas duas forças sejam combinadas em uma pseudo energia livre F segundo

$$F = V_{AE} + V_S, \tag{5.1}$$

a partir da qual as propriedades da distribuição diversidade linguística-área poderiam ser derivadas. Uma vez que, em geral, a energia livre é escrita como

$$F = U - TS, \tag{5.2}$$

o primeiro termo na Equação 5.1 corresponde à energia interna U e o segundo corresponde ao termo entrópico, $-TS$, onde T e S são, respectivamente, a temperatura e a entropia do

[2] "Idiomas são tanto esponjas quanto vírus".

sistema. Como o sistema linguístico não é um sistema termodinâmico em sentido restrito, aqui T é um parâmetro do sistema, não reivindicado ser exatamente a temperatura Kelvin, embora possa, em princípio, estar relacionado com ela.

Para descrever a primeira contribuição para esta pseudo energia livre nós consideramos a interação auto-excludente como

$$V_{AE} = C\rho^2 v. \tag{5.3}$$

Nesta equação, C é uma constante positiva, ρ é a densidade média de diversidade linguística,

$$\rho = \frac{D}{v} \tag{5.4}$$

e v é o volume bidimensional dado pela área A do domínio linguístico circular de raio R,

$$v \sim A \sim R^2. \tag{5.5}$$

Dessa forma

$$\rho \sim \frac{D}{R^2}. \tag{5.6}$$

O expoente 2 na Equação 5.3 considera todos os idiomas como possuindo interações de dois corpos.

À segunda contribuição, o termo entrópico, nós associamos um termo parabólico, Hookeano,

$$V_S = BR^2, \tag{5.7}$$

que privilegia o agrupamento, ou mistura dos idiomas, em contraste com o termo V_{AE} que privilegia a repulsão, tendendo a separar idiomas diferentes. A origem termodinâmica dessa contribuição entrópica está associada à ideia de que próximo ao equilíbrio S é máxima e, portanto, V_S pode ser expandido em série de potências retendo-se apenas o termo quadrático:

$$V_S = \left[-T \left(s + \frac{1}{2} \left(\frac{\partial^2 S}{\partial R^2} \right)_{eq.} R^2 + ... \right) \right], \tag{5.8}$$

onde $s = S(0)$ é uma constante. Assim a Equação 5.1, a menos de uma constante, toma a forma

$$F = \frac{CD^2}{R^2} + BR^2, \tag{5.9}$$

notando-se que

$$B = -\frac{T}{2} \left(\frac{\partial^2 S}{\partial R^2} \right)_{eq.} > 0, \tag{5.10}$$

pois a concavidade $(\partial^2 S/\partial R^2)$ no equilíbrio é negativa. O leitor familiarizado com a física estatística de polímeros reconhecerá na Equação 5.9 ingredientes do modelo de Flory-de Gennes para polímeros auto-excludentes (DE GENNES, 1979).

5.2 Analisando a energia livre proposta

Alguns cuidados devem ser considerados na expressão para V_S para descrever corretamente os aspectos de entropia associados aos idiomas. A constante elástica efetiva, B, Equação 5.10, será definida pelo *ansatz*

$$B = \frac{C'}{D^2},\ C' = bT, \tag{5.11}$$

onde $b > 0$ é um fator de correção formal que poderia ser igualado à unidade redefinindo T apropriadamente. Portanto

$$\frac{b}{D^2} \leftrightarrow -\frac{1}{2}\left(\frac{\partial^2 S}{\partial R^2}\right)_{eq.} \tag{5.12}$$

e notamos que a tendência de agrupar idiomas em uma região geográfica limitada de raio R é enfraquecida quando a diversidade linguística D dentro dela já é grande. Ou seja, a facilidade de confinar vários idiomas em uma região diminui porque a concavidade do potencial de confinamento V_S diminui. Além desse comportamento realista manifestado pela escolha dada na Equação 5.11, surge uma simetria na expressão da pseudo energia livre:

$$F = \frac{CD^2}{R^2} + \frac{C'R^2}{D^2}, \tag{5.13}$$

ou seja, a invariância de F sob as trocas $D \leftrightarrow R$, $C \leftrightarrow C'$. Na Figura 18 apresentamos esquematicamente os termos de auto-exclusão e entrópico, o primeiro tendendo a separar os idiomas e o segundo tendendo a agrupá-los. A superposição desses dois termos levará à relação de equilíbrio procurada representada através da relação $D(A)$.

Após usarmos a Equação 5.11 na Equação 5.13 e minimizarmos em relação a R, temos

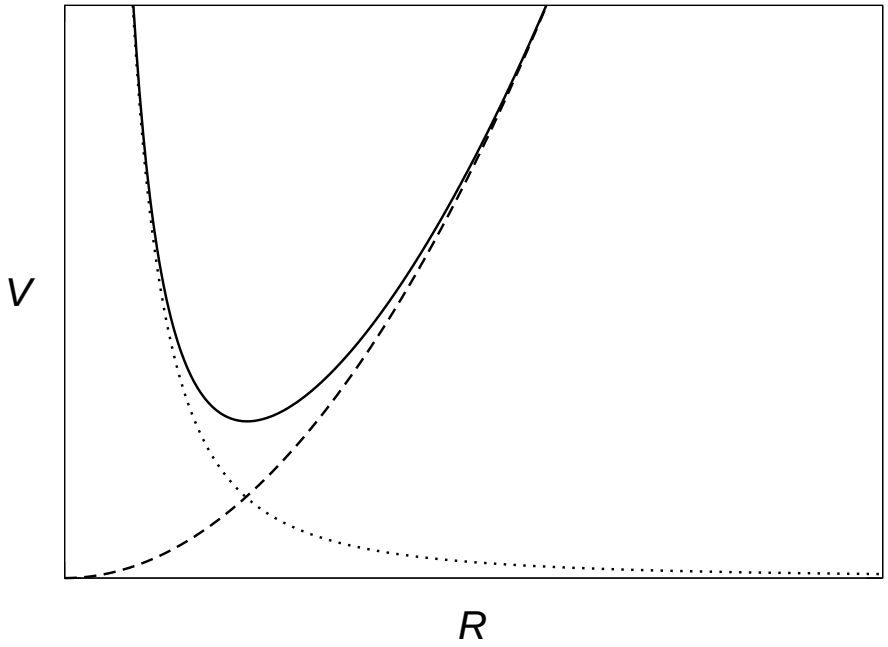
$$D = \left(\frac{C'}{C}\right)^{1/4} R \sim T^{1/4} A^{1/2}. \tag{5.14}$$

Nessas condições, a diversidade linguística escala com a raiz quadrada da área, ou seja, $z = 1/2$ dado que $D \sim A^z$ conforme discutido na Seção 3.1. Este resultado, familiar ao leitor do capítulo anterior, está de acordo com o expoente associado à linha contínua na Figura 16. Ao tomarmos os dados de temperatura média anual por país, disponíveis no *World Bank's Climate Change Knowledge Portal* (WORLD BANK, 2020), torna-se possível observar na Figura 19 a lei de escala

$$D \sim (T^{1/2} A)^{\zeta} \tag{5.15}$$

com o expoente $\zeta = 1/2$. Os dados na Figura 19 incluem todos os 147 países cujas áreas são maiores que 18000 km^2, representando em conjunto aproximadamente 98% da população da Terra. Além disso, a Equação 5.14 indica que $D/A^{1/2}$ decresce quando a temperatura atmosférica diminui, ou seja, com o aumento da latitude, resultado que está qualitativamente de acordo com os dados empíricos (GAVIN; STEPP, 2014).

Figura 18 – Representação esquemática das funções que descrevem as interações de auto-exclusão V_{AE} (linha pontilhada), entrópica V_S (linha tracejada), bem como a pseudo energia livre F (linha contínua) em função do raio R do domínio linguístico considerado.



Fonte: Autor (2021).

Dentro da estrutura apresentada acima, outras opções analíticas para V_{AE} e B estão obviamente associadas a diferentes relações de escala entre a diversidade linguística D e a área A. Assumindo

$$V_{AE} = C\rho^n v, \tag{5.16}$$

com $C > 0$, $n \geq 2$ e

$$B = \frac{bT}{D^m}, \tag{5.17}$$

$m > 0$, a minimização da Equação 5.13 em relação a R leva a

$$D \sim (T^{1/n} A)^{\zeta'} \tag{5.18}$$

com

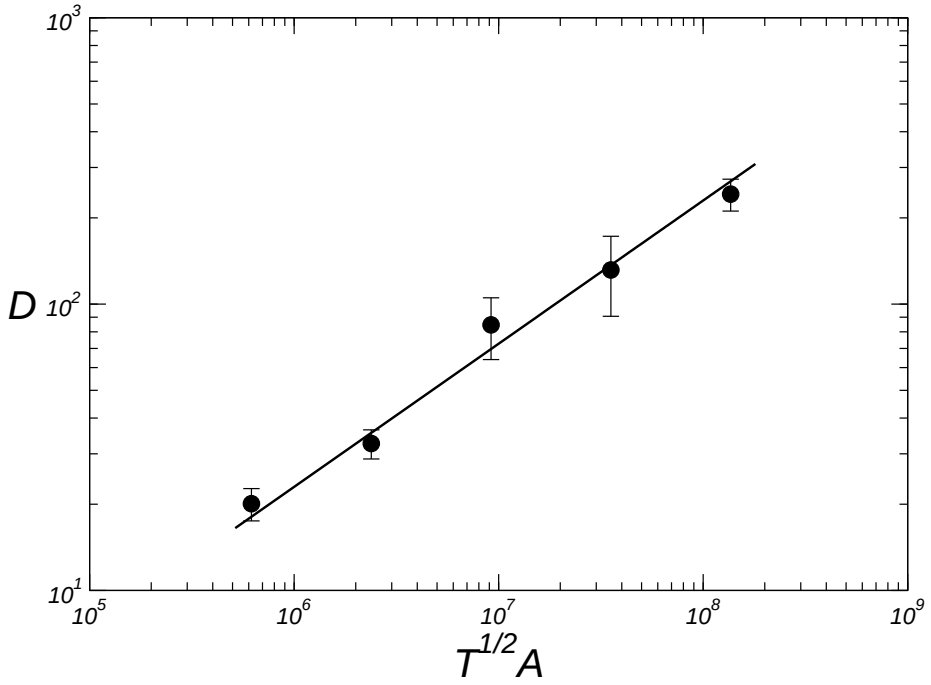
$$\zeta' = \frac{1}{1 + \frac{m}{n}}. \tag{5.19}$$

Assim o expoente ζ' é dependente de um único parâmetro, a razão m/n e se aproxima de 1 para m/n pequeno. Por outro lado, o expoente $\zeta' = 1/2$, é igualmente invariante, segundo o vínculo

$$\frac{1}{2} = \frac{1}{1 + \frac{m}{n}}, \tag{5.20}$$

isto é, desde que $n = m$. Apesar da infinidade de possibilidades para $n = m$, o caso particular $n = 2$ tem uma motivação especial na física porque se refere às muito importantes

Figura 19 – Diversidade linguística D por país em função da raiz quadrada da temperatura média anual T (em Kelvins) multiplicada pela área A (dados categorizados para 147 países). A linha contínua é $D \sim (T^{1/2}A)^{\zeta}$ com $\zeta = 1/2$. Dados extraídos da vigésima edição do *Ethnologue*, dos Indicadores de Desenvolvimento Mundial e do *World Bank's Climate Change Knowledge Portal*.



Fonte: Autor (2021).

interações de dois corpos que incluem fenômenos diversos em áreas como eletromagnetismo, gravitação e mecânica estatística.

Fixando nossa atenção no caso $n = 2$ temos que as Equações 5.18 e 5.19 se tornam

$$D \sim (T^{1/2}A)^{\zeta'} \tag{5.21}$$

e

$$\zeta' = \frac{1}{1 + \frac{m}{2}}. \tag{5.22}$$

Assim, para obtermos o expoente $\zeta' = 1/3$, discutido na Seção 3.1, o expoente m no termo entrópico, Equações 5.7 e 5.17, deve satisfazer $m = 4$, levando à ideia de que a constante B no potencial parabólico atrativo (termo entrópico) diminui rapidamente à medida que a diversidade aumenta dentro de um domínio de raio R. Assim, o expoente $\zeta' = 1/3$ está relacionado a um cenário de enfraquecimento da contribuição entrópica. Isso é consistente com o comportamento $D \sim (T^{(1/2)}A)^{\zeta'}$, $\zeta' = 1/3$, observado na Figura 20. Os dados utilizados para a construção desta figura incluem supracitado conjunto de 147

países com área superior a 18000 km^2 mais outros 39 países com áreas menores. Esses 39 países que possuem uma variabilidade de área de cerca de $2,5$ décadas são em sua maioria de caráter insular, são estados esparsamente distribuídos na Terra e que interagem fracamente com os países vizinhos num cenário em que o termo entrópico é drasticamente reduzido. Note-se também, a partir da Equação 5.19, que ζ' diminui, por exemplo, em um cenário de perda de diversidade linguística, em um futuro não muito hipotético, quando m cresce. O caso extremo $\zeta' \to 0$, que inclui a situação em que a diversidade linguística cresce logaritmicamente com a área, é definido, no quadro teórico deste modelo de campo médio, para $m \gg 1$.

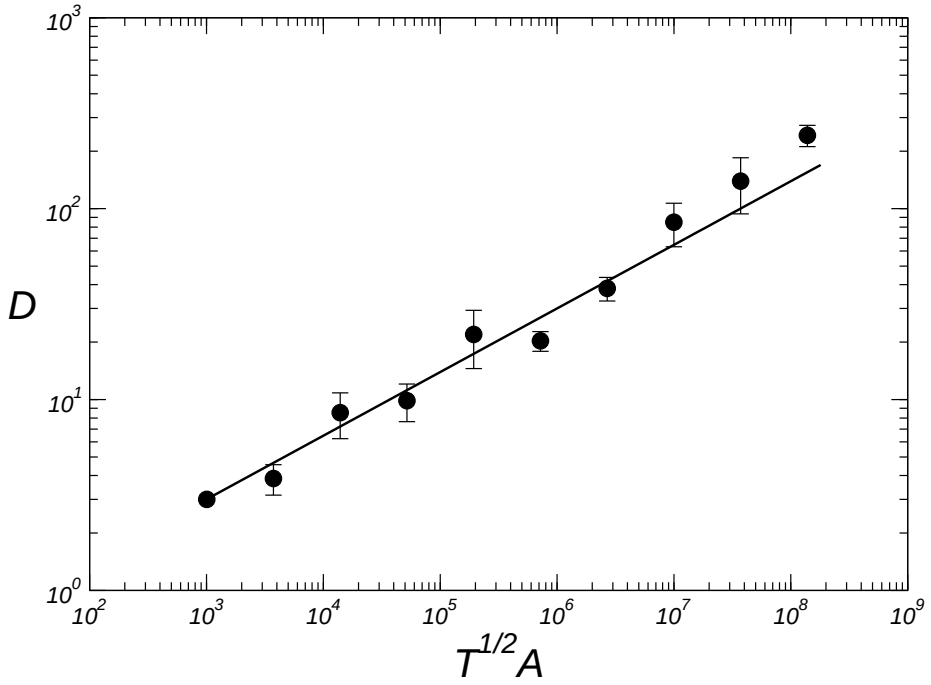
Ainda dentro do modelo apresentado neste capítulo, a redução do termo entrópico, ou seja, a redução da constante de acoplamento entrópico B (Equação 5.17), devido ao aumento de m, significa que as populações são mais refratárias às misturas ou trocas de experiências linguísticas. Em outras palavras, se os idiomas quiserem permanecer puros minimizando as interações com seus vizinhos, o resultado será um mundo com baixa diversidade (idiomas condenados à morte), um resultado bem conhecido dos linguistas (EVANS, 2010). Deve-se notar que esta não é, intrinsecamente, uma limitação dos idiomas, mas dos grupos humanos que falam esses idiomas. Ao contrário, para comunidades linguísticas mais receptivas e abertas às interações com seus vizinhos, ou seja, para m/n pequeno, o expoente z aumenta, e a diversidade linguística por área também aumenta.

5.3 Efeito da temperatura sobre a diversidade linguística e extensão para espécies biológicas

A presença explícita da temperatura na Equação 5.14 convida-nos a explorar algumas consequências dessa abordagem termodinâmica. Do ponto de vista geográfico, podemos observar na Figura 21 que a temperatura média do ar na superfície da Terra é aproximadamente uma função convexa simétrica em relação ao equador da Terra (FEULNER et al., 2013).

É também conhecido que a diversidade linguística é fortemente correlacionada com a latitude (MACE; PAGEL, 1995; NETTLE, 1998; SUTHERLAND, 2003; GAVIN; STEPP, 2014). Há um gradiente latitudinal global na diversidade linguística, Figura 22, com mais idiomas perto do equador do que em latitudes mais altas (HUA et al., 2019). Tal fato também pode ser observado a partir do histograma apresentado na Figura 23 produzido a partir de dados extraídos do *Glottolog* (HAMMARSTRÖM et al., 2020). Padrões latitudinais semelhantes foram reportados para diversidade cultural. Esse gradiente é similar ao padrão geográfico (Figura 24) de aumento da biodiversidade dos pólos ao Equador que, sendo conhecido a alguns séculos, foi quantificado ao longo das últimas décadas (WILLIG; KAUFMAN; STEVENS, 2003; HILLEBRAND, 2004; BROWN, 2014).

Figura 20 – Diversidade linguística D por país em função da raiz quadrada da temperatura média anual T (em Kelvins) multiplicada pela área A (dados categorizados para 186 países). A linha contínua é $D \sim (T^{1/2}A)^{\zeta}$ com $\zeta = 1/3$. Dados extraídos da vigésima edição do *Ethnologue*, dos Indicadores de Desenvolvimento Mundial e do *World Bank's Climate Change Knowledge Portal*.



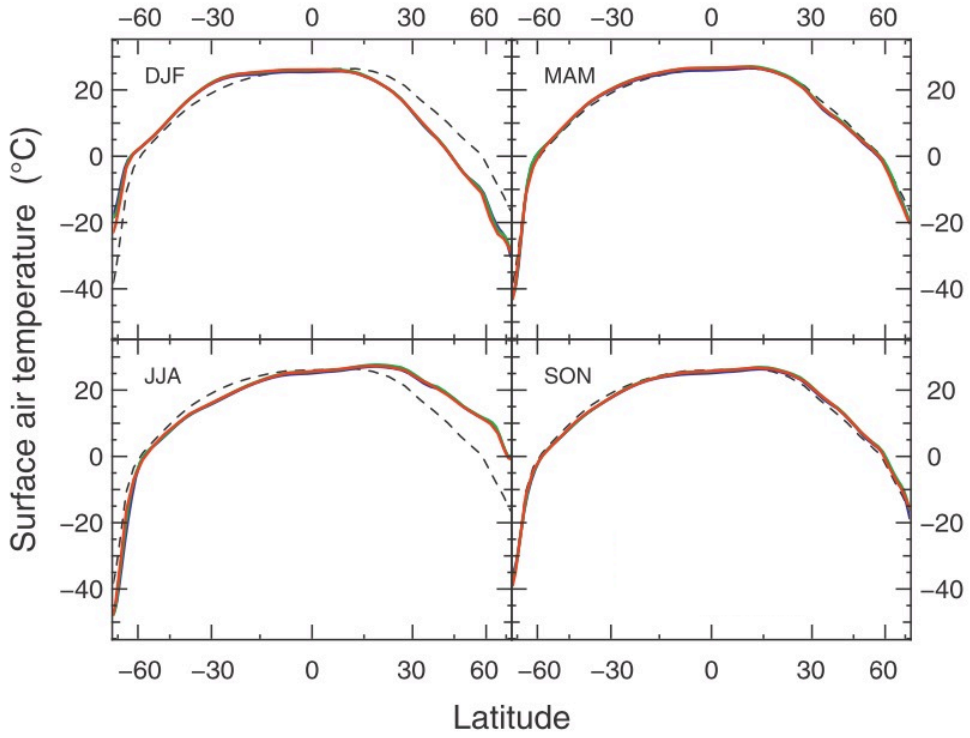
Fonte: Autor (2021).

Portanto, conforme citado nos capítulos precedentes, este é mais um dos diversos padrões associados à diversidade linguística que se assemelham aos padrões de biodiversidade.

O modelo aqui introduzido prevê através da Equação 5.14 ou Equação 5.18, no caso mais geral, que a diversidade linguística aumenta fracamente com a temperatura para áreas fixas, ou seja, D aumenta à medida que diminui a latitude em direção ao equador, de acordo com o observações (HUA et al., 2019). Percebe-se então uma íntima ligação entre a temperatura do ambiente e as diversidades linguística e biológica. O que pode ser compreendido pelo fato de que as taxas metabólicas aumentam exponencialmente com a temperatura (GILLOOLY et al., 2001). Assim, em climas mais quentes, taxas metabólicas mais altas aumentariam as taxas de interações ecológicas e processos evolutivos, e estes, por sua vez, gerariam maior diversidade (BURNSIDE et al., 2012).

Se os padrões de diversidade cultural e biodiversidade são decorrência de processo semelhantes, somos levados à conclusão de que nosso modelo de campo médio com componentes entrópicos e de auto-exclusão pode ser útil para essas situações adicionais. Também é interessante observar que os fatores que afetam a abundância de caçadores-

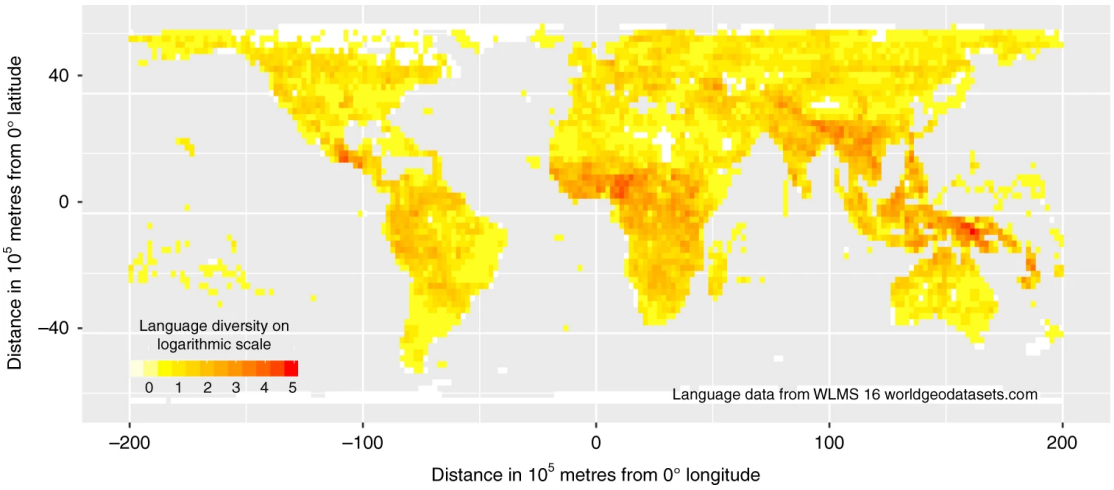
Figura 21 – A curva vermelha apresenta a média anual das temperaturas do ar na superfície terrestre para o período de 1961-1990 em função da latitude a partir de dados do *Climatic Research Unit* (CRU). Cada quadrante representa um trimestre.



Fonte: *On the origin of the surface air temperature difference between the hemispheres in Earth's present-day climate* (FEULNER et al., 2013).

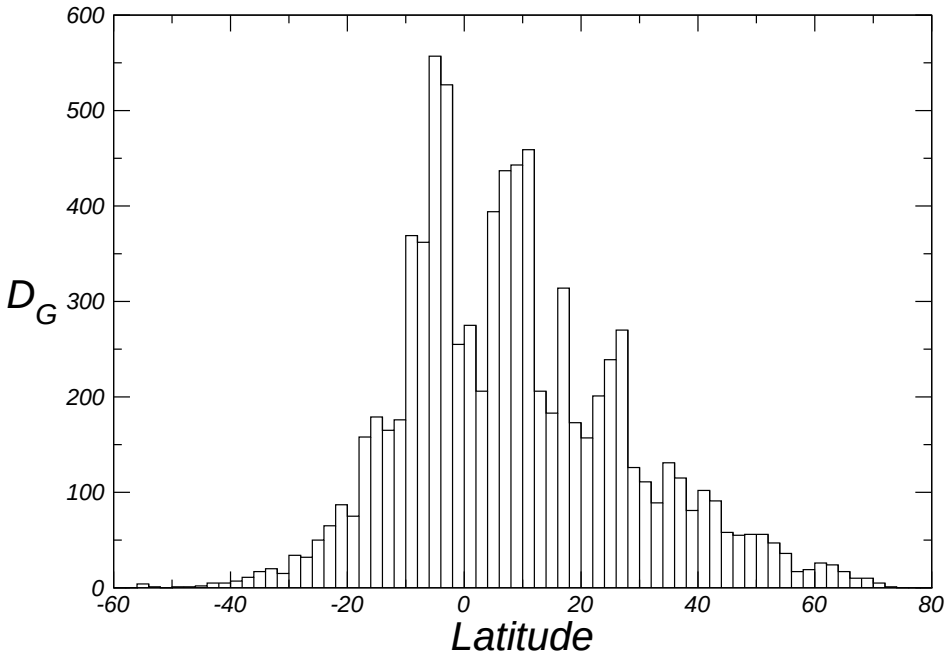
coletores em uma escala global são semelhantes aos que controlam a biodiversidade em escala global (TALLAVAARA; ERONEN; LUOTO, 2018). Apontamos que, embora nossa análise seja baseada em um instantâneo contemporâneo da diversidade linguística e das informações climáticas, os gradientes de temperatura e o gradiente correspondente de biodiversidade existem há dezenas de milhões de anos (MANNION et al., 2014).

Figura 22 – Distribuição global da diversidade linguística.



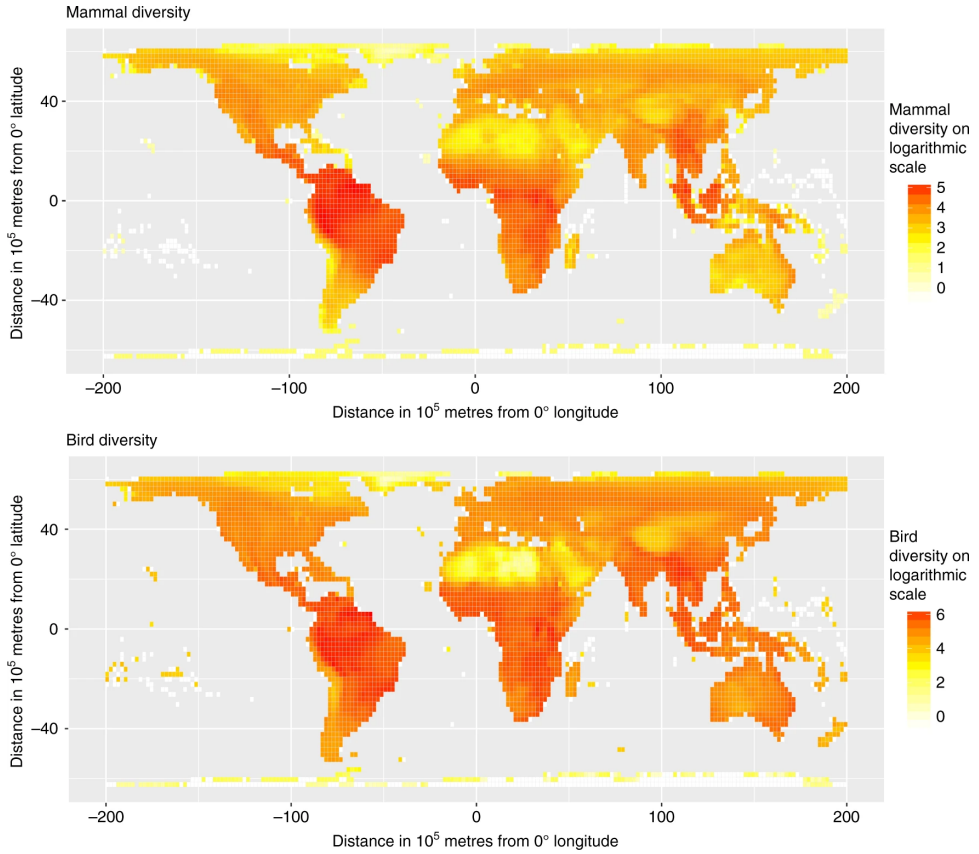
Fonte: *The ecological drivers of variation in global language diversity* (HUA et al., 2019).

Figura 23 – Dependência latitudinal do número D_G de idiomas obtida a partir de dados do *Glottolog* (HAMMARSTRÖM et al., 2020).



Fonte: Autor (2021).

Figura 24 – Distribuição global da diversidade de mamíferos e diversidade de pássaros.



Fonte: *The ecological drivers of variation in global language diversity* (HUA et al., 2019).

6 LEIS DE ESCALA EMERGENTES EM PROCESSOS DE CLASSIFICAÇÃO

"Sólo podemos hablar porque
nuestro idioma no está solo."[1]

Fabio Morábito

Neste capítulo partindo de processos de classificação e ordenamento apresentamos leis de escala hiperbólicas associadas a parâmetros linguísticos, econômicos, geográficos e demográficos. Em contraposição à maior parte das leis de escala reportadas nos capítulos anteriores, são negativos os expoentes que controlam as distribuições discutidas neste capítulo. Inicialmente apresentamos uma medida diferente da diversidade linguística, o número de idiomas entre diferentes países, e comparamos sua distribuição acumulada com aquelas observadas em estudos de fragmentação, percolação e crescimento. Em seguida apresentamos as distribuições acumuladas para o Produto Interno Bruto, área e população dos países. Logo após, são discutidos os resultados das distribuições acumuladas das populações por idioma bem como das populações por grupo étnico. Na quinta seção investigamos as leis de escala emergentes da classificação das famílias linguísticas tanto segundo o número de idiomas quanto de acordo com o número de falantes. A sexta seção contém uma análise da distribuição de tamanhos de idiomas das maiores famílias linguísticas contemporâneas. O capítulo se encerra com um resumo das nossas principais conclusões. Um subgrupo dos resultados apresentados nas duas primeiras seções foi publicado no artigo *"Revisiting scaling relations for linguistic diversity"* (SANTOS; GOMES, 2019) disponível no Apêndice A.

6.1 Distribuição acumulada da diversidade linguística por país

O desenvolvimento da linguagem é apontado como um dos fatores fundamentais para a expansão dos grupos humanos que migraram a partir da origem do ser humano moderno no continente africano (BELLWOOD, 2013). Os diferentes processos históricos pelos quais passaram esses grupos migratórios levaram à uma distribuição não uniforme dos idiomas sobre a superfície terrestre (GAVIN et al., 2013). O agrupamento, nem sempre pacífico, de populações em países reflete essa heterogeneidade também no número de idiomas. A partir dos dados disponíveis na vigésima edição do *Ethnologue* (SIMONS;

[1] "Somente podemos falar porque nossa língua não está só."

FENNIG, 2017), sabemos que a diversidade linguística por país cobre três décadas de variabilidade de modo que, enquanto Papua-Nova Guiné tem falantes de 840 idiomas dentro de suas fronteiras, somente dois idiomas são listados nas Maldivas.

O número de países com pelo menos D diferentes idiomas cada, ou seja, a distribuição acumulada N_D, é definida em termos da distribuição diferencial de países com diversidade linguística D', $n(D')$, como

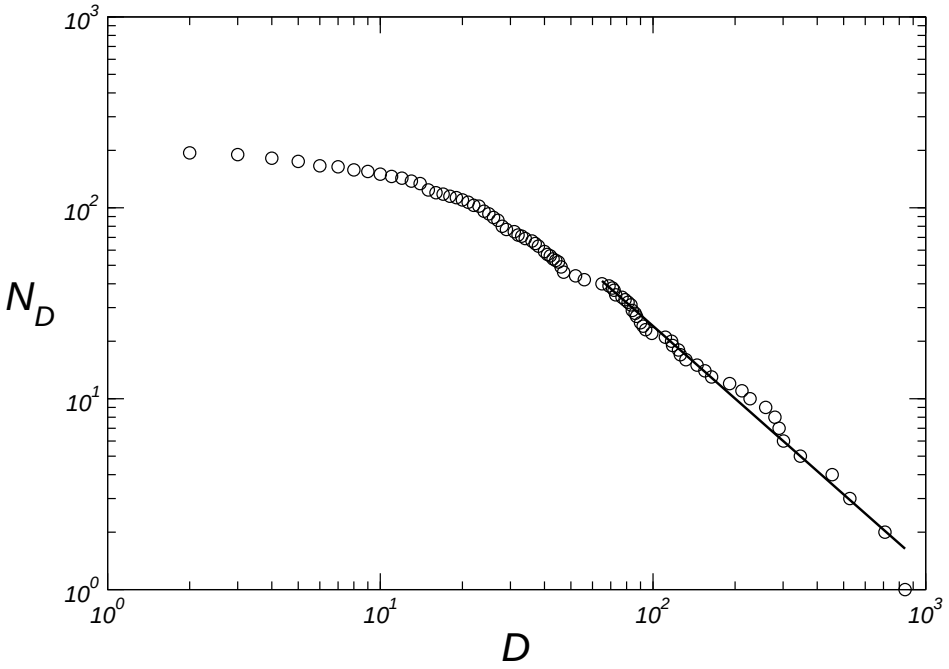
$$N_D = \int_D^\infty n(D')dD'. \tag{6.1}$$

Assintoticamente, se $n(D')$ tiver a simetria de escala, obteremos distribuições hiperbólicas da forma

$$N_D \sim D^{-\delta} \tag{6.2}$$

com $\delta = 1,26 \pm 0,03$ para $D > 60$, i. e. ao longo de pouco mais de uma década de variabilidade em D, conforme observado na Figura 25. Este último expoente é um pouco maior do que o relatado anteriormente (GOMES et al., 1999), $\delta = 1,1$ para $D > 60$. Enquanto este trabalho precedente reportou um comportamento de lei de escala de dois regimes para a distribuição acumulada da diversidade linguística, aqui apontamos apenas uma lei de escala e consequentemente um único expoente δ.

Figura 25 – Número de países N_D com uma diversidade linguística maior que D. A linha contínua é o melhor ajuste para $N_D \sim D^{-\delta}$ de onde $\delta = 1,26 \pm 0,03$ para $D > 60$. Dados extraídos da vigésima edição do *Ethnologue* publicada em 2017.



Fonte: Autor (2021).

Torna-se destacável o fato de que a região de transição entre os dois regimes reportados no último trabalho é a mesma onde observa-se uma pequena descontinuidade

na Figura 25. Muito embora o valor que reportamos para este expoente tenha sido obtido a partir do subconjunto de países cuja diversidade era maior que sessenta idiomas, a partir da Figura 25 podemos observar que este regime ($N_D \sim D^{-1,26}$) é incipiente já para valores inferiores de diversidade, isto é para valores de D inferiores aqueles que formam a descontinuidade supracitada. Tal percepção aponta para o fato de uma maior abrangência deste regime de escala.

A análise precedente é similar àquela utilizada em estudos de fragmentação (TURCOTTE, 1986; ÅSTRÖM; HOLIAN; TIMONEN, 2000), percolação (STAUFFER, 1985) e fenômenos de crescimento (VICSEK, 1992) onde assumindo $n(s)$ como o número de fragmentos ou aglomerados de tamanho s, temos

$$n(s) \sim s^{-\tau}. \tag{6.3}$$

Comparando as Equações 6.2 e 6.3 concluímos que $\delta = \tau - 1$ e consequentemente $\tau = 2,26$ para a diversidade linguística. Este valor do expoente é um pouco maior do que aquele da distribuição de tamanhos de aglomerados de percolação em duas dimensões ($\tau = 187/91$) e é igual, dentro da incerteza, ao obtido em simulações de avalanches em modelos de pilha de areia (MANNA, 1990). Este valor também é próximo dos valores encontrados em uma ampla gama de fenômenos tais como fragmentação nuclear (CAMPI; KRIVINE, 2005), formações geológicas (CAEL; SEEKELL, 2016), estudos de combustíveis (IGLAUER; PALUSZNY; BLUNT, 2013) e avalanches neuronais (YAGHOUBI et al., 2018).

6.2 Distribuições acumuladas do Produto Interno Bruto, área e população por país

Tendo discutido no Capítulo 3 a relação entre a diversidade linguística e medidas de tamanho segundo critérios econômico, geográfico e demográfico dos países, apresentamos nessa seção uma discussão relativa às distribuições acumuladas de tais grandezas. A questão do tamanho de uma unidade sócio-política tem sido objeto de interesse desde os tempos antigos. No Livro II de A Política, Aristóteles aponta que "É difícil - a experiência prova até que é quase impossível - que um Estado ou mesmo uma cidade muito povoada seja bem governada." (ARISTÓTELES, 2006). Nessa esteira, análises têm sido realizadas para compreender o possível balanço entre o benefício do tamanho de um país e os custos provenientes da heterogeneidade (ALESINA; SPOLAORE; WACZIARG, 2005).

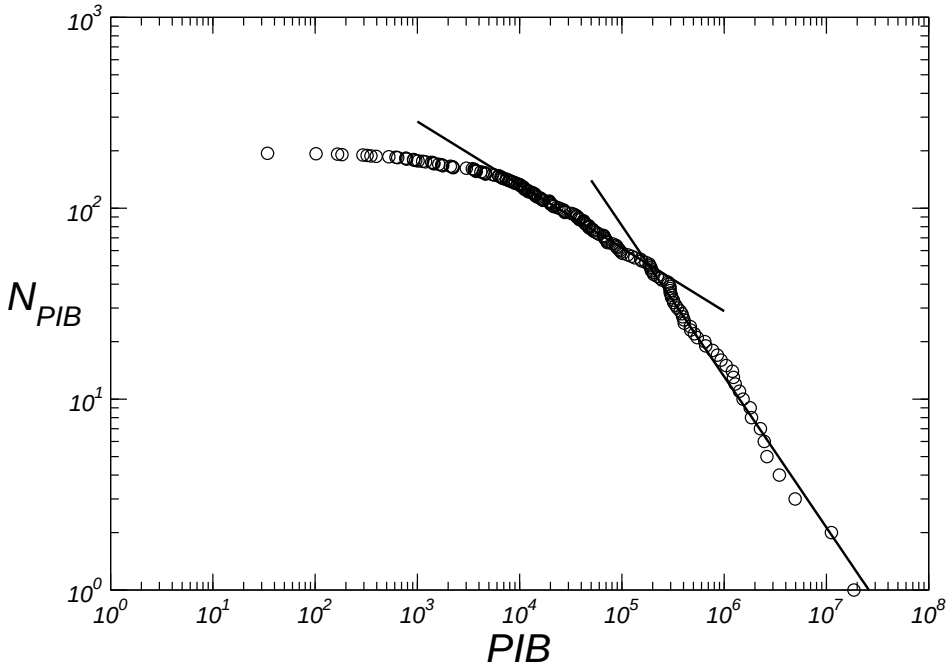
As distribuições de renda são um dos exemplos mais antigos e reconhecidos de lei de escala (CIRILLO, 1978). Como observado para a diversidade linguística, Figura 25, quando investigamos a distribuição acumulada de países como função do Produto Interno Bruto (PIB) relativo ao ano de 2016, a partir dos dados disponíveis nos Indicadores de Desenvolvimento Mundial (WORLD BANK, 2017a), observamos um comportamento

assintótico na forma de uma lei de potência. Sendo N_{PIB} o número de países com um PIB maior que o valor mostrado na abscissa temos

$$N_{PIB} \sim PIB^{-\pi} \tag{6.4}$$

com $\pi = 0,33 \pm 0,01$ para $8 \times 10^3 \leq PIB < 1 \times 10^5$ e $\pi = 0,79 \pm 0,01$ para $1 \times 10^5 \leq PIB$ conforme observado na Figura 26.

Figura 26 – Número de países N_{PIB} com um PIB maior que o valor indicado na abscissa. As linhas contínuas são os melhores ajustes para $N_{PIB} \sim PIB^{-\pi}$ de onde $\pi = 0,33 \pm 0,01$ para $8 \times 10^3 \leq PIB < 1 \times 10^5$ e $\pi = 0,79 \pm 0,01$ para $1 \times 10^5 \leq PIB$. Dados relativos ao ano de 2016 extraídos dos Indicadores de Desenvolvimento Mundial do Banco Mundial.



Fonte: Autor (2021).

A região de duas décadas de variabilidade do PIB onde $\pi \approx 4/5$ corresponde aos cinquenta e oito países que compõem 96% da economia do planeta. Este valor do expoente π indica que quanto maior a economia, menor o número de países capazes de sustentá-la. Podemos também compreender este resultado como apontando para o fato de que é cada vez mais difícil preservar a unidade de países com grande diversidade linguística (Figura 25), embora países de grande PIB possam ter grande diversidade linguística. Ou seja, aumentar a diversidade linguística leva, podemos dizer, a forças repulsivas (examinamos o efeito dessas forças repulsivas no Capítulo 5), as quais inviabilizam a estabilidade de muitos países mas, no entanto, existe um grupo de países que conseguem manter uma alta diversidade linguística num cenário de PIB igualmente alto. Não chega a ser um paradoxo, mas trata-se de uma possibilidade concreta possibilitada pelas leis de escala que definem o comportamento complexo do acoplamento entre o sistema de idiomas e o sistema econômico,

como atualmente distribuído entre todos os países. Este resultado está alinhado àqueles que apontam que a distribuição de renda mundial, mensurada pelo PIB *per capita*, também é descrita segundo leis de escala (DI GUILMI; GAFFEO; GALLEGATI, 2003; FURCERI, 2008) e que tem sido explicada em termos da integração das economias nacionais que é característica da globalização (CRISTELLI; BATTY; PIETRONERO, 2012).

Os quase 130 milhões de quilômetros quadrados de área total de terras emersas em nosso planeta estão distribuídos de modo não uniforme entre aproximadamente duas centenas de países. Com base nos dados disponíveis nos Indicadores de Desenvolvimento Mundial (WORLD BANK, 2017b), sabemos, por exemplo, que a Rússia tem uma área oitocentas mil vezes maior que Nauru, o menor país insular do mundo. Sendo N_A o número de países com área maior que A (medida em quilômetros quadrados) observamos um comportamento assintótico para essa distribuição acumulada que pode ser descrito segundo

$$N_A \sim A^{-\alpha} \tag{6.5}$$

com $\alpha = 0,96 \pm 0,01$ para os cinquenta e cinco países cujas áreas obedecem $3,6 \times 10^5 \leq A \leq 3 \times 10^6$ conforme a Figura 27. O limite superior dessa região foi tomado até a descontinuidade presente na região composta por aqueles seis países de áreas continentais (Rússia, China, Estados Unidos da América, Canadá, Brasil e Austrália) embora a inclusão destes países não altere o valor do expoente α. Quando observamos a região de valores intermediários da área, poderia ser sugerido um outro regime de escala (como na Figura 26) porém, conforme visto na Figura 29 e discutido mais à frente, somente a medida econômica apresenta tal regime duplo de leis de escala.

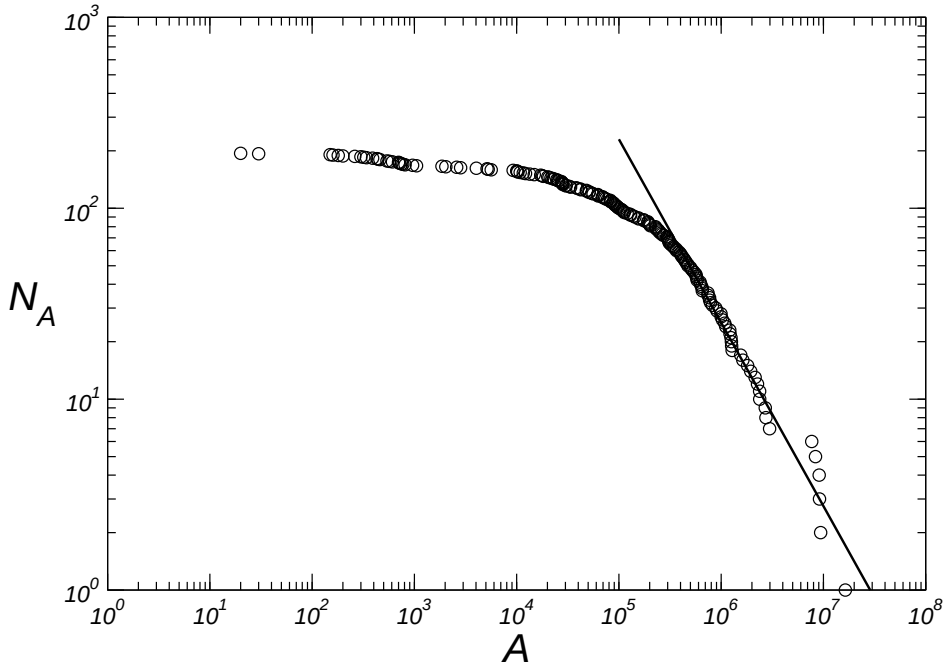
De igual modo, quando investigamos a distribuição acumulada de países como função da população destes, observamos um comportamento assintótico na forma de uma lei de potência. Com N_N sendo o número de países com população maior que N, temos

$$N_N \sim N^{-\nu} \tag{6.6}$$

com $\nu = 0,94 \pm 0,02$ para $N \geq 1,5 \times 10^7$ conforme a Figura 28. Nessa região de duas décadas de variabilidade na população estão contidos cinquenta e oito países. Este valor de ν bem próximo de 1 é tanto igual ao reportado anteriormente (ROSE, 2006) quanto ao supracitado valor de α considerando as incertezas associadas. Apontamos, junto com outros autores (ROSE, 2006; GONZÁLEZ-VAL; SANSO-NAVARRO, 2010), que embora questões sociopolíticas possam alterar drasticamente a área e população das entidades reconhecidas como países, os expoentes decorrentes das análises precedentes não sofrem alterações bruscas ao longo do intervalo de tempo da ordem de um século.

Muito embora movimentos populacionais intranacionais sejam mais comuns do que aqueles que ocorrem entre países, os resultados reportados nessa seção também reforçam que há um comportamento assintótico comum entre países e cidades, conforme apontando

Figura 27 – Número de países N_A com área maior que A. A linha contínua é o melhor ajuste para $N_A \sim A^{-\alpha}$ de onde $\alpha = 0,96 \pm 0,01$ para $3,6 \times 10^5 \leq A \leq 3 \times 10^6$. Dados extraídos dos Indicadores de Desenvolvimento Mundial do Banco Mundial.



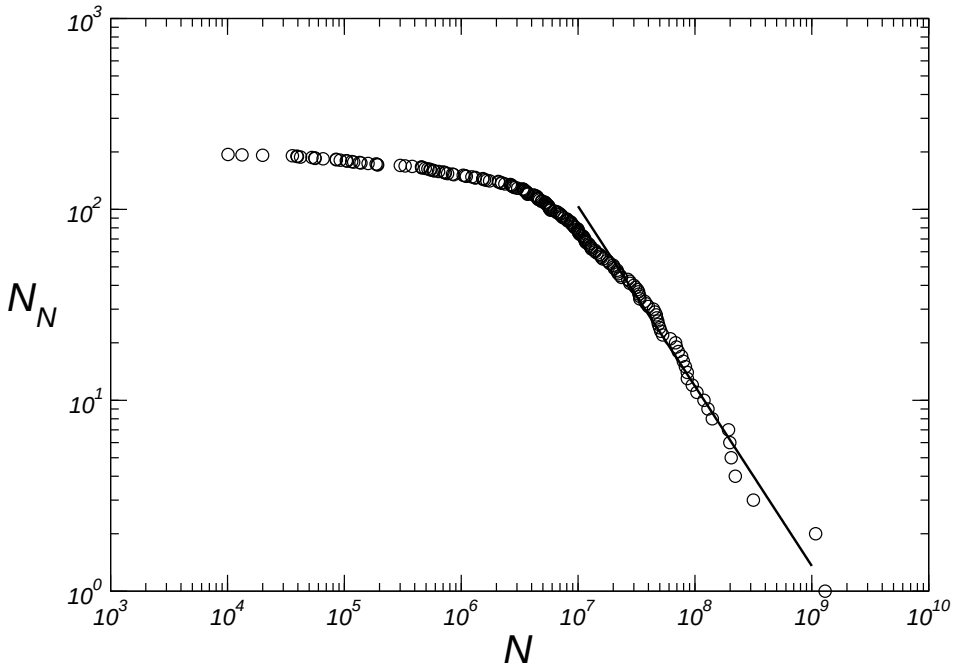
Fonte: Autor (2021).

anteriormente (GONZÁLEZ-VAL; SANSO-NAVARRO, 2010). Ademais é conhecido que o número de pessoas nascidas em um país diferente daquele em que vivem (no período 1960-2000), uma questão contemporânea fundamental, também segue uma distribuição zipfiana (CLEMENTE; GONZÁLEZ-VAL; OLLOQUI, 2011).

Observamos a partir das Figuras 26, 27 e 28 que as três grandezas analisadas nesta seção possuem comportamento assintótico qualitativamente similar. Quando tomamos a distribuição acumulada dos valores relativos dessas grandezas por país, observamos com maior clareza essa similaridade no comportamento colapsado das curvas obtido na Figura 29. Com flutuações presentes na região de grandes áreas, populações e economias (relativas), as três distribuições seguem aproximadamente a mesma curva. Este comportamento para área e população já é conhecido na literatura (TUNCAY, 2008), aqui acrescentamos a informação relativa ao Produto Interno Bruto. Tal observação mostra que, de certa forma, $A \sim N \sim PIB$.

Enquanto para área e população por país reportamos apenas um regime de escala, e consequentemente um expoente, observamos que a medida econômica, de dinâmica histórica mais recente, possui um regime de dupla lei de escala (Figura 29). Ademais, dado que para os grandes países temos $\pi \approx \alpha \approx \nu \approx 1$ (a linha contínua na Figura 29 é $N_P \sim (razão)^{-1}$) podemos associar tais grandezas ao expoente de Fisher $\tau = 187/91$ isto é, compreender a distribuição diferencial dos países segundo tais grandezas como uma aproximação de um processo percolativo em duas dimensões ($\tau = 2,05$).

Figura 28 – Número de países N_N com população maior que N. A linha contínua é o melhor ajuste para $N_N \sim N^{-\nu}$ de onde $\nu = 0,94 \pm 0,02$ para $N \geq 1,5 \times 10^7$. Dados extraídos da vigésima edição do *Ethnologue* publicada em 2017.



Fonte: Autor (2021).

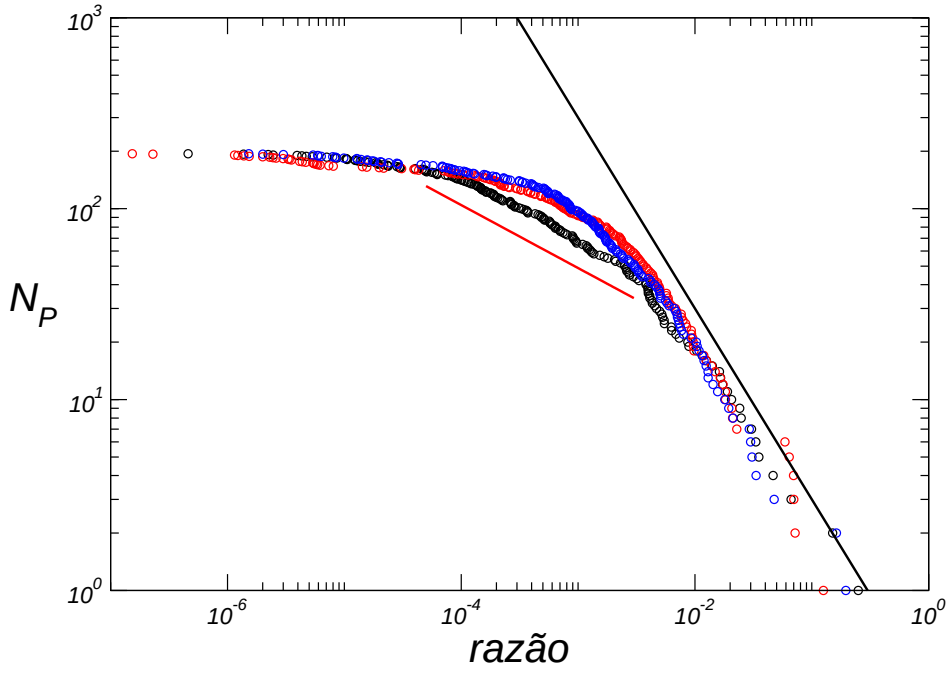
6.3 Distribuições acumuladas da população linguística e população por grupo étnico

Na seção anterior analisamos grandezas que estão vinculadas fortemente à questões geopolíticas. Agora, voltando-nos uma vez mais para as informações referentes ao conjunto de mais de 7000 idiomas, convém investigar a distribuição de tamanhos da população por idioma, ou população linguística, uma vez que essa quantidade não está diretamente limitada às fronteiras nacionais.

Conforme exposto na Seção 1.1, a vigésima edição do *Ethnologue* apresenta 7462[2] idiomas dentre os quais 6681 possuem pelo menos um falante que tem o idioma como seu primeiro idioma. Conforme observado na Figura 30, o histograma das populações destes 6681 idiomas abrange nove ordens de magnitude e é bem ajustado por uma distribuição log-normal. Este ajuste, típico de processos demográficos estocásticos, é um dos fatos empíricos quantitativos mais bem estabelecidos sobre os idiomas existentes (SUTHERLAND, 2003; SCHULZE; STAUFFER, 2005; ZANETTE, 2008; CLINGINGSMITH, 2017). A distribuição de frequência da área de alcance de um idioma também segue uma distribuição log-normal (GAVIN; STEPP, 2014). Esse tipo de distribuição está presente também em outros fenômenos linguísticos (TORRE et al., 2019), bem como em diversos outros sistemas

[2] 7099 idiomas vivos, 360 idiomas mortos e 3 sem informações.

Figura 29 – Número de países N_P com PIB (círculos pretos), área (círculos vermelhos) e população (círculos azuis) maiores que a razão $\frac{PIB_{pais}}{PIB_{Terra}}$; $\frac{Area_{pais}}{Area_{Terra}}$; $\frac{Populacao_{pais}}{Populacao_{Terra}}$. As linhas são $N_P \sim (razão)^{-1}$ (preta) e $N_P \sim (razão)^{-1/3}$ (vermelha).



Fonte: Autor (2021).

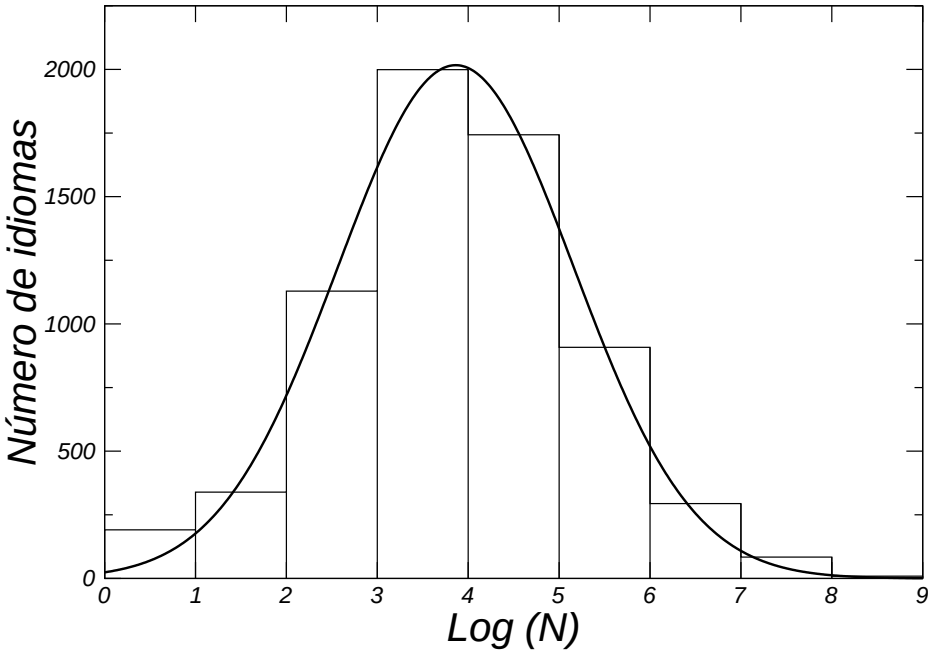
complexos (LIMPERT; STAHEL; ABBT, 2001). A distribuição exibida na Figura 30 tem mediana de 8000 falantes enquanto o tamanho médio de um idioma é 124 vezes maior que o valor dessa mediana. Este fato pode ser compreendido ao observarmos que os dez maiores idiomas são falados por aproximadamente $43,5\%$ da população mundial, conforme exposto na Tabela 2.

Embora o efeito do tamanho da população nos padrões e taxas de evolução dos idiomas seja complexo, foi reportado que idiomas com menos falantes possuem uma taxa de declínio mais acentuada do que idiomas com grandes populações (SUTHERLAND, 2003). Uma possível explicação para este efeito Allee linguístico pode vir da demonstração recente de que idiomas que possuem populações maiores têm taxas mais altas de ganho de novas palavras, enquanto populações menores têm taxas mais altas de perda de palavras (BROMHAM et al., 2015). Esta é mais uma questão que acende um alerta quanto a necessária atenção a ser dada aos idiomas em processo de extinção (HARRISON, 2007; EVANS, 2010).

Uma outra maneira de investigar essa distribuição de tamanhos das populações linguísticas é mediante o estudo da sua distribuição acumulada, isto é do número de idiomas N_L com população linguística maior que N. Observamos que, de modo similar às grandezas discutidas nas seções anteriores, assintoticamente N_L pode ser descrito segundo

$$N_L \sim N^{-\iota}, \tag{6.7}$$

Figura 30 – Distribuição de frequência das populações N dos idiomas vivos em escala logarítmica com uma distribuição log-normal ajustada. Dados extraídos da vigésima edição do *Ethnologue* publicada em 2017.



Fonte: Autor (2021).

com um comportamento composto por dois regimes de escala de modo que $\iota = 0,57 \pm 0,01$ para $1 \times 10^5 \leq N \leq 1 \times 10^7$ e $\iota = 0,97 \pm 0,01$ para $1 \times 10^7 < N \leq 1 \times 10^9$ conforme Figura 31. A região onde ι é próximo a unidade, ou seja, grandes tamanhos de populações linguísticas, é composta pelos 90 maiores idiomas adotados por 80% da população mundial. Por outro lado, a região de tamanhos intermediários de populações, ou seja, $\iota = 0,57$, é composta por 1263 idiomas.

Na Figura 31 estendemos com linhas tracejadas os domínios das leis de escala reportadas acima. O trecho tracejado em azul mostra que o ramo da lei de escala representado pelo regime azul contínuo perde a estabilidade por volta de populações da ordem de 10 milhões de pessoas, indicando que em um sistema finito como a Terra o número de idiomas não poderia continuar crescendo tão rapidamente de modo a ficarmos com um número exorbitante de pessoas falando uma quantidade enorme de idiomas. A mesma coisa é válida em seguida, para o trecho tracejado em vermelho, que deixa de valer próxima a 100 mil pessoas, o qual propiciará a transição para a saturação no número total de idiomas observado.

Além da população linguística, o *Ethnologue* fornece a população de 1280 grupos étnicos de pessoas que se identificam como parte da etnia associada a um idioma específico. As populações que constituem estes grupos étnicos têm tamanhos que se distribuem por oitos décadas de variabilidade. Se tomarmos N_E como o número de grupos étnicos com

Tabela 2 – Dez maiores idiomas de acordo com a população linguística N (número de falantes). O valor $N_{\%}$ aponta o percentual da população mundial que adota o idioma como primeira língua. Dados extraídos da vigésima edição do *Ethnologue* publicada em 2017.

Idioma	**Família**	N	$N_{\%}$
Mandarim	Sino-Tibetana	897902930	$13,5\%$
Espanhol	Indo-Europeia	436667750	$6,6\%$
Inglês	Indo-Europeia	371959910	$5,6\%$
Hindi	Indo-Europeia	260129750	$3,9\%$
Bengali	Indo-Europeia	242315050	$3,6\%$
Português	Indo-Europeia	218765470	$3,3\%$
Russo	Indo-Europeia	153612510	$2,3\%$
Japonês	Japônica	128193360	$1,9\%$
Punjabi	Indo-Europeia	92721700	$1,4\%$
Javanês	Austronesiana	84368500	$1,3\%$

Fonte: Autor (2021).

população maior que N, o comportamento assintótico de N_E pode ser descrito como

$$N_E \sim N^{-\epsilon}. \tag{6.8}$$

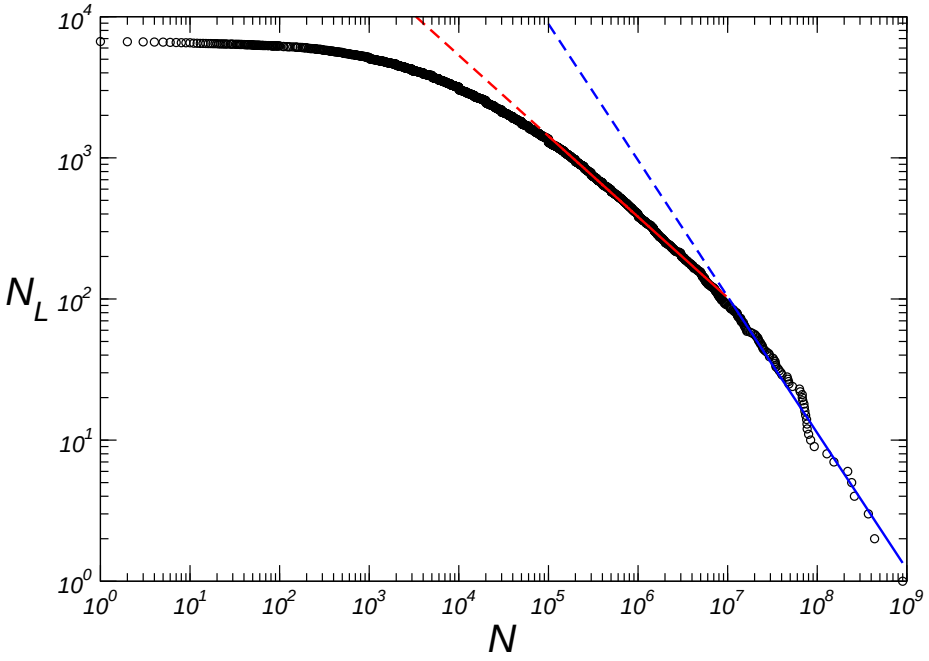
Aqui, similarmente à distribuição acumulada do Produto Interno Bruto e da população linguística, temos um regime de dupla lei de escala com $\epsilon = 0,45 \pm 0,01$ para $1 \times 10^4 \leq N < 5 \times 10^6$ e $\epsilon = 0,78 \pm 0,02$ para $N \geq 5 \times 10^6$ conforme a Figura 32.

6.4 Distribuição de tamanho de famílias de idiomas

Em um paralelo com os estudos biológicos, uma família linguística pode ser compreendida como um grupo formado por idiomas que se desenvolveram a partir de um ancestral comum conhecido como protolíngua (CAMPBELL; MIXCO, 2007). Dentro de uma determinada família, os idiomas compartilham certas características linguísticas como palavras, sons e padrões gramaticais (PERELTSVAIG, 2012). Essa taxonomia da linguagem permite construir árvores linguísticas que guardam diversas semelhanças com árvores genéticas (CAVALLI-SFORZA, 1997). Uma característica importante a ser observada é a amplitude do número de idiomas em cada família que varia da unidade, como na família Carajá, até mais de mil idiomas, como na família Nigero-Congolesa. Com relação a essa diferença de tamanho entre famílias linguísticas, Greenhill listou cinco possíveis explicações: idade da família, tamanho da população, tecnologia (hipótese da dispersão agricultura/linguagem), geografia-e-ecologia e fatores sociais (GREENHILL, 2014).

O *Ethnologue* classifica 6711 idiomas vivos em 141 famílias. Em nossa análise não foram incluídos 388 idiomas classificados nas categorias especiais para línguas construídas,

Figura 31 – Número de idiomas N_L com população maior que N. As retas contínuas fornecem o melhor ajuste para $N_L \sim N^{-\iota}$ com $\iota = 0,57 \pm 0,01$ (reta vermelha) e $\iota = 0,97 \pm 0,01$ (reta azul). Dados extraídos da vigésima edição do *Ethnologue* publicada em 2017.



Fonte: Autor (2021).

crioulos, idiomas de sinais, idiomas isolados, línguas mistas, pidgins e línguas não classificadas. Atribuímos a classificação $r = 1$ para a família Nigero-Congolesa composta por 1526 idiomas, classificação $r = 2$ para a família Austronesiana composta por 1224 idiomas, classificação $r = 3$ para a família Trans-Neo Guineana composta por 478 idiomas e assim seguindo decrescendo sucessivamente. Dessa forma podemos escrever o tamanho cardinal, ou seja, número de idiomas, N_F que compõem uma família de classificação r segundo

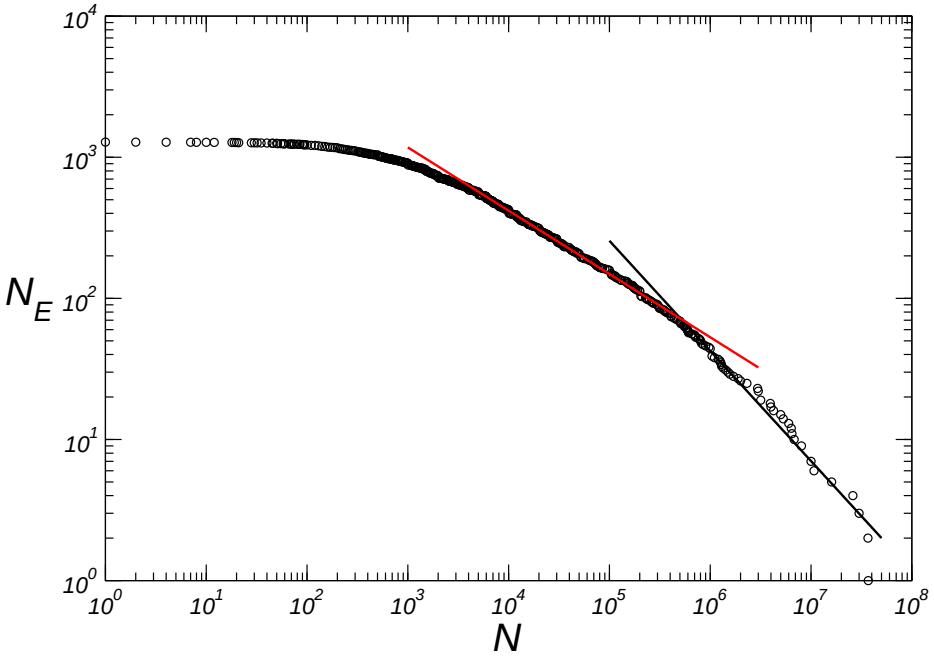
$$N_F \sim r^{-\theta} \tag{6.9}$$

com $\theta = 2$ conforme observado na Figura 33. Este expoente é o mesmo do comportamento assintótico da distribuição de Cauchy.

A distribuição do tamanho de famílias linguísticas que reportamos é similiar a curva de distribuição de famílias biológicas, também chamada de *Hollow Curve* (GREENHILL, 2014). O expoente θ obtido aqui é igual dentro da incerteza ao valor ($\theta = 1,905$) obtido anteriormente por Wichmann (WICHMANN, 2005) porém maior do que o valor ($\theta = 1,38$) obtido por Hammarström a partir de uma base de dados diferente (HAMMARSTRÖM, 2010). Este nosso resultado se contrapõe àquele de Zanette, que duas décadas atrás, analisando um conjunto de dezessete famílias, propôs que N_F decairia exponencialmente com a classificação r da família (ZANETTE, 2001).

Hammarström aponta também que uma vez que a diferenciação linguística ocorre

Figura 32 – Número de grupos étnicos N_E com população maior que N. As retas contínuas fornecem o melhor ajuste para $N_E \sim N^{-\epsilon}$ com $\epsilon = 0,45 \pm 0,01$ (reta vermelha) e $\epsilon = 0,78 \pm 0,02$ (reta preta). Dados extraídos da vigésima edição do *Ethnologue* publicada em 2017.



Fonte: Autor (2021).

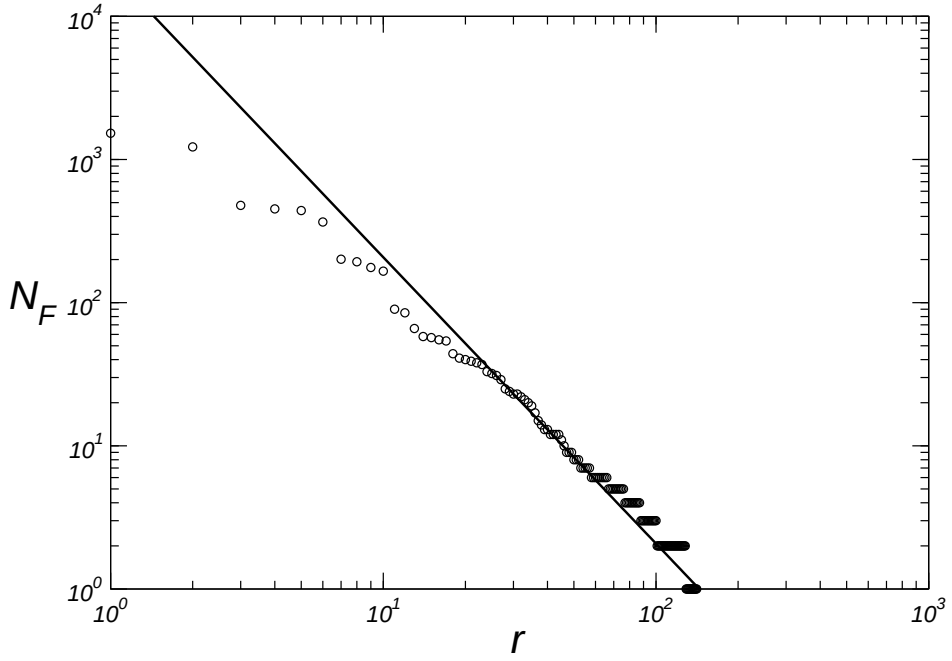
principalmente por meio de migração humana, o tamanho cardinal de uma família pode ser considerado uma medida da propagação difusiva de uma família. A partir dessa perspectiva podemos compreender que apenas um pequeno número de famílias linguísticas estão difundidas por grandes áreas sobre a superfície terrestre. Paralelamente, é possível compreender que a maior parte das famílias têm um pequeno alcance geográfico.

Uma outra maneira de medir o tamanho de uma família linguística é tomando o número N_S de falantes totais dos idiomas membros de cada família. Num tratamento similar ao apresentado ao longo deste capítulo, temos que N_S escala com a classificação r da família segundo

$$N_S \sim r^{-\psi}. \tag{6.10}$$

A Figura 34 aponta $\psi = 1,6$ para as doze maiores famílias linguísticas segundo o número de falantes. Essas famílias contém os 4603 idiomas que são adotados como primeira língua por $98,75\%$ da população mundial. Este conjunto corresponde a $68,59\%$ dos idiomas vivos atualmente. Após essa região observamos uma descontinuidade que é seguida por uma região para a qual não reportamos o expoente mas que visualmente decresce com uma taxa maior que aquela vista na região das maiores famílias. Essa segunda região é composta pelas famílias ameaçadas de extinção e moribundas (WHALEN; SIMONS, 2012) como, por exemplo as famílias Tupiana, Uto-asteca e Nambiquara. Tais famílias linguísticas

Figura 33 – Número de idiomas N_F por família linguística como função da classificação r segundo o número de idiomas. A linha contínua é a curva $N_F \sim r^{-\theta}$ com $\theta = 2$. Dados extraídos da vigésima edição do *Ethnologue* publicada em 2017.



Fonte: Autor (2021).

podem estar ligadas a eventos históricos que eliminaram bruscamente muitos membros das populações como guerras, pestes e genocídio. Conjecturamos que na ausência destes eventos traumáticos, o regime de escala se estenderia pelo segmento tracejado exposto na Figura 34.

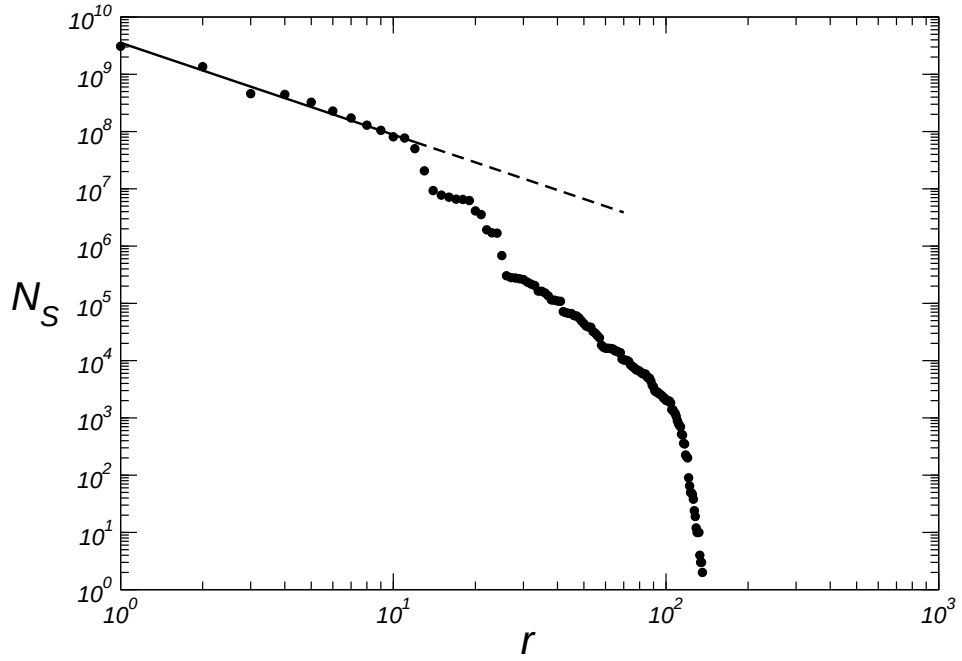
Salientamos ainda que o processo crescente de extinção de famílias linguísticas é um evento ainda mais danoso que a extinção de idiomas. Conforme apontam Campbell e Belew,

> The loss of a specific language may be likened to the loss of a single species, say the Bengal tiger or the Right whale. However, the extinction of whole families of languages is a tragedy analogous to the loss of whole branches of the animal kingdom, say to the loss of all felines or all cetaceans.[3] (CAMPBELL; BELEW, 2018).

Portanto, conhecer quais famílias estão se extinguindo mais rapidamente permite entender quais idiomas em rota de extinção devem ser priorizados.

[3] "A perda de um idioma específico pode ser comparada à perda de uma única espécie, digamos, o tigre de bengala ou a baleia-franca. No entanto, a extinção de famílias inteiras de línguas é uma tragédia análoga à perda de ramos inteiros do reino animal, digamos, a perda de todos os felinos ou de todos os cetáceos."

Figura 34 – Número de falantes N_S por família linguística como função da classificação r segundo o número de falantes. A linha contínua é a curva $N_S \sim r^{-\psi}$ com $\psi = 1,6$. Dados extraídos da vigésima edição do *Ethnologue* publicada em 2017.



Fonte: Autor (2021).

6.5 Um olhar adicional sobre distribuições de Zipf para famílias linguísticas

A Figura 33 permite ver que apenas dez famílias são compostas por mais de cem idiomas. Convém perguntar se o regime observado na Figura 34 é observado também intrafamílias. Se sim, isso implicaria que fazendo a classificação dos idiomas segundo a população linguística N devemos observar

$$N \sim r^{-\kappa} \tag{6.11}$$

onde r é a classificação do idioma e κ é o expoente característico. A Tabela 3 apresenta o valor deste expoente para as catorze maiores famílias segundo o número de idiomas. Abaixo discutimos cada família especificamente.

A maior família linguística, a Nigero-Congolesa, é composta por 1526 idiomas e compreende quase todos os idiomas nativos da África abaixo do Saara. Essa família é caracterizada por $\kappa = 1,2$ conforme a Figura 35 (a). A outra família composta por mais de mil idiomas é a Austronesiana que tem seus 1224 idiomas espalhados desde a Indonésia

passando pela ilha da Nova Guiné até a Ilha de Páscoa. Conforme observado na Figura 35 (b) esta família, originária da região de Taiwan, tem $\kappa = 1,6$. A ilha de Nova Guiné também é o lar da família Trans-Neo Guineana que é composta por 478 idiomas e é caracterizada por $\kappa = 1,1$ como exibido na Figura 35 (c). A família Sino-Tibetana possui 452 idiomas, porém conta com mais de 380 vezes mais falantes do que a família Trans-Neo Guineana, e conforme Figura 35 (d), tem $\kappa = 1,7$. A família Indo-Europeia inclui quase todos os idiomas da Europa bem como muitos idiomas no continente Asiático. Entre os vinte maiores idiomas do mundo, onze são pertencentes a esta família e no total são mais de três bilhões de falantes distribuidos em 440 idiomas, conforme exposto na Figura 35 (e), tal que $\kappa = 1,7$.

Os 366 idiomas que compõem a família Afro-Asiática provavelmente descendem da língua falada pelos grupos humanos que migraram do continente Africano para o Oriente Médio há mais de 50000 anos. Foi em fenício, um idioma dessa família, que o primeiro alfabeto fonético foi construído. Diferente das cinco maiores famílias que foram caracterizadas por dois valores do expoente ($\kappa = 1,15 \pm 0,05$ e $\kappa = 1,65 \pm 0,05$), e de acordo com Figura 35 (f), a família Afro-Asiática tem $\kappa = 2,6$, sendo este valor o maior entre as catorze maiores famílias linguísticas. Ao sul da região dos idiomas Afro-Asiáticos se encontra outra família cujo valor de κ também é distinto daqueles dois característicos das cinco maiores famílias. Composta por mais de 200 idiomas, a família Nilo-Saariana tem $\kappa = 1,4$, conforme a Figura 36 (a).

A família Australiana embora muito diversa em número de idiomas, 193 no total, tem decrescido bastante em número de falantes nos últimos séculos, sendo a menor família segundo este critério dentre as famílias aqui discutidas. Composta por idiomas menores do que dez mil falantes, conforme a Figura 36 (b), temos $\kappa = 1,6$, valor igual aquele reportado para a família Austroneniana. A família Otomangueana característica do atual território mexicano, e que também teve sua população reduzida após processos colonizadores, tem $\kappa = 1,1$, conforme a Figura 36 (c). Este valor é igual aquele reportado para a família Trans-Neo Guineana.

A família Austro-Asiática, falada do sudeste da Ásia até o leste da Índia, é a nona família com mais falantes no mundo e, conforme a Figura 36 (d), tem $\kappa = 2,0$. O sudeste asiático é também a região da família Tai-Kadai, que possuindo cerca de metade do número de idiomas da família Austro-Asiática, é caracterizada por $\kappa = 1,2$ conforme a Figura 37 (a). A família Dravidiana, típica do sul da Ásia, tem uma população linguística de mais de duzentos milhões de falantes. Pertencem a esta família muitos dos idiomas remanescentes após a expansão e domínio dos idiomas Indo-Europeus na região da Índia. De acordo com a Figura 37 (b) a família Dravidiana é caracterizada por $\kappa = 2,0$, expoente igual aquele da família Austro-Asiática.

Os processos decorrentes da colonização afetaram de modo irreversível a distribui-

ção de idiomas no continente americano. Duas das famílias linguísticas mais atingidas foram a Tupiana, na América do Sul, e a Uto-Asteca, na América do Norte. Estas duas famílias são caracterizadas por $\kappa = 2,1$, conforme a Figura 37 (c) e (d) respectivamente. Juntamente com as famílias Austro-Asiática e Dravidiana, essas duas famílias constituem um quadrupleto caracterizado por $\kappa = 2,05 \pm 0,05$.

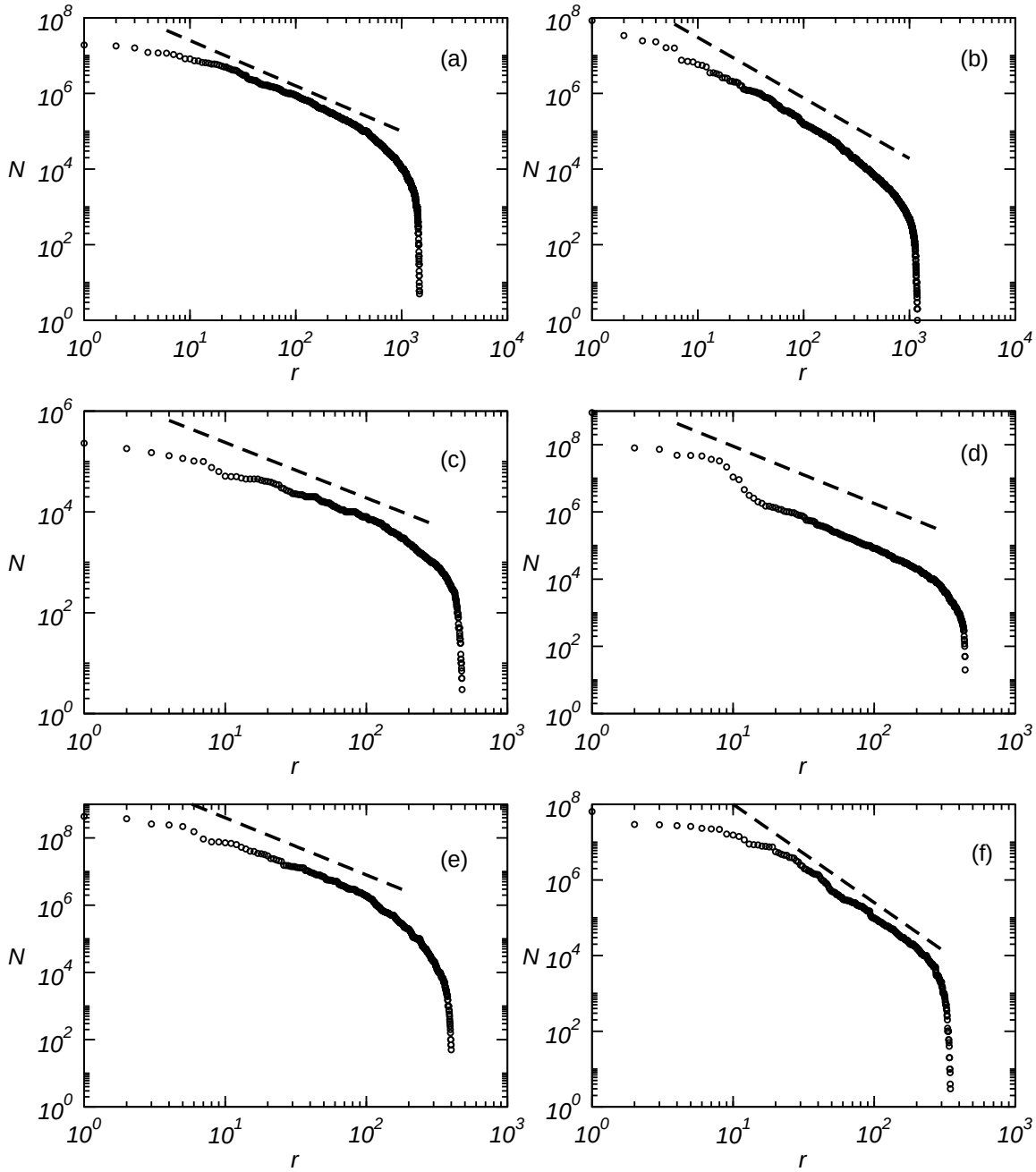
Futuras investigações, com particular ênfase nos processos migratórios humanos, deverão buscar compreender (*i*) porque as doze das maiores famílias linguísticas constituem três quadrupletos bem caracterizados segundo os valores dos expoentes ($\kappa = 1,15 \pm 0,05$, $\kappa = 1,65 \pm 0,05$ e $\kappa = 2,05 \pm 0,05$) e (*ii*) porque as famílias Afro-Asiática e Nilo-Saariana, ambas do continente africano, possuem valores de expoentes diferentes daqueles dos três quadrupletos supracitados.

Tabela 3 – Catorze maiores famílias linguísticas segundo o número de idiomas N_F. O valor r indica a classificação do idioma segundo este ordenamento e r_s aponta a classificação da família segundo a população linguística N. O valor de κ é o expoente da lei de escala $N \sim r^{-\kappa}$ (Figuras 35, 36 e 37). Dados extraídos da vigésima edição do *Ethnologue* publicada em 2017.

r	**Família**	N_F	N **(em milhões)**	r_s	κ
01	Nigero-Congolesa	1475	458,90	03	1,2
02	Austronesiana	1224	324,88	05	1,6
03	Trans-Neo Guineana	478	3,55	21	1,1
04	Sino-Tibetana	452	1355,71	02	1,7
05	Indo-Europeia	440	3077,11	01	1,7
06	Afro-Asiática	366	444,85	04	2,6
07	Nilo-Saariana	201	50,33	12	1,4
08	Australiana	193	0,04	12	1,6
09	Otomangueana	176	1,68	24	1,1
10	Austro-Asiática	166	104,99	09	2,0
11	Tai-Kadai	90	80,1	10	1,2
12	Dravidiana	85	228,1	06	2,0
13	Tupiana	66	6,2	19	2,1
14	Uto-Asteca	58	1,9	22	2,1

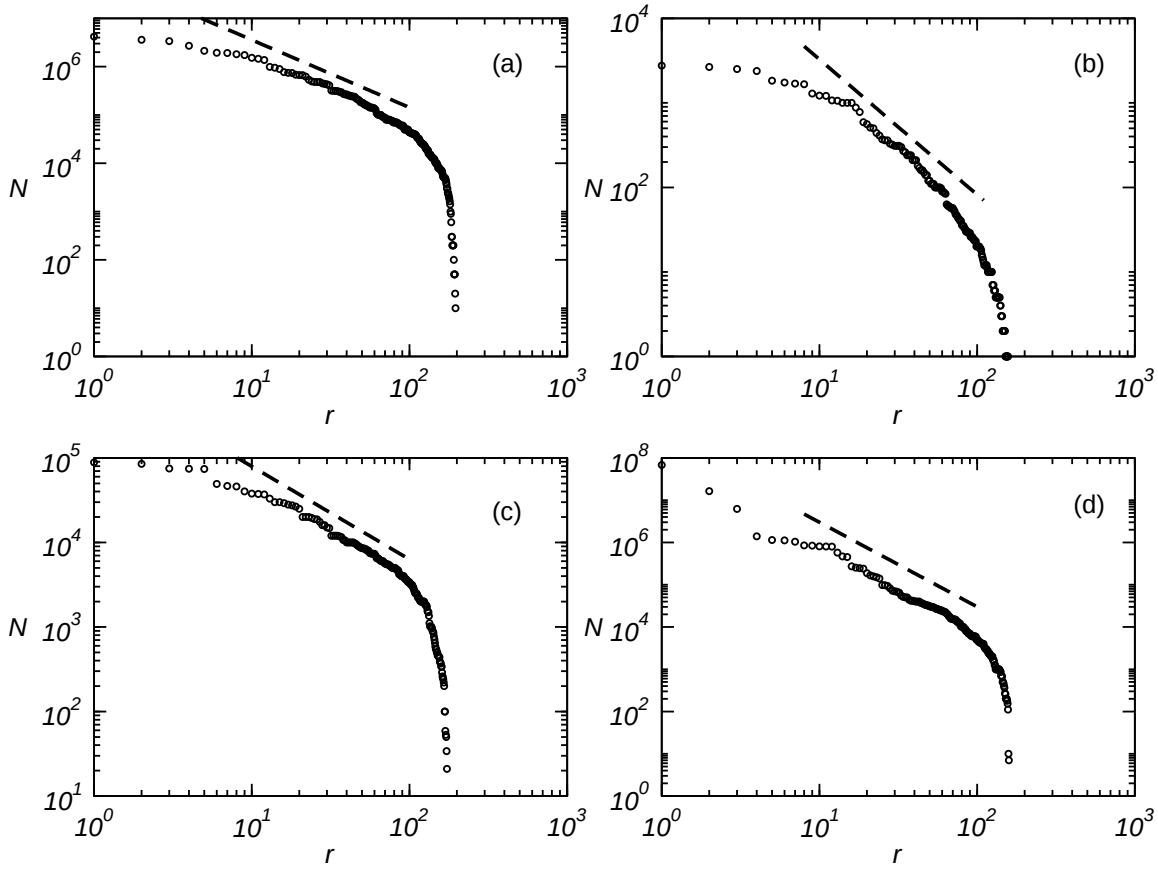
Fonte: Autor (2021).

Figura 35 – Número de falantes N por idioma como função da classificação r. As linhas tracejadas fornecem o ajuste para $N \sim r^{-\kappa}$ para as famílias: (a) Nigero-Congolesa ($\kappa = 1,2$); (b) Austronesiana ($\kappa = 1,6$); (c) Trans-Neo Guineana ($\kappa = 1,1$); (d) Sino-Tibetana ($\kappa = 1,7$); (e) Indo-Europeia ($\kappa = 1,7$) e (f) Afro-Asiática ($\kappa = 2,6$). Dados extraídos da vigésima edição do *Ethnologue* publicada em 2017.



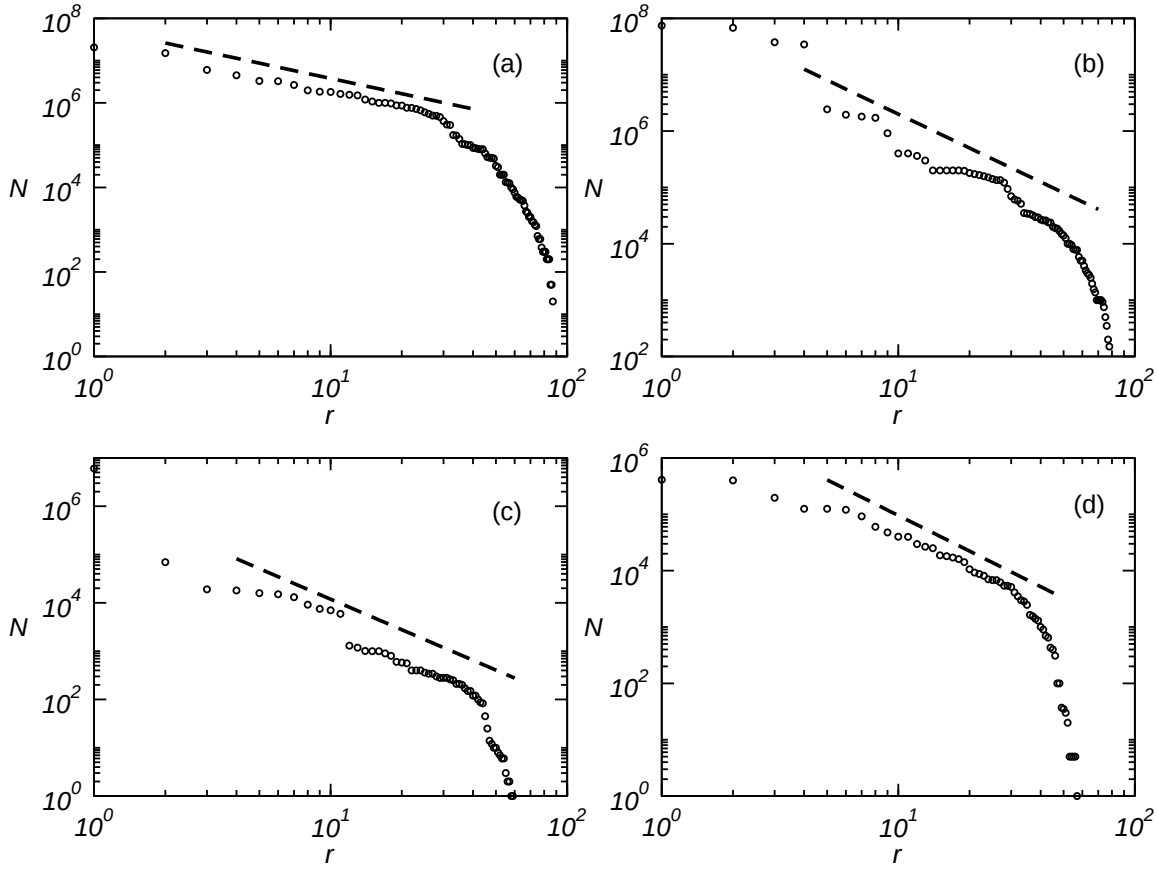
Fonte: Autor (2021).

Figura 36 – Número de falantes N por idioma como função da classificação r. As linhas tracejadas fornecem o ajuste para $N \sim r^{-\kappa}$ para as famílias: (a) Nilo-Saariana ($\kappa = 1,4$); (b) Australiana ($\kappa = 1,6$); (c) Otomangueana ($\kappa = 1,1$) e (d) Austro-Asiática ($\kappa = 2,0$). Dados extraídos da vigésima edição do *Ethnologue* publicada em 2017.



Fonte: Autor (2021).

Figura 37 – Número de falantes N por idioma como função da classificação r. As linhas tracejadas fornecem o ajuste para $N \sim r^{-\kappa}$ para as famílias: (a) Tai-Kadai ($\kappa = 1, 2$); (b) Dravidiana ($\kappa = 2, 0$); (c) Tupiana ($\kappa = 2, 1$) e (d) Uto-Asteca ($\kappa = 2, 1$). Dados extraídos da vigésima edição do *Ethnologue* publicada em 2017.



Fonte: Autor (2021).

6.6 Conclusões

Neste capítulo reportamos leis de escala para as distribuições acumuladas da diversidade linguística (Figura 25), área (Figura 27) e população (Figura 28) dos países, atualizando alguns resultados apresentados duas décadas atrás (GOMES et al., 1999). Os expoentes associados à tais distribuições de tamanho são, respectivamente, $\delta = 1,26$, $\alpha = 0,96$ e $\nu = 0,94$. Observamos que a distribuição acumulada de tamanhos de PIB (Figura 26), cuja dinâmica histórica é mais recente, é bem descrita por uma dupla lei de escala sendo $\pi = 0,33$ e $\pi = 0,79$ os expoentes na primeira e segunda região, respectivamente. Apresentamos que, assintoticamente, as distribuições dos parâmetros geográfico, demográfico e econômico possuem um comportamento qualitativo similiar (Figura 29) cujo valor do expoente se aproxima ao da percolação em duas dimensões.

Mostramos em seguida que tanto a distribuição acumulada da população linguística (Figura 31) quanto a da população étnica (Figura 32), aqui apresentada pela primeira vez na literatura, são descritas por leis de potência de regime duplo com expoentes não muito diferentes.. Depois reproduzimos a lei de potência do tipo Zipf (Figura 33) com expoente 2 para as famílias linguísticas classificadas segundo o número de idiomas, reportada inicialmente por Wichmann (WICHMANN, 2005), e apresentamos a inédita lei de potência (Figura 34) com expoente $1,6$ para as famílias linguísticas classificadas segundo o número de falantes.

Por fim, foram apresentados os resultados inéditos relativos às distribuições de Zipf (Figuras 35, 36 e 37) para as catorze maiores famílias linguísticas contemporâneas. Apontamos que, com exceção das famílias Afro-Asiática e Nilo-Saariana, temos três quadrupletos agrupados segundo o expoente das distribuições de Zipf. O primeiro composto pelas famílias Trans-Neo Guineana, Otomangueana, Nigero-Congolesa e Tai-Kadai é caracterizado por $\kappa = 1,15 \pm 0,05$. Com $\kappa = 1,65 \pm 0,05$, o segundo quadrupleto reúne as famílias Austronesiana, Australiana, Sino-Tibetana e Indo-Europeia. Já o terceiro quadrupleto com as famílias Austro-Asiática, Dravidiana, Tupiana e Uto-Asteca tem $\kappa = 2,05 \pm 0,05$.

7 CONCLUSÕES

"Me pregunto cuál será la lengua de mi senilidad,
si en ella caigo, y en qué lengua moriré."[1]

Sylvia Molloy

Nesta tese apresentamos nossas contribuições ao estudo de alguns aspectos da linguagem. Exploramos certas características da distribuição da diversidade linguística na Terra, com particular atenção dada a relação diversidade linguística-área (SANTOS; GOMES, 2019; SANTOS; GOMES, 2020; SANTOS; GOMES, 2021a; SANTOS; GOMES, 2021b). Além de um assunto de interesse teórico, a diversidade linguística, e de modo mais amplo a diversidade cultural, guarda interessantes conexões com a biodiversidade. Ademais, diante da ação antrópica, a preservação da diversidade biocultural será um tema recorrente ao longo do presente século, como já exemplificado pela Assembleia Geral das Nações Unidas que declarou o ano de 2008 como Ano Internacional das Línguas e o ano de 2019 como Ano Internacional das Línguas Indígenas.

Iniciamos revisando a relação diversidade linguística D e área A a partir dos registros disponíveis no *Ethnologue* para mais de 7000 idiomas distribuídos em quase duas centenas de países (Seção 3.1). Apontamos que a relação de escala $D \sim A^z$ é robusta, seja quando analisamos os dados não categorizados (Figura 9) ou quando agrupamos os países em categorias segundo a área (Figura 10). O expoente obtido foi $z = 0,33 \pm 0,03$ para cinco décadas de variabilidade da área. Relação de escala similar é observada entre a diversidade de espécies biológicas e a área (HOBOHM, 2003). Como consequência desta lei de escala, apontamos a redução da densidade ρ_D de idiomas vivos com o aumento da área, $\rho_D \sim A^{-0,67}$ (Figura 11), que é um comportamento típico de sistemas de natureza fractal.

Assumindo que os deslocamentos das populações migrantes primitivas deveriam combinar algum tipo de aleatoriedade com algum tipo de auto-exclusão, propusemos um modelo fractal para justificar o supracitado expoente $z = 1/3$ (Seção 3.3). Tomando a caminhada aleatória auto-excludente em uma superfície bidimensional como o motivo estrutural típico na base destes deslocamentos e observando que o conjunto de idiomas está distribuído em um conjunto fractal de dimensão $d' = 2z = 0,66$, mostramos que a complementaridade entre as estruturas fractais da distribuição de idiomas e das trajetórias das populações migrantes poderia levar a uma estabilidade razoável do sistema de idiomas. Dessa forma, sugerimos que a relação de escala $D - A$ atualmente observada pode ser compreendida como uma relíquia de tempos primordiais em que populações humanas

[1] "Pergunto-me qual será a língua da minha senilidade – se ela me pegar – e em que língua morrerei."

migravam perfazendo caminhadas que se aproximavam de caminhadas aleatórias auto-excludentes em duas dimensões. Conjecturamos que esse tipo de argumento pode ser útil em algumas situações de interesse biológico como exemplificado pela dinâmica de uma presa que busca fugir de um predador cujo deslocamento se aproxime de uma caminhada aleatória auto-excludente.

A segunda lei de escala alométrica explorada foi aquela que relaciona a diversidade linguística com um parâmetro demográfico, aqui quantificado pela população N dos países (Seção 3.2). Reportamos que $D \sim N^{\nu}$ com $\nu = 0,41 \pm 0,03$ para pouco mais de cinco décadas de variabilidade da população (Figura 13). Extrapolando o referido ajuste, especulamos que uma população da ordem de quinhentas pessoas definiria o conjunto mínimo relacionado à existência de um idioma. Mostramos que a partir dessas duas leis alométricas recuperamos $N \sim A^{0,80}$. Este resultado é tanto igual ao obtido diretamente a partir dos dados $N \times A$ quanto ao apresentado na literatura. Por fim, isto nos permitiu afirmar que a densidade populacional terrestre ρ_N não é constante mas segue a relação de escala $\rho_N \sim A^{-0,20}$ e que a população humana está distribuída sobre a superfície terrestre em um conjunto fractal de dimensão $\delta = 1,6 \pm 0,2$. Este valor nos traz à mente aquele da dimensão do esqueleto do aglomerado de percolação em duas dimensões[2], um aspecto que leva a reflexões à luz da explicação do expoente z comentada no parágrafo anterior.

Reportamos a existência de uma relação de escala entre o tamanho da economia de um país, medida pelo Produto Interno Bruto (PIB), e o número de idiomas falado em seu território (Seção 3.4). De maneira semelhante às relações com os aspectos geográficos e demográficos, mostramos que $D \sim (PIB)^{\zeta}$ com $\zeta = 0,31 \pm 0,03$ ao longo de mais de cinco décadas de variabilidade do parâmetro econômico (Figura 15). Portanto, os valores dos expoentes que relacionam $D \times A$ e $D \times PIB$ são os mesmos dentro das barras de erro e são robustos independentemente da categorização. Este resultado que aponta que as maiores economias estão estatisticamente relacionadas a uma maior diversidade linguística está alinhado ao crescente reconhecimento de que uma maior presença da diversidade cultural está positivamente correlacionada com maiores benefícios comerciais e econômicos. Não nos furtamos, no entanto, do reconhecimento do caráter polêmico que frequentemente tende a assumir as discussões que buscam relacionar os índices econômicos à qualidade de vida e outros indicadores sociais.

No Capítulo 4, apontamos que $D \sim A^{z}$, com $z = 1/2$, se nos restringirmos ao conjunto dos 147 maiores países caracterizados por área superior a 18000 km^2 (Figura 16). De acordo com muitos outros casos em que um procedimento de maximização é responsável pelo surgimento de leis de escala, introduzimos então um modelo heurístico que emerge de um tipo de quadro variacional para resolver o problema da relação entre diversidade linguística e área para este conjunto de países. Mostramos que a introdução desse *ansatz*

[2] $D_B = 1,60 \pm 0,05$ (HERRMANN; HONG; STANLEY, 1984).

aritmético-geométrico simples foi tanto capaz de reproduzir a lei de escala com $z = 1/2$, quanto mostrar que a área associada à existência de um idioma é a região suficiente para garantir a vida de aproximadamente quinhentos caçadores-coletores conforme discutido na Seção 3.2. Este tamanho de população coincide com estimativas feitas por Hamilton et al. (HAMILTON et al., 2007), também reportadas em Gavin et al. (GAVIN et al., 2013). Gostaríamos de notar que este tamanho mínimo de uma população não é tão diferente do número mínimo estimado de migrantes extraplanetários (MARIN; BELUFFI, 2018; SALOTTI, 2020).

Propusemos no Capítulo 5 um modelo teórico simples que nos permitiu compreender melhor as características específicas da distribuição da diversidade linguística-área, em termos de forças entrópicas e de auto-exclusão. A linguagem compartilha com outros fenômenos biológicos e culturais, o entendimento de que duas forças opostas podem atuar gerando diversidade. É um fato notável que os dados empíricos explicitados na Figura 23 concordam, de forma qualitativa, evidentemente, com $D \sim T^{1/4}A^{1/2}$ (Equação 5.14), no sentido de que os detalhes mais sutis da dependência da diversidade linguística com a temperatura (do ar/ecossistema) emergem de forma *ab initio* da estrutura intrínseca deste modelo termodinâmico. Além disso, o mesmo modelo fornece um arcabouço teórico para entender as relações de escala entre biodiversidade e área em termos de forças entrópicas (representadas pelo fator m na expressão a seguir) e de auto-exclusão (representadas pelo fator n nas expressões a seguir) por meio do uso das expressões $D \sim (T^{1/n}A)^{\zeta'}$ e $\zeta' = \frac{1}{1+\frac{m}{n}}$, em particular para $n = 2$ (Equações 5.18 e 5.19).

No Capítulo 6 apresentamos leis de escala emergentes em processos de classificação e ordenamento. Apontamos que tanto países, quanto idiomas, grupos étnicos e famílias linguísticas seguem à leis de escalas similares àquelas observadas em diversos sistemas ecológicos e urbanos (ROSE, 2006; BETTENCOURT et al., 2007; BATTY, 2008; TUNCAY, 2008). Iniciamos reportando funções hiperbólicas associadas à distribuições acumuladas de parâmetros linguísticos (Figura 25), econômicos (Figura 26), geográficos (Figura 27) e demográficos (Figura 28) dos países. Os expoentes associados à tais distribuições de tamanho foram $\delta = 1,26$ (idiomas), $\alpha = 0,96$ (áreas) e $\nu = 0,94$ (populações). Já para o Produto Interno Bruto reportamos uma dupla lei de escala sendo os expoentes na primeira e segunda região $\pi = 0,33$ e $\pi = 0,79$, respectivamente. Apontamos um comportamento qualitativo assintótico similar para as distribuições dos parâmetros geográfico, demográfico e econômico dos países (Figura 29). Sugerimos que no regime de grandes países é possível associar tais grandezas ao expoente de Fisher $\tau = 187/91$ de modo que somos tentados a compreender a distribuição dos países segundo tais grandezas como uma aproximação de um processo percolativo em duas dimensões (STAUFFER, 1985).

As distribuições acumuladas de idiomas (Figura 31) bem como de grupos étnicos (Figura 32) em função da população foram apresentadas e ambas foram descritas segundo

leis de potência de regime duplo. Enquanto para os idiomas apontamos $\iota = 0,57$ e $= 0,97$, reportamos para os grupos étnicos $\epsilon = 0,45$ e $\epsilon = 0,78$. Investigamos as leis de escala emergentes da classificação das famílias linguísticas segundo o número de idiomas (Figura 33) e o número de falantes (Figura 34). Para o primeiro caso, reproduzimos a curva do tipo Zipf com expoente $\theta = 2$ enquanto para o segundo apresentamos a inédita lei de potência com expoente $1,6$. Por fim, analisamos a distribuição de tamanhos de idiomas das catorze maiores famílias linguísticas contemporâneas e mostramos que doze destas famílias podem ser agrupadas, segundo o expoente das distribuições de Zipf, em três grupos caracterizados por expoentes $\kappa = 1,15 \pm 0,05$, $\kappa = 1,65 \pm 0,05$ e $\kappa = 2,05 \pm 0,05$. As duas exceções a estas classificações foram as famílias Afro-Asiática ($\kappa = 2,6$) e Nilo-Saariana ($\kappa = 1,4$) cuja evidente proximidade geográfica e antiguidade podem indicar um caminho para compreensão deste caráter distinto.

A partir das discussões precedentes novas perguntas emergem. Registramos aqui alguns dos questionamentos que povoam nossa mente enquanto concluímos este trabalho: Além do Produto Interno Bruto, existe(m) outro(s) índice(s) econômico(s) relacionado(s) mais claramente com a diversidade linguística? Como adquirir informações sobre a distribuição de tamanho dos idiomas e das suas famílias ao longo da história da humanidade? De que forma a ocorrência de grandes expansões ou contrações demográficas podem modificar os modelos propostos, ou podem neles ser acomodadas, sobretudo aquele discutido no Capítulo 5? Quais fatores levam um idioma à dominância em uma região? E, como estimular a diversidade linguística e cultural em um mundo aparentemente cada vez mais avesso à heterogeneidade?

REFERÊNCIAS

ÅSTRÖM, J. A.; HOLIAN, B. L.; TIMONEN, J. Universality in fragmentation. *Phys. Rev. Lett.*, American Physical Society, v. 84, p. 3061–3064, 2000. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevLett.84.3061>. Citado na página 70.

ABRAMS, D. M.; STROGATZ, S. H. Modelling the dynamics of language death. *Nature*, v. 424, n. 6951, p. 900–900, 2003. Disponível em: <https://doi.org/10.1038/424900a>. Citado 2 vezes nas páginas 57 e 58.

ALESINA, A.; SPOLAORE, E.; WACZIARG, R. Chapter 23 - Trade, Growth and the Size of Countries. In: AGHION, P.; DURLAUF, S. N. (Ed.). *Handbook of Economic Growth*. Elsevier, 2005. v. 1, p. 1499–1542. ISBN 1574-0684. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1574068405010233>. Citado na página 70.

ALTMANN, E. G.; GERLACH, M. Statistical laws in linguistics. In: ESPOSTI, M. D.; ALTMANN, E. G.; PACHET, F. (Ed.). *Creativity and Universality in Language*. Springer International Publishing, 2016. p. 7–26. ISBN 978-3-319-24403-7. Disponível em: <https://doi.org/10.1007/978-3-319-24403-7_2>. Citado na página 31.

ALTMANN, G. Prolegomena to Menzerath's law. *Glottometrika*, Bochum, v. 2, n. 2, p. 1–10, 1980. Citado na página 31.

ANDERSON, P. W. More is different. *Science*, American Association for the Advancement of Science, v. 177, n. 4047, p. 393–396, 1972. ISSN 0036-8075. Disponível em: <https://science.sciencemag.org/content/177/4047/393>. Citado na página 23.

ARISTÓTELES. *Metafísica*. Porto Alegre: Globo, 1969. Citado na página 45.

ARISTÓTELES. *A política*. São Paulo: Martins Fontes, 2006. Citado na página 70.

ARRHENIUS, O. Species and area. *Journal of Ecology*, Wiley, British Ecological Society, v. 9, n. 1, p. 95–99, 1921. ISSN 00220477, 13652745. Disponível em: <http://www.jstor.org/stable/2255763>. Citado na página 28.

AUERBACH, F. Das gesetz der bevölkerungskonzentration. *Petermanns Geographische Mitteilungen*, v. 59, p. 73–76, 1913. Citado na página 28.

AUSTIN, P. K.; SALLABANK, J. *The Cambridge Handbook of Endangered Languages*. Cambridge: Cambridge University Press, 2011. ISBN 9780521882156. Disponível em: <https://doi.org/10.1017/CBO9780511975981>. Citado 2 vezes nas páginas 45 e 54.

AXELSEN, J. B.; MANRUBIA, S. River density and landscape roughness are universal determinants of linguistic diversity. *Proceedings of the Royal Society B: Biological Sciences*, Royal Society, v. 281, n. 1784, p. 20133029, 2014. Disponível em: <https://doi.org/10.1098/rspb.2013.3029>. Citado na página 51.

BAGGS, I.; FREEDMAN, H. I. A mathematical model for the dynamics of interactions between a unilingual and a bilingual population: persistence versus extinction. *The Journal of Mathematical Sociology*, Routledge, v. 16, n. 1, p. 51–75, 1990. Disponível em: <https://doi.org/10.1080/0022250X.1990.9990078>. Citado 2 vezes nas páginas 57 e 58.

BAK, P. *How Nature Works: the science of self-organized criticality*. New York, NY: Springer New York, 1996. Citado na página 22.

BATTY, M. The size, scale, and shape of cities. *Science*, American Association for the Advancement of Science, v. 319, n. 5864, p. 769–771, 2008. ISSN 0036-8075. Disponível em: <https://science.sciencemag.org/content/319/5864/769>. Citado na página 90.

BELLWOOD, P. *First Migrants: Ancient Migration in Global Perspective*. Malden, MA, Oxford and Chichester: Wiley-Blackwell, 2013. Citado na página 68.

BELLWOOD, P.; RENFREW, C. (Ed.). *Examining the Farming/Language Dispersal Hypothesis*. Cambridge: McDonald Institute for Archaeological Research, 2002. Citado na página 54.

BENTZ, C. et al. The evolution of language families is shaped by the environment beyond neutral drift. *Nature Human Behaviour*, v. 2, n. 11, p. 816–821, 2018. Disponível em: <https://doi.org/10.1038/s41562-018-0457-6>. Citado 2 vezes nas páginas 50 e 51.

BERWICK, R. C.; CHOMSKY, N. *Por que apenas nós?* São Paulo: Editora Unesp, 2017. Citado 2 vezes nas páginas 21 e 22.

BETTENCOURT, L. M. A. et al. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, National Academy of Sciences, v. 104, n. 17, p. 7301–7306, 2007. ISSN 0027-8424. Disponível em: <https://www.pnas.org/content/104/17/7301>. Citado 2 vezes nas páginas 19 e 90.

BIAN, C. et al. Scaling laws and model of words organization in spoken and written language. *Europhysics Letters*, IOP Publishing, v. 113, n. 1, p. 18002, 2016. Disponível em: <http://dx.doi.org/10.1209/0295-5075/113/18002>. Citado na página 31.

BÍBLIA Sagrada: Nova Versão Transformadora. 1. ed. São Paulo: Mundo Cristão, 2016. Citado na página 7.

BOCQUET-APPEL, J.-P. When the world's population took off: the springboard of the neolithic demographic transition. *Science*, American Association for the Advancement of Science, v. 333, n. 6042, p. 560–561, 2011. ISSN 0036-8075. Disponível em: <https://doi.org/10.1126/science.1208880>. Citado na página 54.

BONGAARTS, J. Human population growth and the demographic transition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, Royal Society, v. 364, n. 1532, p. 2985–2990, 2009. Disponível em: <https://doi.org/10.1098/rstb.2009.0137>. Citado na página 54.

BROCK, W. Scaling in economics: a reader's guide. *Industrial and Corporate Change*, v. 8, n. 3, p. 409–446, 1999. ISSN 0960-6491. Disponível em: <https://dx.doi.org/10.1093/icc/8.3.409>. Citado 2 vezes nas páginas 19 e 45.

BROMHAM, L. et al. Rate of language evolution is affected by population size. *Proceedings of the National Academy of Sciences*, National Academy of Sciences, v. 112, n. 7, p. 2097–2102, 2015. ISSN 0027-8424. Disponível em: <https://doi.org/10.1073/pnas.1419704112>. Citado na página 75.

BROWN, J. H. Why are there so many species in the tropics? *Journal of Biogeography*, John Wiley & Sons, Ltd, v. 41, n. 1, p. 8–22, 2014. Disponível em: <https://doi.org/10.1111/jbi.12228>. Citado na página 63.

BURNSIDE, W. R. et al. Human macroecology: linking pattern and process in big-picture human ecology. *Biological Reviews*, John Wiley & Sons, Ltd, v. 87, n. 1, p. 194–208, 2012. Disponível em: <https://doi.org/10.1111/j.1469-185X.2011.00192.x>. Citado 2 vezes nas páginas 36 e 64.

BUTLER, R. P. et al. Catalog of nearby exoplanets. *The Astrophysical Journal*, IOP Publishing, v. 646, n. 1, p. 505–522, 2006. Disponível em: <http://dx.doi.org/10.1086/504701>. Citado na página 55.

CAEL, B. B.; SEEKELL, D. A. The size-distribution of Earth's lakes. *Scientific Reports*, v. 6, n. 1, p. 29633, 2016. Disponível em: <https://doi.org/10.1038/srep29633>. Citado na página 70.

CAMPBELL, L.; BELEW, A. *Cataloguing the World's Endangered Languages*. New York: Routledge, 2018. Citado na página 80.

CAMPBELL, L.; MIXCO, M. J. *A Glossary of Historical Linguistics*. Edinburgh: Edinburgh University Press, 2007. Citado na página 77.

CAMPI, X.; KRIVINE, H. Zipf's law in multifragmentation. *Physical Review C*, American Physical Society, v. 72, n. 5, p. 057602–, 2005. Disponível em: <https://link.aps.org/doi/10.1103/PhysRevC.72.057602>. Citado na página 70.

CARDILLO, M.; BROMHAM, L.; GREENHILL, S. J. Links between language diversity and species richness can be confounded by spatial autocorrelation. *Proceedings of the Royal Society B: Biological Sciences*, Royal Society, v. 282, n. 1809, p. 20142986, 2015. Disponível em: <https://doi.org/10.1098/rspb.2014.2986>. Citado 2 vezes nas páginas 36 e 37.

CAVALLI-SFORZA, L.; CAVALLI-SFORZA, F. *Quem somos?* São Paulo: Editora Unesp, 2002. Citado 2 vezes nas páginas 21 e 53.

CAVALLI-SFORZA, L. L. Genes, peoples, and languages. *Proceedings of the National Academy of Sciences*, v. 94, n. 15, p. 7719, 1997. Disponível em: <http://www.pnas.org/content/94/15/7719.abstract>. Citado na página 77.

CHEN, Y. The rank-size scaling law and entropy-maximizing principle. *Physica A: Statistical Mechanics and its Applications*, v. 391, n. 3, p. 767–778, 2012. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0378437111005577>. Citado na página 55.

CHOMSKY, N. *Que tipo de criaturas somos nós?* Petropólis: Vozes, 2018. Citado na página 22.

CIORAN, E. *Aveux et anathèmes*. Paris: Gallimard, 1987. Citado na página 45.

CIRILLO, R. *The Economics of Vilfredo Pareto*. London: Routledge, 1978. Citado 3 vezes nas páginas 27, 45 e 70.

CLAUSET, A.; SHALIZI, C. R.; NEWMAN, M. E. J. Power-law distributions in empirical data. *SIAM Review*, Society for Industrial and Applied Mathematics, v. 51, n. 4, p. 661–703, 2009. Disponível em: <https://doi.org/10.1137/070710111>. Citado na página 28.

CLEMENTE, J.; GONZÁLEZ-VAL, R.; OLLOQUI, I. Zipf's and Gibrat's laws for migrations. *The Annals of Regional Science*, v. 47, n. 1, p. 235–248, 2011. Disponível em: <https://doi.org/10.1007/s00168-010-0367-7>. Citado na página 73.

CLINGINGSMITH, D. Are the world's languages consolidating? The dynamics and distribution of language populations. *The Economic Journal*, John Wiley & Sons, Ltd, v. 127, n. 599, p. 143–176, 2017. Disponível em: <https://doi.org/10.1111/ecoj.12257>. Citado na página 74.

COUTINHO, K.; GOMES, M. A. F.; ADHIKARI, S. K. Robust scaling in fragmentation from d = 1 to 5. *Europhysics Letters*, IOP Publishing, v. 18, n. 2, p. 119–124, 1992. Disponível em: <http://dx.doi.org/10.1209/0295-5075/18/2/006>. Citado na página 39.

CRISTELLI, M.; BATTY, M.; PIETRONERO, L. There is more than a power law in Zipf. *Scientific Reports*, v. 2, n. 1, p. 812, 2012. Disponível em: <https://doi.org/10.1038/srep00812>. Citado na página 72.

DARWIN, C. *The descent of man, and selection in relation to sex*. London: John Murray, 1871. Citado na página 36.

DE GENNES, P. G. *Scaling Concepts in Polymer Physics*. Ithaca/London: Cornell University Press, 1979. ISBN 9780801412035. Citado 2 vezes nas páginas 43 e 59.

DE OLIVEIRA, V. M. et al. Bounded fitness landscapes and the evolution of the linguistic diversity. *Physica A: Statistical Mechanics and its Applications*, v. 368, n. 1, p. 257–261, 2006. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378437106001610>. Citado na página 50.

DE OLIVEIRA, V. M.; GOMES, M. A. F.; TSANG, I. Theoretical model for the evolution of the linguistic diversity. *Physica A: Statistical Mechanics and its Applications*, v. 361, n. 1, p. 361 – 370, 2006. ISSN 0378-4371. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378437105007053>. Citado 4 vezes nas páginas 50, 54, 57 e 58.

DESMET, K.; ORTUÑO-ORTÍN, I.; WACZIARG, R. Linguistic cleavages and economic development. In: GINSBURGH, V.; WEBER, S. (Ed.). *The Palgrave Handbook of Economics and Language*. London: Palgrave Macmillan UK, 2016. p. 425–446. ISBN 978-1-137-32505-1. Disponível em: <https://doi.org/10.1007/978-1-137-32505-1_16>. Citado na página 46.

DESMET, P.; COWLING, R. Using the species–area relationship to set baseline targets for conservation. *Ecology and Society*, v. 9, n. 2, p. [online], 2004. Disponível em: <http://www.ecologyandsociety.org/vol9/iss2/art11/>. Citado na página 39.

DI GUILMI, C.; GAFFEO, E.; GALLEGATI, M. Power law scaling in world income distribution. *Economics Bulletin*, Vanberbilt University, v. 15, p. 1–7, 2003. Disponível em: <http://hdl.handle.net/10453/9990>. Citado na página 72.

DURRETT, R.; LEVIN, S. Spatial models for species-area curves. *Journal of Theoretical Biology*, v. 179, n. 2, p. 119–127, 1996. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0022519396900533>. Citado 2 vezes nas páginas 42 e 44.

EBERHARD, D. *Em defesa das línguas minoritárias do Brasil*. Associação Internacional de Linguística SIL-Brasil, 2013. Disponível em: <https://www.silbrasil.org.br/resources/archives/76953>. Citado na página 17.

EDWARDS, H. M. *Riemann's Zeta Function*. New York: Dover, 2001. Citado na página 34.

ESTOUP, J. *Gammes sténographiques: méthode & exercices pour l'acquisition de la vitesse*. Paris: Institut Sténographique, 1916. Citado na página 28.

EVANS, N. *Dying Words: Endangered Languages and What They Have to Tell Us*. Malden, MA: Wiley, 2010. (The Language Library). ISBN 9780631233053. Citado 2 vezes nas páginas 63 e 75.

FALCONER, K. *Fractal Geometry: Mathematical Foundations and Applications*. West Sussex: Wiley, 2014. Citado 3 vezes nas páginas 24, 26 e 43.

FEULNER, G. et al. On the origin of the surface air temperature difference between the hemispheres in Earth's present-day climate. *Journal of Climate*, v. 26, n. 18, p. 7136–7150, 2013. Disponível em: <https://doi.org/10.1175/JCLI-D-12-00636.1>. Citado 2 vezes nas páginas 63 e 65.

FRAINER, A. et al. Opinion: cultural and linguistic diversities are underappreciated pillars of biodiversity. *Proceedings of the National Academy of Sciences*, p. 202019469, 2020. Disponível em: <http://www.pnas.org/content/early/2020/10/06/2019469117.abstract>. Citado na página 16.

FURCERI, D. Zipf's law and world income distribution. *Applied Economics Letters*, Routledge, v. 15, n. 12, p. 921–923, 2008. Disponível em: <https://doi.org/10.1080/13504850600972261>. Citado na página 72.

GABAIX, X. Power laws in economics and finance. *Annual Review of Economics*, Annual Reviews, v. 1, n. 1, p. 255–294, 2009. Disponível em: <https://doi.org/10.1146/annurev.economics.050708.142940>. Citado na página 45.

GALILEI, G. *Diálogo sobre os dois máximos sistemas do mundo ptolomaico e copernicano*. São Paulo: Associação Filosófica Scientiae Studia: Editora 34, 2011. Citado na página 21.

GALILEI, G. *Duas novas ciências*. São Paulo: Nova Stella; Ched, s/d. Citado 2 vezes nas páginas 19 e 27.

GAVIN, M. C. et al. Toward a Mechanistic Understanding of Linguistic Diversity. *BioScience*, v. 63, n. 7, p. 524–535, 2013. ISSN 0006-3568. Disponível em: <https://doi.org/10.1525/bio.2013.63.7.6>. Citado 3 vezes nas páginas 50, 68 e 90.

GAVIN, M. C.; STEPP, J. R. Rapoport's rule revisited: Geographical distributions of human languages. *PLOS ONE*, Public Library of Science, v. 9, n. 9, p. e107623–, 2014. Disponível em: <https://doi.org/10.1371/journal.pone.0107623>. Citado 3 vezes nas páginas 60, 63 e 74.

GILLOOLY, J. F. et al. Effects of size and temperature on metabolic rate. *Science*, v. 293, n. 5538, p. 2248, 2001. Disponível em: <http://science.sciencemag.org/content/293/5538/2248.abstract>. Citado na página 64.

GINSBURGH, V.; WEBER, S. *How Many Languages Do We Need?* Princeton University Press, 2011. ISBN 9780691136899. Disponível em: <https://www.jstor.org/stable/j.ctt7s9tj>. Citado na página 46.

GOLDENFELD, N.; KADANOFF, L. P. Simple lessons from complexity. *Science*, American Association for the Advancement of Science, v. 284, n. 5411, p. 87–89, 1999. ISSN 0036-8075. Disponível em: <https://science.sciencemag.org/content/284/5411/87>. Citado na página 23.

GOMES, M. A. F. et al. Scaling relations for diversity of languages. *Physica A: Statistical Mechanics and its Applications*, v. 271, n. 3, p. 489 – 495, 1999. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0378437199002496>. Citado 8 vezes nas páginas 16, 32, 33, 36, 37, 50, 69 e 87.

GONG, T.; SHUAI, L.; ZHANG, M. Modelling language evolution: examples and predictions. *Physics of Life Reviews*, v. 11, n. 2, p. 280–302, 2014. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1571064513001863>. Citado na página 50.

GONZÁLEZ-VAL, R.; SANSO-NAVARRO, M. Gibrat's law for countries. *Journal of Population Economics*, v. 23, n. 4, p. 1371–1389, 2010. Disponível em: <https://doi.org/10.1007/s00148-009-0246-7>. Citado 2 vezes nas páginas 72 e 73.

GORENFLO, L. J. et al. Co-occurrence of linguistic and biological diversity in biodiversity hotspots and high biodiversity wilderness areas. *Proceedings of the National Academy of Sciences*, v. 109, n. 21, p. 8032, 2012. Disponível em: <http://www.pnas.org/content/109/21/8032.abstract>. Citado 2 vezes nas páginas 36 e 37.

GREENHILL, S. J. Demographic correlates of language diversity. In: BOWERN, C.; EVANS, B. (Ed.). *The Routledge Handbook of Historical Linguistics*. Amsterdam: Routledge, 2014. p. 557–578. ISBN 978-0-415-52789-7. Citado 2 vezes nas páginas 77 e 78.

GREENHILL, S. J. et al. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences*, v. 114, n. 42, p. E8822, 2017. Disponível em: <http://www.pnas.org/content/114/42/E8822.abstract>. Citado na página 50.

GRIN, F.; SFREDDO, C.; VAILLANCOURT, F. *The Economics of the Multilingual Workplace*. New York: Routledge, 2013. Citado na página 47.

HAMILTON, M. J. et al. The complex structure of hunter–gatherer social networks. *Proceedings of the Royal Society B: Biological Sciences*, Royal Society, v. 274, n. 1622, p. 2195–2203, 2007. Disponível em: <https://doi.org/10.1098/rspb.2007.0564>. Citado na página 90.

HAMMARSTRÖM, H. A full-scale test of the language farming dispersal hypothesis. *Diachronica*, John Benjamins, v. 27, n. 2, p. 197–213, 2010. ISSN 0176-4225. Disponível em: <https://www.jbe-platform.com/content/journals/10.1075/dia.27.2.02ham>. Citado na página 78.

HAMMARSTRÖM, H. *Ethnologue* 16/17/18th editions: A comprehensive review. *Language*, Linguistic Society of America, v. 91, n. 3, p. 723–737, 2015. ISSN 00978507, 15350665. Disponível em: <http://www.jstor.org/stable/24672170>. Citado na página 15.

HAMMARSTRÖM, H. et al. *Glottolog 4.3*. Jena: [s.n.], 2020. Max Planck Institute for the Science of Human History. Disponível em: <https://glottolog.org/>. Citado 3 vezes nas páginas 18, 63 e 66.

HARRISON, K. D. *When Languages Die*. New York: Oxford University Press, 2007. Citado na página 75.

HAUSER, M. D. et al. The mystery of language evolution. *Frontiers in Psychology*, v. 5, p. 401, 2014. ISSN 1664-1078. Disponível em: <https://www.frontiersin.org/article/10.3389/fpsyg.2014.00401>. Citado na página 22.

HEAPS, H. S. *Information Retrieval: Computational and Theoretical Aspects*. Orlando: Academic Press, Inc., 1978. ISBN 0123357500. Citado na página 31.

HERDAN, G. The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics. *Biometrika*, v. 45, n. 1-2, p. 222–228, 1958. Disponível em: <https://doi.org/10.1093/biomet/45.1-2.222>. Citado na página 31.

HERÓDOTO. *História*. Rio de Janeiro: Nova Fronteira, 2019. Citado na página 49.

HERRMANN, H. J.; HONG, D. C.; STANLEY, H. E. Backbone and elastic backbone of percolation clusters obtained by the new method of 'burning'. *Journal of Physics A: Mathematical and General*, IOP Publishing, v. 17, n. 5, p. L261–L266, 1984. Disponível em: <https://doi.org/10.1088%2F0305-4470%2F17%2F5%2F008>. Citado 2 vezes nas páginas 41 e 89.

HILLEBRAND, H. On the generality of the latitudinal diversity gradient. *The American Naturalist*, The University of Chicago Press, v. 163, n. 2, p. 192–211, 2004. Disponível em: <https://doi.org/10.1086/381004>. Citado na página 63.

HOBOHM, C. Characterization and ranking of biodiversity hotspots: centres of species richness and endemism. *Biodiversity & Conservation*, v. 12, n. 2, p. 279–287, 2003. Disponível em: <https://doi.org/10.1023/A:1021934910722>. Citado 2 vezes nas páginas 39 e 88.

HOGAN-BRUN, G. *Linguanomics*. London: Bloomsbury Academic, 2017. Citado na página 47.

HUA, X. et al. The ecological drivers of variation in global language diversity. *Nature Communications*, v. 10, n. 1, p. 2047, 2019. Disponível em: <https://doi.org/10.1038/s41467-019-09842-2>. Citado 5 vezes nas páginas 36, 63, 64, 66 e 67.

IGLAUER, S.; PALUSZNY, A.; BLUNT, M. J. Simultaneous oil recovery and residual gas storage: a pore-level analysis using in situ x-ray micro-tomography. *Fuel*, v. 103, p. 905–914, 2013. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0016236112005236>. Citado na página 70.

KLEIBER, M. Body size and metabolism. *Hilgardia*, v. 6, n. 11, p. 315–353, 1932. Disponível em: <https://doi.org/10.3733/hilg.v06n11p315>. Citado na página 27.

LADYMAN, J.; LAMBERT, J.; WIESNER, K. What is a complex system? *European Journal for Philosophy of Science*, v. 3, n. 1, p. 33–67, 2013. Disponível em: <https://doi.org/10.1007/s13194-012-0056-8>. Citado na página 22.

LEWIS, M.; SIMONS, G. Assessing endangerment: expanding Fishman's gids. *Revue Roumaine de Linguistique*, v. 55, 2010. Disponível em: <https://www.lingv.ro/RRL-2010.html>. Citado na página 16.

LI, B.; MADRAS, N.; SOKAL, A. D. Critical exponents, hyperscaling, and universal amplitude ratios for two- and three-dimensional self-avoiding walks. *Journal of Statistical Physics*, v. 80, n. 3, p. 661–754, 1995. Disponível em: <https://doi.org/10.1007/BF02178552>. Citado na página 43.

LIMPERT, E.; STAHEL, W. A.; ABBT, M. Log-normal distributions across the sciences. *BioScience*, v. 51, n. 5, p. 341–352, 2001. ISSN 0006-3568. Disponível em: <https://doi.org/10.1641/0006-3568(2001)051%5B0341:LNDATS%5D2.0.CO;2>. Citado na página 75.

LOH, J.; HARMON, D. A global index of biocultural diversity. *Ecological Indicators*, v. 5, n. 3, p. 231 – 241, 2005. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1470160X0500018X>. Citado na página 37.

LOTKA, A. J. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, Washington Academy of Sciences, v. 16, n. 12, p. 317–323, 1926. ISSN 00430439. Disponível em: <http://www.jstor.org/stable/24529203>. Citado na página 27.

LOVEJOY, S.; SCHERTZER, D.; LADOY, P. Fractal characterization of inhomogeneous geophysical measuring networks. *Nature*, v. 319, n. 6048, p. 43–44, 1986. Disponível em: <https://doi.org/10.1038/319043a0>. Citado na página 43.

MACE, R.; PAGEL, M. A latitudinal gradient in the density of human languages in North America. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, Royal Society, v. 261, n. 1360, p. 117–121, 1995. Disponível em: <https://doi.org/10.1098/rspb.1995.0125>. Citado na página 63.

MAFFI, L. Linguistic, cultural, and biological diversity. *Annual Review of Anthropology*, Annual Reviews, v. 34, n. 1, p. 599–617, 2005. Disponível em: <https://doi.org/10.1146/annurev.anthro.34.081804.120437>. Citado na página 37.

MANDELBROT, B. How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science*, American Association for the Advancement of Science, v. 156, n. 3775, p. 636–638, 1967. ISSN 0036-8075. Disponível em: <https://science.sciencemag.org/content/156/3775/636>. Citado na página 24.

MANDELBROT, B. *Les objets fractals*. Paris: Flammarion, 1975. Citado na página 24.

MANDELBROT, B. *Objectos Fractais*. 2. ed. Lisboa: Gradiva, 1998. Citado na página 24.

MANDELBROT, B. *The fractalist: memoir of a scientific maverick*. New York: Pantheon Books, 2012. Citado na página 23.

MANNA, S. S. Large-scale simulation of avalanche cluster distribution in sand pile model. *Journal of Statistical Physics*, v. 59, n. 1, p. 509–521, 1990. Disponível em: <https://doi.org/10.1007/BF01015580>. Citado na página 70.

MANNION, P. D. et al. The latitudinal biodiversity gradient through deep time. *Trends in Ecology & Evolution*, v. 29, n. 1, p. 42–50, 2014. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0169534713002358>. Citado na página 65.

MARIN, F.; BELUFFI, C. Computing the minimal crew for a multi-generational space travel towards Proxima Centauri b. *Journal of the British Interplanetary Society*, v. 71, p. 45–52, 2018. Citado 2 vezes nas páginas 55 e 90.

MATACIC, C. World's largest linguistics database is getting too expensive for some researchers. *Science*, 2020. Disponível em: <doi:10.1126/science.abb2422>. Citado na página 15.

MAUPERTUIS, P.-L. M. de. *Réflexions philosophiques sur l'origine des langues et la signification des mots*. [S.l.: s.n.], 1740. Citado na página 49.

MITCHELL, M. *Complexity: A Guided Tour*. New York: Oxford University Press, Inc., 2009. ISBN 0195124413. Citado na página 23.

MOORE, J. L. et al. The distribution of cultural and biological diversity in Africa. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, Royal Society, v. 269, n. 1501, p. 1645–1653, 2002. Disponível em: <https://doi.org/10.1098/rspb.2002.2075>. Citado 2 vezes nas páginas 36 e 37.

MUFWENE, S. S. *The Ecology of Language Evolution*. Cambridge: Cambridge University Press, 2001. ISBN 9780521791380. Disponível em: <https://www.cambridge.org/core/books/ecology-of-language-evolution/16CE992F71B9A066F508A3A74BE4DDE5>. Citado na página 36.

NETTLE, D. Explaining global patterns of language diversity. *Journal of Anthropological Archaeology*, v. 17, n. 4, p. 354–374, 1998. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0278416598903282>. Citado na página 63.

NETTLE, D. Linguistic fragmentation and the wealth of nations: the Fishman-Pool hypothesis reexamined. *Economic Development and Cultural Change*, The University of Chicago Press, v. 48, n. 2, p. 335–348, 2000. Disponível em: <https://doi.org/10.1086/452461>. Citado na página 45.

NEWMAN, M. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, Taylor & Francis, v. 46, n. 5, p. 323–351, 2005. Disponível em: <https://doi.org/10.1080/00107510500052444>. Citado 2 vezes nas páginas 28 e 35.

PAGEL, M. Human language as a culturally transmitted replicator. *Nature Reviews Genetics*, v. 10, n. 6, p. 405–415, 2009. Disponível em: <https://doi.org/10.1038/nrg2560>. Citado na página 36.

PAGEL, M.; MACE, R. The cultural wealth of nations. *Nature*, v. 428, n. 6980, p. 275–278, 2004. Disponível em: <https://doi.org/10.1038/428275a>. Citado 2 vezes nas páginas 36 e 37.

PARISI, G. Complex systems: a physicist's viewpoint. *Physica A: Statistical Mechanics and its Applications*, v. 263, n. 1, p. 557–564, 1999. Disponível em: <https://www.sciencedirect.com/science/article/pii/S037843719800524X>. Citado na página 22.

PASTOR-SATORRAS, R.; WAGENSBERG, J. The maximum entropy principle and the nature of fractals. *Physica A: Statistical Mechanics and its Applications*, v. 251, n. 3, p. 291 – 302, 1998. ISSN 0378-4371. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378437197005712>. Citado na página 55.

PATRIARCA, M. et al. Modeling two-language competition dynamics. *Advances in Complex Systems*, World Scientific Publishing Co., v. 15, n. 03n04, p. 1250048, 2012. Disponível em: <https://doi.org/10.1142/S0219525912500488>. Citado 2 vezes nas páginas 57 e 58.

PATRIARCA, M.; HEINSALU, E.; LEONARD, J. L. *Languages in Space and Time: Models and Methods from Complex Systems Theory*. Cambridge: Cambridge University Press, 2020. ISBN 9781108480659. Disponível em: <https://www.cambridge.org/core/books/languages-in-space-and-time/BB26B97D9BDB4A8E203519124F26003A>. Citado na página 58.

PERELTSVAIG, A. *Languages of the world: an introduction*. Cambridge: Cambridge University Press, 2012. Citado na página 77.

PETERSON, M. A. Galileo's discovery of scaling laws. *American Journal of Physics*, American Association of Physics Teachers, v. 70, n. 6, p. 575–580, 2002. Disponível em: <https://doi.org/10.1119/1.1475329>. Citado na página 27.

PIETRONERO, L. Complexity ideas from condensed matter and statistical physics. *Europhysics News*, v. 39, n. 6, p. 26–29, 2008. Disponível em: <https://doi.org/10.1051/epn:2008603>. Citado na página 23.

PIMENTEL, D.; PIMENTEL, M. H. *Food, energy, and society*. Boca Raton: CRC Press, 2008. Citado na página 53.

PINTO, C. M. A.; LOPES, A. M.; MACHADO, J. A. T. A review of power laws in real life phenomena. *Communications in Nonlinear Science and Numerical Simulation*, v. 17, n. 9, p. 3558–3578, 2012. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1007570412000354>. Citado na página 28.

PLOTKIN, J. B. et al. Species-area curves, spatial aggregation, and habitat specialization in tropical forests. *Journal of Theoretical Biology*, v. 207, n. 1, p. 81–99, 2000. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0022519300921581>. Citado 2 vezes nas páginas 42 e 44.

POOL, J. National development and language diversity. In: *In Joshua A. Fishman (Ed.), Advances in the sociology of language*. The Hague: Mouton, 1972. p. 213–230. Citado na página 45.

RENFREW, C. *Archaeology and language*. London: Penguin Group, 1989. Citado 2 vezes nas páginas 42 e 54.

ROBERTS, S.; WINTERS, J. Linguistic diversity and traffic accidents: lessons from statistical studies of cultural traits. *PLOS ONE*, Public Library of Science, v. 8, n. 8, p. 1–13, 2013. Disponível em: <https://doi.org/10.1371/journal.pone.0070902>. Citado na página 44.

ROMAINE, S. Preserving endangered languages. *Language and Linguistics Compass*, John Wiley & Sons, Ltd (10.1111), v. 1, n. 1-2, p. 115–132, 2007. Disponível em: <https://doi.org/10.1111/j.1749-818X.2007.00004.x>. Citado 2 vezes nas páginas 45 e 54.

ROSE, A. K. Cities and countries. *Journal of Money, Credit and Banking*, v. 38, n. 8, p. 2225–2245, 2006. Disponível em: <http://www.jstor.org/stable/4123049>. Citado 2 vezes nas páginas 72 e 90.

ROSENZWEIG, M. L. *Species diversity in space and time*. Cambridge: Cambridge University Press, 1995. Citado 2 vezes nas páginas 28 e 37.

RUBNER, M. Ueber den einfluss der körpergrösse auf stoff- und kraftwechsel. *Zeitschrift für Biologie*, Urban & Schwarzenberg München Lehmann 1865-1971, München ; Berlin [u.a.], v. 19, p. 535–562, 1883. ISSN 0372-8366. Disponível em: <http://uri.gbv.de/document/gvk:ppn:129463736>. Citado na página 27.

SALMON, P. B. Origin of language debate in the eighteenth century. In: OERNER, E. F. K.; ASHER, R. E. (Ed.). *Concise History of the Language Sciences*. Amsterdam: Pergamon, 1995. p. 184–187. ISBN 978-0-08-042580-1. Disponível em: <http://www.sciencedirect.com/science/article/pii/B9780080425801500350>. Citado na página 49.

SALOTTI, J.-M. Minimum number of settlers for survival on another planet. *Scientific Reports*, v. 10, n. 1, p. 9700, 2020. Disponível em: <https://doi.org/10.1038/s41598-020-66740-0>. Citado 2 vezes nas páginas 55 e 90.

SANTOS, M. R. F.; GOMES, M. A. F. Revisiting scaling relations for linguistic diversity. *Physica A: Statistical Mechanics and its Applications*, v. 532, p. 121821, 2019. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378437119310684>. Citado 4 vezes nas páginas 36, 50, 68 e 88.

SANTOS, M. R. F.; GOMES, M. A. F. A heuristic model for the scaling linguistic diversity-area. *Physica A: Statistical Mechanics and its Applications*, v. 555, p. 124622, 2020. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378437120303034>. Citado 2 vezes nas páginas 49 e 88.

SANTOS, M. R. F.; GOMES, M. A. F. Entropic-self-avoiding mean field model for the scaling linguistic diversity-area. Manuscrito submetido para publicação. 2021. Citado na página 88.

SANTOS, M. R. F.; GOMES, M. A. F. A simple model with biological significance for the scaling linguistic diversity-area. Manuscrito submetido para publicação. 2021. Citado 2 vezes nas páginas 42 e 88.

SAUSSURE, F. de. *Curso de linguística geral*. São Paulo: Cultrix, 2012. Citado na página 58.

SCHMIDT-NIELSEN, K. *Scaling: why is animal size so important?* Cambridge: Cambridge University Press, 1984. Citado na página 27.

SCHROEDER, M. *Fractals, chaos, power laws*. New York: Dover, 2009. Citado na página 24.

SCHULZE, C.; STAUFFER, D. Monte Carlo simulation of the rise and the fall of languages. *International Journal of Modern Physics C*, World Scientific Publishing Co., v. 16, n. 05, p. 781–787, 2005. Disponível em: <https://doi.org/10.1142/S0129183105007479>. Citado na página 74.

SERENO, M. I. Four analogies between biological and cultural/linguistic evolution. *Journal of Theoretical Biology*, v. 151, n. 4, p. 467–507, 1991. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0022519305803662>. Citado na página 51.

SHANNON, C. E. Prediction and entropy of printed english. *The Bell System Technical Journal*, v. 30, n. 1, p. 50–64, 1951. Citado na página 30.

SIMONS, G. F.; FENNIG, C. D. *Ethnologue: Languages of the World*. 20. ed. Dallas: SIL International, 2017. Disponível em: <http://www.ethnologue.com/20>. Citado 4 vezes nas páginas 15, 37, 50 e 69.

SING, D. K. et al. A continuum from clear to cloudy hot-Jupiter exoplanets without primordial water depletion. *Nature*, v. 529, n. 7584, p. 59–62, 2016. Disponível em: <https://doi.org/10.1038/nature16068>. Citado na página 55.

SMITH, J. M.; SZATHMARY, E. *The major transitions in evolution*. Oxford: Oxford University Press, 1997. Citado na página 22.

STANLEY, H. E. *Introduction to Phase Transitions and Critical Phenomena*. Oxford: Oxford University Press, 1971. Citado na página 53.

STAUFFER, D. *Introduction to percolation theory*. London and Philadelphia: Taylor & Francis, 1985. Citado 2 vezes nas páginas 70 e 90.

STAUFFER, D.; SCHULZE, C. Microscopic and macroscopic simulation of competition between languages. *Physics of Life Reviews*, v. 2, n. 2, p. 89–116, 2005. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1571064505000114>. Citado na página 58.

STORCH, D.; MARQUET, P.; BROWN, J. *Scaling Biodiversity*. Cambridge: Cambridge University Press, 2007. ISBN 9780521876025. Disponível em: <https://www.cambridge.org/core/books/scaling-biodiversity/E956EF65E3D7727A0A836AF38A40A893>. Citado na página 28.

SUTHERLAND, W. J. Parallel extinction risk and global distribution of languages and species. *Nature*, v. 423, n. 6937, p. 276–279, 2003. Disponível em: <https://doi.org/10.1038/nature01607>. Citado 5 vezes nas páginas 37, 48, 63, 74 e 75.

TALLAVAARA, M.; ERONEN, J. T.; LUOTO, M. Productivity, biodiversity, and pathogens influence the global hunter-gatherer population density. *Proceedings of the National Academy of Sciences*, v. 115, n. 6, p. 1232, 2018. Disponível em: <http://www.pnas.org/content/115/6/1232.abstract>. Citado na página 65.

TATTERSALL, I. *Masters of the planet: the search for our human origins*. New York: St. Martin's Press, 2012. Citado na página 22.

THURNER, S.; KLIMEK, P.; HANEL, R. *Introduction to the Theory of Complex Systems*. Oxford: Oxford University Press, 2018. Citado na página 23.

TORRE, I. G. et al. On the physical origin of linguistic laws and lognormality in speech. *Royal Society Open Science*, v. 6, n. 8, p. 191023, 2019. Disponível em: <https://doi.org/10.1098/rsos.191023>. Citado 3 vezes nas páginas 30, 31 e 74.

TRIANTIS, K. A.; GUILHAUMON, F.; WHITTAKER, R. J. The island species–area relationship: biology and statistics. *Journal of Biogeography*, John Wiley & Sons, Ltd, v. 39, n. 2, p. 215–231, 2012. Disponível em: <https://doi.org/10.1111/j.1365-2699.2011.02652.x>. Citado na página 44.

TUNCAY, Ç. Model of world: her cities, languages and countries. *International Journal of Modern Physics C*, v. 19, n. 03, p. 471–484, 2008. Disponível em: <https://doi.org/10.1142/S0129183108012261>. Citado 2 vezes nas páginas 73 e 90.

TURCOTTE, D. L. Fractals and fragmentation. *Journal of Geophysical Research*, American Geophysical Union (AGU), v. 91, n. B2, p. 1921, 1986. Disponível em: <https://doi.org/10.1029%2Fjb091ib02p01921>. Citado na página 70.

TURVEY, S. T.; PETTORELLI, N. Spatial congruence in language and species richness but not threat in the world's top linguistic hotspot. *Proceedings of the Royal Society B: Biological Sciences*, Royal Society, v. 281, n. 1796, p. 20141644, 2014. Disponível em: <https://doi.org/10.1098/rspb.2014.1644>. Citado na página 37.

UPADHYAY, R. K.; HASNAIN, S. I. Linguistic diversity and biodiversity. *Lingua*, v. 195, p. 110–123, 2017. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0024384117302449>. Citado na página 36.

VICSEK, T. *Fractal Growth Phenomena*. World Scientific, 1992. 528 p. ISBN 978-981-02-0668-0. Disponível em: <https://doi.org/10.1142/1407>. Citado na página 70.

WEIL, D. N.; SHARMA, A. *Economic growth*. 3. ed. Boston: Pearson Education, 2013. Citado na página 45.

WEST, G. *Scale*. New York: Penguin Press, 2017. Citado 2 vezes nas páginas 19 e 28.

WEST, G. B.; BROWN, J. H.; ENQUIST, B. J. A general model for the origin of allometric scaling laws in biology. *Science*, American Association for the Advancement of Science, v. 276, n. 5309, p. 122–126, 1997. ISSN 0036-8075. Disponível em:

<https://science.sciencemag.org/content/276/5309/122>. Citado 2 vezes nas páginas 19 e 27.

WHALEN, D. H.; SIMONS, G. F. Endangered language families. *Language*, Linguistic Society of America, v. 88, n. 1, p. 155–173, 2012. ISSN 00978507, 15350665. Disponível em: <http://www.jstor.org/stable/41348886>. Citado na página 79.

WHITFIELD, J. Across the curious parallel of language and species evolution. *PLOS Biology*, Public Library of Science, v. 6, n. 7, p. e186–, 2008. Disponível em: <https://doi.org/10.1371/journal.pbio.0060186>. Citado 2 vezes nas páginas 36 e 49.

WICHMANN, S. On the power-law distribution of language family sizes. *Journal of Linguistics*, Cambridge University Press, v. 41, n. 1, p. 117–131, 2005. Disponível em: <https://doi.org/10.1017/S002222670400307X>. Citado 2 vezes nas páginas 78 e 87.

WILDER, B. T. et al. The importance of indigenous knowledge in curbing the loss of language and biodiversity. *BioScience*, v. 66, n. 6, p. 499–509, 2016. ISSN 0006-3568. Disponível em: <https://doi.org/10.1093/biosci/biw026>. Citado 2 vezes nas páginas 16 e 45.

WILLIG, M. R.; KAUFMAN, D. M.; STEVENS, R. D. Latitudinal gradients of biodiversity: pattern, process, scale, and synthesis. *Annual Review of Ecology, Evolution, and Systematics*, Annual Reviews, v. 34, n. 1, p. 273–309, 2003. Disponível em: <https://doi.org/10.1146/annurev.ecolsys.34.012103.144032>. Citado na página 63.

WORLD BANK. *World Development Indicators (2017) - GDP (current US$)*. 2017. Disponível em: <https://data.worldbank.org/indicator/NY.GDP.MKTP.CD>. Citado 2 vezes nas páginas 18 e 70.

WORLD BANK. *World Development Indicators (2017) - Land area (sq. km)*. 2017. Disponível em: <https://data.worldbank.org/indicator/AG.LND.TOTL.K2>. Citado 2 vezes nas páginas 18 e 72.

WORLD BANK. *Climate Change Knowledge Portal*. 2020. Disponível em: <https://climateknowledgeportal.worldbank.org/>. Citado 2 vezes nas páginas 18 e 60.

YAGHOUBI, M. et al. Neuronal avalanche dynamics indicates different universality classes in neuronal cultures. *Scientific Reports*, v. 8, n. 1, p. 3417, 2018. Disponível em: <https://doi.org/10.1038/s41598-018-21730-1>. Citado na página 70.

YULE, G. U. II.—A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, Royal Society, v. 213, n. 402-410, p. 21–87, 1925. Disponível em: <https://doi.org/10.1098/rstb.1925.0002>. Citado na página 27.

ZANETTE, D. H. Self-similarity in the taxonomic classification of human languages. *Advances in Complex Systems*, World Scientific Publishing Co., v. 04, n. 02n03, p. 281–286, 2001. Disponível em: <https://doi.org/10.1142/S0219525901000206>. Citado na página 78.

ZANETTE, D. H. Demographic growth and the distribution of language sizes. *International Journal of Modern Physics C*, World Scientific Publishing Co., v. 19, n. 02, p. 237–247, 2008. Disponível em: <https://doi.org/10.1142/S0129183108012042>. Citado na página 74.

ZHANG, J.; YU, T. Allometric scaling of countries. *Physica A: Statistical Mechanics and its Applications*, v. 389, n. 21, p. 4887–4896, 2010. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378437110006072>. Citado 2 vezes nas páginas 19 e 40.

ZHOU, Z.; SZYMANSKI, B. K.; GAO, J. Modeling competitive evolution of multiple languages. *PLOS ONE*, Public Library of Science, v. 15, n. 5, p. e0232888–, 2020. Disponível em: <https://doi.org/10.1371/journal.pone.0232888>. Citado 2 vezes nas páginas 57 e 58.

ZIPF, G. K. *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley, 1949. Citado 3 vezes nas páginas 28, 29 e 58.

ZIPF, G. K. *The Phycho-Biology of Language*. Cambridge: MIT Press Classic, 1965. Citado 2 vezes nas páginas 28 e 30.

APÊNDICE A – ARTIGOS PUBLICADOS EM PERIÓDICOS

Neste apêndice apresentamos os dois artigos publicados até o presente momento sobre os temas da tese. Os resultados apresentados no primeiro, *"Revisiting scaling relations for linguistic diversity"*, publicado no periódico *Physica A* em 2019, são discutidos nos Capítulos 3 e 6. Os resultados apresentados no segundo, *"A heuristic model for the scaling linguistic diversity-area"*, publicado no mesmo periódico em 2020, são discutidos no Capítulo 4.

Physica A 532 (2019) 121821

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Revisiting scaling relations for linguistic diversity

M.R.F. Santos, M.A.F. Gomes *

Departamento de Física, Universidade Federal de Pernambuco, 50670-901 Recife, PE, Brazil

HIGHLIGHTS

- Scaling relation between linguistic diversity (D) and Gross Domestic Product (GDP).
- Same exponents relate D × Area and D × GDP, independent of the binarization.
- The cumulative distributions of D and GDP exhibit the same asymptotic behavior.

ARTICLE INFO

Article history:
Received 26 November 2018
Received in revised form 10 June 2019
Available online 21 June 2019

Keywords:
Linguistic diversity
Economy
Scaling laws

ABSTRACT

The relations between linguistic diversity and geographic and ecological parameters have been studied over the last decades. Like biological species many languages are in risk of extinction. The effect of these extinctions must be felt in the loss of knowledge contained in such languages and their consequences on the economy could be noticed. Here a detailed statistical analysis shows that in addition to the recognized diversity of languages–area relation there is a remarkable scaling relation between the size of the economy of a country, as measured by the Gross Domestic Product (*GDP*), and the number of languages spoken in its territory. In addition, it is shown that the cumulative distributions of linguistic diversity and *GDP* exhibit the same type of asymptotic behavior.

1. Introduction

With the recent proliferation of digital databases a good deal of skepticism and criticism must be exercised when examining the large number of studies that look for quantitative relations between cultural, demographic, economic, and geographical variables. If this is not made, spurious positive correlations could be advanced. Within this framework, Roberts and Winters have recently pointed out the risk of over emphasizing such correlations [1]. To illustrate, those authors show that a positive correlation emerge, for instance, between linguistic diversity and the percentual number of annual traffic fatalities. They observe, in addition, that, unfortunately, some of these studies are receiving media attention without a widespread understanding of the complexities of the issue, and there is a risk that poorly controlled studies could affect policy. In spite of the inherent risk of misunderstand the relation among variables in complex phenomena, some important socio-economic scaling relations are widely known, useful, and now are on a more satisfactory theoretical basis, and as an example we can cite the original Pareto's law of income in Economics [2] which says that the distribution of income is very uneven, with the richest population controlling the largest part of the resources [3].

On the other hand, in a very distant scenario at the dawn of the theoretical conceptions that define the foundation for the Western world, it may be opportune to recall the opening words in Aristotle's Metaphysics "All men by nature are

* Corresponding author.
E-mail addresses: maelyson@df.ufpe.br (M.R.F. Santos), mafg@ufpe.br (M.A.F. Gomes).

https://doi.org/10.1016/j.physa.2019.121821
0378-4371/

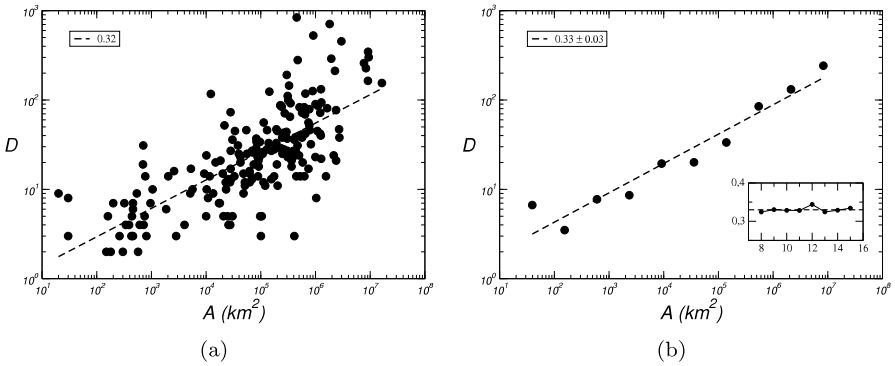


Fig. 1. Linguistic diversity as a function of land area. The dashed lines are best fits whose slope gives the exponent $z = 0.32$ ($r = 0.73$, $n = 194$) for the raw data (a), and $z = 0.33 \pm 0.03$ ($r = 0.97$) for binned data (b).

actuated with the desire of knowledge" [4]. This is a type of statement meaning that the fact of living must be put as a kind of knowledge. It is tempting to complement this widely known thought with an idea of uncertain origin although frequently associated with the Romanian-born philosopher Emil Cioran, that "one does not inhabit a country; one inhabits a language" [5]. If some plausibility exists in these two statements, it will be reasonable to conclude that a world with large linguistic diversity represents a world with more knowledge. Although recent studies have followed research into the amount of information in texts and languages [6,7], this is not a scope of our work. In the meantime, most people studying the languages at risk of extinction seem to believe that the loss of linguistic diversity represents not only a reduction in the repertoire of (cultural) world views, but also a reduction in biological, ecological, geographical, and technological information, to name only a few aspects [8,9]. As the economic activity generally privileges knowledge, we are led to conclude that linguistic diversity may be associated with some basic economic indicator. This hypothesis was investigated by Pool [10] which analyzed the relationship between the fraction of the population, the largest native-language group, and the gross domestic product per person per year for 133 countries. As well as re-examined by Nettle [11]. Still, according to Desmet et al. "the relation between linguistic diversity and the level of economic development has been somewhat understudied" [12]. Despite possible methodological ambiguities, the *GDP* has for decades been the most widely used indicator to measure the economic activity of any country [13]. The argument advanced in the preceding lines led us to the subject of the present work, which is the examination in detail of the existence of a connection between economic activity, symbolized by *GDP*, and linguistic diversity.

2. Results and discussion

We used the linguistic diversity data from the twentieth edition of the Ethnologue [14], published in 2017, which lists 7099 languages spoken in 236 countries. For each country is given the corresponding total diversity D which is defined as the number of living languages used as a first language. The country with the greatest linguistic diversity is Papua New Guinea with 840 living languages followed by Indonesia and Nigeria with 709 and 527 living languages, respectively. However, the fact that all these countries do not have the largest *GDP*s is immaterial in a statistical study involving a much more extended ensemble of countries.

We first analyze the relation between linguistic diversity D and the land area A (in square kilometers) of the countries. The data on land areas were obtained from The World Bank's World Development Indicators [15]. Because information about the *GDP* is only available for 194 countries, as discussed below, we present the relation between linguistic diversity and the land area for these 194 countries instead of the 236 listed in the Ethnologue.

In biology and biogeography the species–area relationship (SAR) is an important tool in the study of species diversity and the shapes of species–area curves have been discussed [16]. The power law curve model is one of the most used descriptions. Scaling laws appear in the modeling of various complex phenomena transcending the boundaries of physics [17–19].

As frequently observed with biological species [20], our analysis of relation between linguistic diversity (D) and area (A), as shown in Fig. 1, is well described by a power law scaling dependence of type

$$D \sim A^z \tag{1}$$

where the exponent $z = 0.32$ is displayed from Fig. 1(a). This exponent is somewhat smaller to that presented by Gomes et al. [21] but equal within the statistical uncertainty to that reported by Loh et al. [22]. A much more clear trendline emerges after the division of the countries into 10 groups (bins) according to the area and then calculate the average diversity of living languages in each bin. After this separation, the exponent obtained was $z = 0.33 \pm 0.03$ (Fig. 1(b)). The dependence between the exponent z with the number of boxes used in the separation was analyzed. By varying the

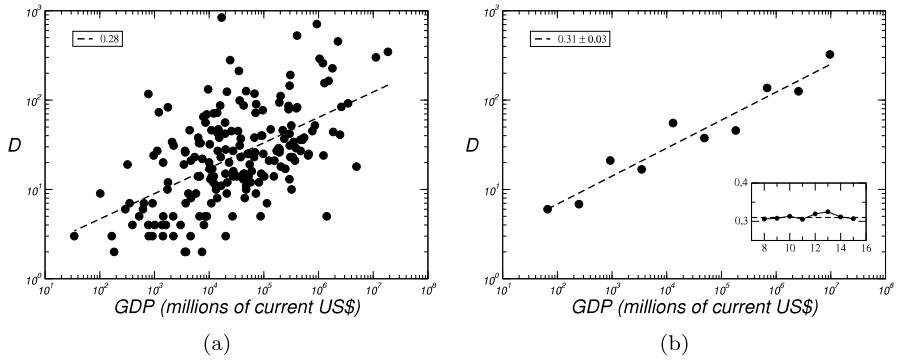


Fig. 2. Linguistic diversity as a function of *GDP*. The dashed lines are best fits whose slope gives the exponent $\zeta = 0.28$ ($r = 0.56$, $n = 194$) for the raw data (a), and $\zeta = 0.31 \pm 0.03$ ($r = 0.96$) for binned data (b).

number of bins between 8 and 15, we see in the inset in Fig. 1(b) that the z value fluctuates slightly around the value presented above. The agreement between the result without separation in bins and the result when the number of bins varies, attests the robustness of the scaling relation between the linguistic diversity and the land area. This exponent z is comparable to the ones we observe in ecology for the relation between species diversity and area [23,24].

In order to study a possible relation between linguistic diversity and economic performance, we chose as indicator the *GDP*. The data for *GDP*, relative to 2016, are in units of millions of current US dollars and were obtained from The World Bank's World Development Indicators [25]. A limiting factor is that *GDP* data is available for a smaller set of countries (194) when compared with the linguistic diversity.

In a similar way to the approach discussed above, we propose that

$$D \sim (GDP)^{\zeta} \tag{2}$$

where the exponent $\zeta = 0.28$ is obtained from Fig. 2(a). In addition, we divide the countries into 10 groups (bins) but this time according to the *GDP* and then calculate the average diversity of living languages in each bin. After this separation, a significative increase in correlation was observed and the exponent obtained was $\zeta = 0.31 \pm 0.03$ (Fig. 2(b)), in conformity with the exponent z of Eq. (1). Here we also analyze the dependence between the value of the exponent ζ with the number of bins used in the separation of the countries. In the inset of Fig. 2(b) presenting the results of the variation of the number of bins between 8 and 15, we see that the zeta exponent is robust and independent of binarization. It is evident that the largest economies are statistically related to a greater linguistic diversity. The cause of this scaling law may be a consequence that GDP tends to increase with area, but a more direct and synergistic coupling between linguistic diversity and economic activity and therefore in GDP could be present.

A possible relation between linguistic diversity and the *GDP* per capita was additionally investigated. The data about this economic parameter was also obtained from the World Bank. However, a global analysis like the one presented in Fig. 2 does not point to the existence of a power-law relation between linguistic diversity and the *GDP* per capita. This finding is similar to that pointed out by Sutherland [26].

Finally, the cumulative distribution of the linguistic diversity and the *GDP* among the various countries was investigated. Asymptotically the number of countries N with a linguistic diversity greater than D is given by

$$N \sim D^{-B}, \tag{3}$$

with $B = 1.26 \pm 0.03$ for $D > 60$, as shown in Fig. 3(a). This last exponent is somewhat larger than that previously reported [21]. Fig. 3(a) shows that it is increasingly uncommon to preserve the unity of countries with great linguistic diversity. In correspondence with this graph, in Fig. 3(b) we have the accumulated distribution for the *GDP*. As previously, we observe an asymptotic behavior in the form of a power law

$$N \sim (GDP)^{-\beta}, \tag{4}$$

with $\beta = 0.80 \pm 0.01$ for *GDP* greater than 10^5 million dollars. This indicates that the larger the economy the smaller the number of countries able to support it. There is a compromise, so: D tends to increase with *GDP* but it is proportionally difficult in general to preserve the unity of countries with great linguistic diversity.

3. Conclusions

The linguistic diversity is a subject of general theoretical interest in physics and correlated sciences [21,26–30]. In summary, from the discussion presented here the following main conclusions emerge: (i) there is a robust scaling relation between linguistic diversity and *GDP* over more than five decades in this variable. (ii) The exponents that relate $D \times A$

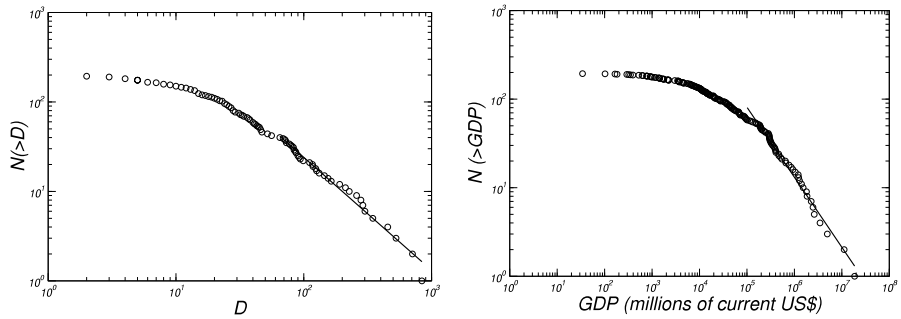


Fig. 3. (a) Number of countries with a linguistic diversity greater than D as a function of D. The solid line is a best fit whose slope gives the exponent $B = 1.26 \pm 0.03$ for $60 < D < 840$, totalizing 40 countries. (b) Number of countries with a *GDP* greater than *GDP* as a function of *GDP*. The solid line is a best fit whose slope gives the exponent $\beta = 0.80 \pm 0.01$ for the symptotic region totalizing 58 countries.

and $D \times GDP$ are the same within the error bars, and they are robust independent of the binarization. (iii) It is not easy to maintain countries with a great linguistic diversity, albeit in spite of this constraint, these countries tend to present a very strong economic performance as measured by the *GDP*, still obeying the same scaling law (Eq. (2)). Of course, we cannot rule out that the relationship between linguistic diversity and GDP could be caused by the fact that this latter variable tends to increase with the area of the country. However, a more direct and synergistic coupling between linguistic diversity and economic activity and therefore in GDP may also be present. From the point of view of planners or policy makers, the results of this work may introduce some warning about possible economic implications of the impoverishment of language diversity. To conclude, as a matter of reflection, we must always be fully aware of the controversial nature that frequently tends to take on the discussions that try to relate economic indices to quality of life and other social indicators.

Funding

The work is supported by National Council for Scientific and Technological Development (CNPq), Brazil, National Council for the Improvement of Higher Education (CAPES), Brazil, PROEX 534/2018 N^o. 23038.003382/2018-39 and PRONEX from MCT/CNPq/FACEPE, Brazil, N^o. APQ-1330-1.05/10.

References

[1] S. Roberts, J. Winters, Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits, PLoS ONE 8 (8) (2013) 1–13.
[2] R. Cirillo, The Economics of Vilfredo Pareto, Routledge, 1978.
[3] B.M. Boghosian, Kinetics of wealth and the Pareto law, Phys. Rev. E 89 (2014) 042804.
[4] Aristotle, The Metaphysics, Vol. I, Dover, 2007.
[5] E.M. Cioran, Anathemas and Admirations, Arcade, 2012.
[6] M.A. Montemurro, Quantifying the information in the long-range order of words: Semantic structures and universal linguistic constraints, Cortex 55 (2014) 5–16.
[7] C. Bentz, D. Alikaniotis, M. Cysouw, R. Ferrer-i Cancho, The entropy of words—Learnability and expressivity across more than 1000 languages, Entropy 19 (6) (2017) 1–32.
[8] P.K. Austin, J. Sallabank, Cambridge Handbooks in Language and Linguistics, Cambridge University Press, Cambridge, 2011.
[9] S. Romaine, Preserving endangered languages, Lang. Linguist. Compass 1 (12) (2007) 115–132.
[10] J. Pool, National development and language diversity, in: J.A. Fishman (Ed.), Advances in the Sociology of Language, The Hague: Mouton, 1972, pp. 213–230.
[11] D. Nettle, Linguistic fragmentation and the wealth of nations: The Fishman–Pool hypothesis reexamined, Econom. Dev. Cult. Chang. 48 (2) (2000) 335–348.
[12] K. Desmet, I. Ortuño-Ortín, R. Wacziarg, in: V. Ginsburgh, S. Weber (Eds.), The Palgrave Handbook of Economics and Language, Palgrave Macmillan UK, London, 2016, pp. 425–446.
[13] D.N. Weil, A. Sharma, Economic Growth, third ed., Pearson Education, 2013.
[14] G.F. Simons, C.D. Fennig, Ethnologue: Languages of the World, twentieth ed., SIL International, 2017.
[15] World Bank, World development indicators, 2017, Land area (sq. km).
[16] E. Tjørve, Shapes and functions of species–area curves: a review of possible models, J. Biogeogr. 30 (6) (2003) 827–835.
[17] G.B. West, J.H. Brown, B.J. Enquist, A general model for the origin of allometric scaling laws in biology, Science 276 (5309) (1997) 122–126.
[18] W.A. Brock, Scaling in economics: a reader's guide, Ind. Corp. Change 8 (3) (1999) 409–446.
[19] L.M.A. Bettencourt, The origins of scaling in cities, Science 340 (6139) (2013) 1438–1441.
[20] M.L. Rosenzweig, Species Diversity in Space and Time, Cambridge, 1995.
[21] M.A.F. Gomes, G.L. Vasconcelos, I.J. Tsang, I.R. Tsang, Scaling relations for diversity of languages, Physica A 271 (3) (1999) 489–495.
[22] J. Loh, D. Harmon, A global index of biocultural diversity, Ecol. Indic. 5 (3) (2005) 231–241.
[23] P. Desmet, R. Cowling, Using the species–area relationship to set baseline targets for conservation, Ecol. Soc. 9 (2) (2004) [online].

[24] C. Hobohm, Characterization and ranking of biodiversity hotspots: centres of species richness and endemism, Biodivers. Conserv. 12 (2) (2003) 279–287.
[25] World Bank, World development indicators, 2017, GDP (current US$).
[26] W.J. Sutherland, Parallel extinction risk and global distribution of languages and species, Nature 423 (2003) 276–279.
[27] D. Nettle, Explaining global patterns of language diversity, J. Anthropol. Archaeol. 17 (4) (1998) 354–374.
[28] V.M. de Oliveira, M.A.F. Gomes, I.R. Tsang, Theoretical model for the evolution of the linguistic diversity, Physica A 361 (1) (2006) 361–370.
[29] J.B. Axelsen, S. Manrubia, River density and landscape roughness are universal determinants of linguistic diversity, Proc. R. Soc. Lond. Ser. B: Biol. Sci. 281 (1788) (2014) 1–8.
[30] R.K. Upadhyay, S.I. Hasnain, Linguistic diversity and biodiversity, Lingua 195 (2017) 110–123.

Physica A 555 (2020) 124622

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

A heuristic model for the scaling linguistic diversity-area

M.R.F. Santos, M.A.F. Gomes *

Departamento de Física, Universidade Federal de Pernambuco, 50670-901 Recife, PE, Brazil

ARTICLE INFO

Article history:
Received 19 December 2019
Available online 25 April 2020

Keywords:
Complex systems
Linguistic diversity
Scaling laws

ABSTRACT

Here is introduced a simple heuristic model which emerges from a type of variational framework aimed to solve the problem of the relation between linguistic diversity and area from a maximization procedure. We show that this model reproduces the robust scaling law for linguistic diversity versus area observed today on Earth for the largest 147 countries with area superior to 18,000 km^2.

1. Introduction

The interest in the origin and in the diversification of languages can be tracked back at least to the ancient Egypt of the Pharaohs, as commented in §II, Book II, of Herodotus The Histories [1], in the V century before our era. In modern times, 2200 years after, in the XVIII century the mathematician and philosopher Pierre-Louis Moreau de Maupertuis, pioneer of the namesake variational principle and of the linguistic studies, published in 1740 *Réflexions philosophiques sur l'origine des langues et la signification des mots* [2]. In the present work we are interested in develop a simple model inspired in very general arguments that has as consequence the reproduction of the scaling relation between the linguistic diversity and area observed today on the Earth for the largest 147 countries with areas superior to 18,000 km^2 representing more than 99% of the population of the planet. This is accomplished with the introduction of the ansatz discussed in Section 2 which emerges from a type of variational framework aimed to solve the problem of the relation between linguistic diversity and area from a maximization procedure. Although interested in variational principles and in the linguistic diversity, Maupertuis could not develop a similar procedure as reported here because of the lack of data on the distribution of languages on Earth in the XVIII century. Today and in the last decades the situation is very different, and the subject of the linguistic diversity has attracted much of the attention of many researchers beyond the linguistic community [3–7]. A similar rise of interest occurred in the study of complex phenomena where scaling laws develop a fundamental role [8–10].

2. The maximum diversity model and discussion

One of the most basic relations involving the linguistic diversity, D, concerns the robust power law scaling relating D with the area, A, where these languages are spoken, $D \sim A^z$, along almost six decades of variability in area [11,12]. Here the linguistic diversity is defined as the number of living languages used as a first language. Recent data from Ethnologue [13] led to the plot exhibited in Fig. 1 [12]. Looking at this figure one could conclude that, irrespective binarization, $z = 1/3$ within typical fluctuations of 1%, along more than five orders of magnitude in area.

However, an alternative scaling description could be advanced, as shown by dotted and continuous lines in Fig. 2. In this last case, $z = 1/2$, along 2.5 decades in area (continuous line), in a very representative interval that consider small, medium and large countries, with areas larger than 18,000 km^2, corresponding to 147 countries from the total of 194

* Corresponding author.
E-mail addresses: maelyson@df.ufpe.br (M.R.F. Santos), mafg@ufpe.br (M.A.F. Gomes).

https://doi.org/10.1016/j.physa.2020.124622
0378-4371/

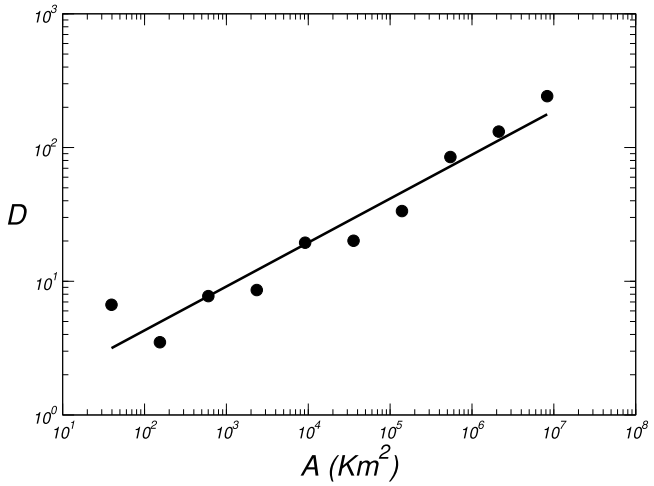


Fig. 1. Log–log plot of the linguistic diversity D as a function of the area A. See Ref. [12] for detail. After binarization of the data, the continuous line refers to the overall fit along all area classes and gives $z = 0.33$.

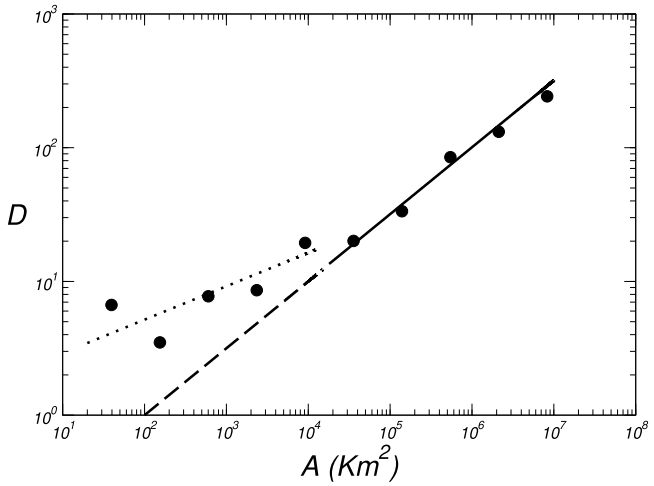


Fig. 2. The same data as in Fig. 1 fitted in a different way with the continuous (dotted) line representing the scaling $D \sim A^z$ with $z = 1/2$ ($z = 1/4$) for the largest (smallest) 147 (47) countries with areas larger (smaller) than 18,000 km^2. For the extrapolation represented by the dashed line see Section 2, fifth paragraph.

states, leaving the remaining 47 small countries (dotted line), most of them of insular nature, as belonging to a different group associated with another $D-A$ power law scaling, along 3 decades in area, with $z = 1/4$. That group of 147 countries is quite representative in area. The sum of its areas corresponds to more than 99% of the landmasses of the Earth. Thus, the focus of the present work is to examine a possible genesis for the scaling $D \sim A^{1/2}$ observed for states belonging, in general, to the large continuous mass of lands forming the great continents.

It is interesting to ask if after a long maturation period the global linguistic diversity observed today for a large fraction of all the present countries on Earth has evolved to a maximum linguistic diversity possible compatible with the geographic, economic, political, and ethnic-cultural constraints. Obviously any *ab initio* attack to this problem is a very complex task, perhaps never achievable. Our objective is to present a simple zero-order model that presents an answer to this question which is able to be statistically verified.

Here, it is assumed that the distribution of languages on a certain local domain on the surface of the Earth in a distant past was distributed on a set $S = \{a_i\}$ of disconnected domains whose total area a_i is associated with a single language i from the total repertoire of languages $1, 2, \ldots, i, \ldots, D$, where D is the total linguistic diversity (Fig. 3). Each area a_i stands for the sum of all areas of settlements or regions where the language i is spoken. Thus, in the beginning it is conjectured that there is a single mother language and $a_1 = L^2$ represents the area available to all humans speaking the primordial language (Fig. 3(a)). The migrational movements can culminate in geographical separation leading to the language diversification, so that the area associated to the initial language is separated, for example, into two distinct areas,

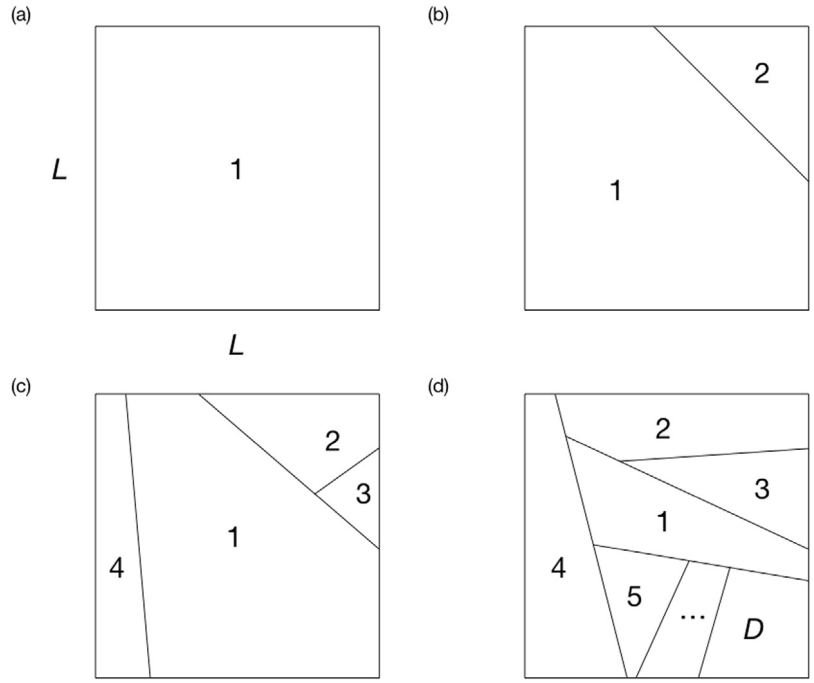


Fig. 3. Schematic illustration of the dynamics of evolution of areas a_i where languages are spoken, for a near fixed total available area L^2 or, equivalently, a near fixed population. See Section 2, fourth and fifth paragraphs.

each one corresponding to a different language (Fig. 3(b)). Such population groups speaking these two languages may be further divided and isolated at a later time to form new sparse domains (Fig. 3(c)). Those sparse domains were separated by unpopulated regions, as mountain ranges, deserts and water courses, that prevented clashes between speakers of different languages in ancient times. These effectively unoccupied spaces are schematically symbolized in Fig. 3 by the network of continuous rectilinear segments separating the various domains of areas $a_i s$. Over time, different groups can enter in collaboration or in war, and the idioms can enter in expansion, fusion, extinction or can preserve the respective identities with variable degree of linguistic exchanges. In our model, it is assumed that the dynamics of these complex historical and ethnic-cultural processes over millennia can lead to the mosaic of linguistic diversity exhibited today in our planet.

The linguistic domains previously introduced are thus assumed to be disjoint and submitted to the sum rule $A = \sum_i a_i$, where A is the total populated area on Earth, which, in turn, is a small fraction of the total available land not covered by waters on Earth. Within the framework presented in this section, the condition to achieve maximum diversity in the past is solved with the *ansatz* $a_i = \alpha.i$, where α is a constant with dimension of area. In these terms, the sum rule, which represents the sum of area terms is an arithmetic progression of areas α, 2α, 3α, 4α, ..., $D\alpha$ leading to

$$A = \alpha\,(1 + 2 + \cdots + D) = \alpha \frac{D(D+1)}{2}, \tag{1}$$

which is equivalent to the scaling

$$D \approx \left(\frac{2}{\alpha}\right)^{1/2} A^{1/2}, \tag{2}$$

for $D \gg 1$. That is, the resulting scaling law found in this maximization process is just that embodied in the continuous fit of Fig. 2. From the extrapolated dashed line in Fig. 2, is obtained $D \approx 1$ for $A \approx 100$ km^2 therefore we can estimate $\alpha = 200$ km^2 or 20,000 ha, an area usually associated with the region sufficient to assure the life of approximately 500 hunter-gatherers [14]. Equivalently, this means in terms of our model that in some distant past the total occupied areas a_i associated with a different language i defined an arithmetic progression correspondingly populated by groups with approximately $500i$ humans, with $i = 1, 2, 3, \ldots$. It is tantalizing to associate these groups of near 500 humans with quanta of migrant people. In this way, the previous ansatz is equivalent to the quantization of the areas (that is, the values of the populated areas are given by integer multiples of a characteristic area α) or the quantization of the number of migrant groups speaking a same tongue within these areas.

The advantage of this simple model, with few and reasonable assumptions, is twofold: (a) it gives the scaling observed today for the relation between linguistic diversity and area, valid for all countries with appreciable areas ($A \geq 18{,}000$

km^2), i.e. along almost three decades in area. (b) It suggests that the exponent 1/2 can be connected with a maximization principle as outlined in the beginning of this section. It should be noted that the exponent 1/2 is present in various physical systems, it is common in important scaling phenomena, including critical phenomena in condensed matter physics [15]. We suppose that there is a primordial epoch in which the linguistic diversity was already large ($D \gg 1$) corresponding to a constant area so that our argument is initially applicable in that reference primordial time. With the passage of the centuries, the total population, the boundary of the several populated areas, the linguistic diversity, and the total area changes with a very complex dynamics, including intermixing of populations speaking different idioms in a same region, but the signature $D \sim A^{1/2}$ remained stable as a relic of that epoch at least in an appreciable fraction of the inhabited area presently associated with countries with areas superior to 18,000 km^2.

It is fair to ask if the model presented above had the opportunity to materialize on Earth. We are led to imagine that at all times when the population was stationary (or near stationary) and, correspondingly, the occupied area was stationary (or near stationary) we had an opportunity for this model to be effectively active. Once reached the situation characterized by the scaling $D \sim A^{1/2}$, the dynamics could be locked in this configuration of maximal linguistic diversity till to the present. Even though in the last two centuries the human population has grown unprecedentedly, the population growth rate of Homo sapiens has been low for most of its history [16]. This discussion becomes even more relevant when it is observed that in the last years the relations between the prehistoric farming development and growth in population size have been discussed. Demographic expansions were investigated [17] and the Neolithic Demographic Transition have been related with the "farming/language dispersal hypothesis" [18]. It is noteworthy that the effects of neolithic population expansions can still be perceived in current patterns of language diversity [19].

One final question arises: Can the evolution of linguistic distribution in the future result in a reduction in the diversity-area scaling exponent from 1/2 to 1/4 for the region comprising countries with an area of over 18,000 km^2? We believe so, and this reduction in diversity, which is already discussed in the literature [20], has also been suggested from computer simulations [21].

3. Conclusions

In agreement with many other cases where a maximization procedure is responsible for the emergence of scaling laws [22,23], we have introduced a simple arithmetic–geometric ansatz leading to the scaling $D \sim A^{1/2}$ observed today for the linguistic diversity, D, found in areas A satisfying $A \geq 18{,}000$ km^2 on Earth. We conjecture that the universe of languages has undergone an evolution in the remote past, in which linguistic diversity has increased in a complex manner, leaving us with the scaling law currently observed as a signature of that primordial distribution of languages.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work is supported by National Council for Scientific and Technological Development (CNPq), Brazil, National Council for the Improvement of Higher Education (CAPES), Brazil, Programa de Excelência Acadêmica (PROEX), Brazil 534/2018 Nº. 23038.003382/2018-39 and Programa de Apoio a Núcleos de Excelência (PRONEX), Brazil from MCT/CNPq/FACEPE, Nº. APQ-1330-1.05/10.

References

[1] Herodotus, The histories, Oxford University Press, 2008.
[2] P.-L.M. de Maupertuis, Réflexions philosophiques sur l'origine des langues et la signification des mots, 1740.
[3] D. Nettle, Explaining global patterns of language diversity, J. Anthropol. Archaeol. 17 (4) (1998) 354–374, http://dx.doi.org/10.1006/jaar.1998.0328.
[4] V.M. de Oliveira, P.R.A. Campos, M. Gomes, I.R. Tsang, Bounded fitness landscapes and the evolution of the linguistic diversity, Physica A 368 (1) (2006) 257–261, http://dx.doi.org/10.1016/j.physa.2005.11.058.
[5] S.J. Greenhill, C.-H. Wu, X. Hua, M. Dunn, S.C. Levinson, R.D. Gray, Evolutionary dynamics of language systems, Proc. Natl. Acad. Sci. 114 (42) (2017) E8822–E8829, http://dx.doi.org/10.1073/pnas.1700388114.
[6] R.K. Upadhyay, S.I. Hasnain, Linguistic diversity and biodiversity, Lingua 195 (2017) 110–123, http://dx.doi.org/10.1016/j.lingua.2017.06.002.
[7] X. Hua, S.J. Greenhill, M. Cardillo, H. Schneemann, L. Bromham, The ecological drivers of variation in global language diversity, Nature Commun. 10 (1) (2019) 2047, http://dx.doi.org/10.1038/s41467-019-09842-2.
[8] G.B. West, J.H. Brown, B.J. Enquist, A general model for the origin of allometric scaling laws in biology, Science 276 (5309) (1997) 122–126, http://dx.doi.org/10.1126/science.276.5309.122.
[9] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509, http://dx.doi.org/10.1126/science.286.5439.509.
[10] C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics, Rev. Modern Phys. 81 (2) (2009) 591–646, http://dx.doi.org/10.1103/RevModPhys.81.591.
[11] M.A.F. Gomes, G.L. Vasconcelos, I.J. Tsang, I.R. Tsang, Scaling relations for diversity of languages, Physica A 271 (3) (1999) 489–495, http://dx.doi.org/10.1016/S0378-4371(99)00249-6.

[12] M.R.F. Santos, M.A.F. Gomes, Revisiting scaling relations for linguistic diversity, Physica A 532 (2019) 121821, http://dx.doi.org/10.1016/j.physa.2019.121821.
[13] G.F. Simons, C.D. Fennig, Ethnologue: Languages of the World, twentieth ed., SIL International, 2017.
[14] D. Pimentel, M.H. Pimentel, Food, Energy, and Society, CRC Press, 2008.
[15] H.E. Stanley, Introduction to Phase Transitions and Critical Phenomena, Oxford University Press, 1971.
[16] J. Bongaarts, Human population growth and the demographic transition, Philos. Trans. R. Soc. B 364 (1532) (2009) 2985–2990, http://dx.doi.org/10.1098/rstb.2009.0137.
[17] J.-P. Bocquet-Appel, When the world's population took off: The springboard of the neolithic demographic transition, Science 333 (6042) (2011) 560–561, http://dx.doi.org/10.1126/science.1208880.
[18] P. Bellwood, C. Renfrew (Eds.), Examining the Farming/Language Dispersal Hypothesis, McDonald Institute for Archaeological Research, 2002.
[19] C. Renfrew, Archaeology and Language, Penguin Group, 1989.
[20] P.K. Austin, J. Sallabank, Cambridge Handbooks in Language and Linguistics, Cambridge University Press, Cambridge, 2011.
[21] V.M. de Oliveira, M.A.F. Gomes, I.R. Tsang, Theoretical model for the evolution of the linguistic diversity, Physica A 361 (1) (2006) 361–370, http://dx.doi.org/10.1016/j.physa.2005.06.069.
[22] R. Pastor-Satorras, J. Wagensberg, The maximum entropy principle and the nature of fractals, Physica A 251 (3) (1998) 291–302, http://dx.doi.org/10.1016/S0378-4371(97)00571-2.
[23] Y. Chen, The rank-size scaling law and entropy-maximizing principle, Physica A 391 (3) (2012) 767–778, http://dx.doi.org/10.1016/j.physa.2011.07.010.